Towards Rigorously Tested & Reliable Machine Learning for Health

By Michael Karl Oberst

B.A., Harvard University (2012)
S.M., Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by:	Michael Karl Oberst Department of Electrical Engineering and Computer Science June 7, 2023
Certified by:	David Sontag Professor of Electrical Engineering and Computer Science Thesis Supervisor
Accepted by:	Leslie A. Kolodziejski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students

Towards Rigorously Tested & Reliable Machine Learning for Health

by

Michael Karl Oberst

Submitted to the Department of Electrical Engineering and Computer Science
on June 7th, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

When can we rely on machine learning in high-risk domains like healthcare? In the long-term, we want machine learning systems to be as reliable as any FDA-approved medication or diagnostic test. Building reliable models is complicated by the need for causal reasoning and robust performance. To support decision-making, we want to draw causal conclusions about the impact of model recommendations (e.g., will recommending a particular drug lead to better patient outcomes?). Moreover, we want our models to perform well across different hospitals and patient populations, including those that differ from the hospitals / populations seen during model development.

These objectives run into limitations of what our data can tell us without further assumptions. For instance, we only observe outcomes for the treatments that were actually prescribed to patients, not all possible treatments. Similarly, we do not observe performance on every conceivable hospital where a model might be deployed, but only on the (typically much more limited) data we have access to.

In this thesis, I approach these challenges using tools from causality and statistics, incorporating external knowledge into the process of both model validation and design. External knowledge can come from a variety of sources, including human experts (e.g., clinicians) or gold-standard data (e.g., from randomized trials). First, I introduce methods for assessing and improving the credibility of causal inference, including methods to help domain experts "sanity check" the causal reasoning of models for decision-making, identify under-represented populations in causal analyses, and incorporate limited experimental data to improve the credibility of causal conclusions. Second, I introduce tools for building robust predictive models by incorporating domain knowledge of plausible variation across environments: Both estimating worst-case predictive performance (e.g., accuracy) of models under domain-specific changes in the data generating process, as well as optimizing models to obtain optimal worst-case performance.

Thesis Supervisor: David Sontag
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I want to thank my advisor, David Sontag. I feel blessed to have had an advisor that pushed me to ask the right questions in my research, gave me the freedom to pursue whatever questions I thought were worth asking, and was always there to support me as a person. I'm also grateful for the support of my thesis committee members, Tommi Jaakkola and Jonas Peters, who gave invaluable feedback on this thesis.

I also want to thank all the wonderful folks here at MIT with whom I've had the opportunity to do research, including Christina X. Ji, Fredrik D. Johansson, Helen Zhou, Ilker Demirel, Justin Lim, Ming-Chieh Shih, Sooraj Boominathan, and Zeshan Hussain. Thank you all for the joy of debating and kicking around ideas among friends, whether at a whiteboard or over Zoom.

I've had many inspiring collaborators outside of MIT, without whom much of the work presented in this thesis would not have been possible, including Alexander D'Amour (Google), Dennis Wei (IBM), Gabriel Brat (Beth Israel Deaconess Medical Center), Jonas Peters (ETH Zurich), Kush R. Varshney (IBM), Leora Horwitz (NYU Langone), Minmin Chen (Google), Nikolaj Thams (University of Copenhagen), Sanjat Kanjilal (Mass General Brigham), Saul Blecker (NYU Langone), Steve Yadlowsky (Google), Tian Gao (IBM), and Yuyan Wang (Google). I want to thank Sanjat Kanjilal in particular for his mentorship and for his role as ever-available sounding board for my ideas, and Leora Horwitz and Saul Blecker for many fruitful and insightful conversations. In addition, I want to thank Nikolaj Thams and Jonas Peters for a wonderful experience collaborating, first across time-zones and then in-person when we had the benefit of hosting Nikolaj in Cambridge.

I also want to thank the entire Clinical Machine Learning group, who made the shared experience of research a joy, including those I never got a chance to write papers with: Ahmed Alaa, Chandler Squires, Elizabeth Bondi-Kelly, Hunter Lang, Hussein Mozannar, Irene Chen, Monica Agrawal, Rahul Krishnan, Rebecca Boiarsky, Shannon Shen, and Uri Shalit. The lab has truly felt like a home over the years, and I thank you all for making it feel that way. Whether attending the weddings of lab members, engaging in late-night conversations during conferences at the group Airbnb, or just enjoying the camaraderie in E25, it's been a truly unique and rewarding experience.

I owe an immense debt of gratitude to my mentors at Harvard who convinced me to apply to graduate school in the first place, and who wrote the letters of recommendation that made this experience possible: Edo Airoldi, Joe Blitzstein, and Alex D'Amour. I'll never forget the enthusiasm with which Edo offered to guide me on my PhD application process, despite being many years out of school. I want to thank Alex in particular for being one of my role models in research, and I have greatly enjoyed the chance to work together in various capacities over the years.

One of the many benefits of living in Boston has been all the friendships I've gotten to enjoy outside of MIT. To my best friend Tim Chin, I want to thank you for all the sage wisdom, late-night bike rides, spontaneous adventures, and for our Halloween tradition of going to haunted farms. To my college roommates, I'm glad that we've gotten to stay close over the years, and I'm grateful for our shared adventures in Boston and beyond: Ryan Solis, Christian Chauvet, Eric Newcomer, Nima Khavanin, Aidan Shapiro-Leighton, and John Bedell. I also want to thank the various members of my community at Alethia Church, both past and present, including AJ Unander, Chris Chestnut, Levi Kedowide, Matthew Johnson, Myles Olaja, and Nixon Maitre. Finally, in keeping with the stereotype of a nerdy MIT PhD student, I picked up the classic game of Dungeons & Dragons during my PhD. Playing D&D has been a surprising source of camaraderie and friendship, and I want to thank all the friends I've played with in many long-running campaigns during my PhD, including Aditya Mahalingam-Dhingra, Annie Johnson, Ellie Hastings, Jack Montgomery, Jackie Bruleigh, Matthew Johnson, Matthew Teal, Stacy Jankowski, and Tim Wright.

Finally, I want to thank the people who have been closest to me throughout this journey: My girlfriend and partner Erica, and my family. Erica and I met early on during my PhD while volunteering, and she has been an incredible source of love, support, and patience. I'm grateful for all the adventures we've shared, from hiking and skiing all over Montana, kayaking the Charles River in the summer, and visiting family in Taiwan for Chinese New Year. Thank you for keeping me grounded, and I can't wait to see what life has in store for us next.

I've also been blessed to be close to family in Boston during my journey, and I want to thank them for all the support they've given me over the many years of my PhD: My mother (Carla) and father (Tom), my siblings (Sarah, Matthew, and Aaron), my brother-in-law (Aidan), and my niece and nephew (Emma and Oliver). Thank you all for being there during the highs and the lows, and for helping me make it to the end of this journey in one piece.

This thesis is dedicated to my niece Emma and my nephew Oliver, who were born during my PhD. It has been a privilege and a joy to be your uncle, and as you grow older, I hope you'll discover and treasure the joy of learning.

Contents

II Robust prediction via causal knowledge 211

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and context

What is the role of data in improving healthcare? What role does machine learning, statistics, and routinely-collected data have to play in decision-making for healthcare? There is an increasingly large amount of data available from routine clinical practice, often referred to as "observational" data, to distinguish it from "experimental" data collected in clinical trials. For instance, in the United States, since passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, there has been a dramatic rise in the availability of electronic healthcare data via the adoption of electronic health records (Adler-Milstein and Jha, 2017).

Data collected during routine care has the potential to personalize decision-making (Kanjilal et al., 2020), help triage patients (Cummings et al., 2021), alert clinicians to patients that are at risk of deterioration (Singh et al., 2021; Adams et al., 2022), help automate routine tasks in radiology like quantifying cardiac function (He et al., 2023), and so on. Observational data is also an important complement to data collected during clinical trials, which is often restricted to narrow populations. For instance, Fehrenbacher et al. (2009) found, in a sample of patients with non-small-cell lung cancer (NSCLC) at Kaiser Permanente, that only 34% of patients would have been eligible to participate in the clinical trials that inform current treatment guidelines. Similarly, Travers et al. (2007) found that, across 17 major clinical trials used to inform treatment for asthma, a median of 6% of individuals currently being treated for asthma would have qualified for inclusion. Against this backdrop, the U.S. Food and Drug Administration (FDA) has sought to create pathways for the use of observational data in regulatory approvals (FDA, 2018). For instance, Prograf, a drug designed to prevent organ rejection in transplant patients, was approved for use in lung transplantation based on data from outside of a randomized trial (FDA, 2021), having been initially approved for other transplants based on data from clinical trials.

In this thesis, we consider the use of historical data to generate *predictions* that inform decision-making, whether those predictions are personalized to individuals (e.g., "this patient is likely to need intensive care") or given in broad terms (e.g., "this drug is generally effective for treating a given disease"). The focus of this thesis is on methods for making those predictions more reliable. For our purposes, we use the term "reliability" to capture the idea that predictions should be accurate in the practical contexts in which they are used.

Examples of unreliable predictions in healthcare To illustrate what we mean by "reliability", it is useful to consider examples of unreliable predictions. Today, machine learning and data-driven systems for decision-making are actively used in healthcare, but can fail in unanticipated ways.

As a cautionary example, consider the Epic Sepsis Model, a proprietary prediction model used in hundreds of hospitals to detect sepsis. Sepsis is a deadly syndrome, triggered by infection, that contributes to the death of hundreds of thousands of patients per year (CDC, 2022). A recent study found that this model had far lower performance in practice (0.63 AUC[1]) than originally found by the company (0.76–0.83 AUC), and an alarmingly high rate of false positives (Wong et al., 2021). The

[1] AUC (Area Under the receiver operating characteristic Curve) is a common measure of accuracy, which describes the probability that the model will correctly distinguish between a randomly chosen positive and negative example.

company now recommends that the model be trained on hospital-specific data, and that individual hospitals should conduct their own analyses of performance (Ross, 2022). Meanwhile, the Food and Drug Administration (FDA) recently released guidance indicating that similar models may be subject to increased regulatory scrutiny (FDA, 2022). In part, this under-performance was due to an over-reliance on correlations in training data that were not present during deployment, such as indicators that clinicians had already begun treatment of the condition (Ross, 2021).

In medical imaging, Oakden-Rayner et al. (2020) consider the task of classifying pneumothorax (a collapsed lung) from chest X-rays, and observe on that on a commonly-used dataset, existing metrics of performance are misleading. While the models they examine achieve an AUC of 0.87 overall, this seemingly superb performance masks the fact that 80% of pneumothorax cases in the evaluation dataset contain chest drains, a treatment for the condition. Evaluating the model on these cases alone yields an AUC of 0.94, while evaluating the model on untreated cases (a more realistic setting in a real deployment) yields a far lower AUC of 0.77.

More recently, attempts to apply machine learning during the COVID-19 crisis have met with mixed results. Sun et al. (2022) describe the development and validation of a model that uses chest X-rays to diagnose COVID-19, training on a combination of publicly available datasets and data from the University of Minnesota health system. Their model achieves superb performance (AUC of 0.96) on publicly available datasets, but an AUC of 0.8 on data from the University of Minnesota health system. Meanwhile, when attempting to use the model in other locations, they observe substantial drops in performance, with an AUC of 0.72 and 0.66 at the health systems of Indiana University and Emory University respectively.

In another notable example of using data to inform decision-making, during early March 2020, a small-scale observational study of patients in France suggested that an anti-malarial drug, hydroxychloroquine, was an effective treatment for COVID-19, particularly in combination with azithromycin, an antibiotic (Gautret et al., 2020). At a time when the medical community was scrambling to find effective treatments for COVID-19, this study contributed to a substantial increase in the use of the drug, as well as an Emergency Use Authorization (EUA) by the FDA in late March 2020 (Saag, 2020). Shortly thereafter, large-scale randomized trials such as the Recovery Trial (RECOVERY Collaborative Group et al., 2020) demonstrated no benefit of the drug for hospitalized patients, and larger observational studies established evidence for the poor safety profile of this particular combination of treatments in certain patients (Lane et al., 2020). The FDA withdrew the EUA in June 2020 (FDA, 2020) citing in part the new data provided by large-scale randomized trials.

Understanding why predictions can be unreliable In all of the examples above, the relevant predictions were simply not accurate in practice, despite retrospective analyses that initially suggested promising results. Determining whether or not our predictions are reliable is more complex than it might first appear, because the *context in which a prediction is used* often differs from the data that we have available to us when developing our predictions.

For instance, we may have data from patients treated according to current practice, but we are interested in making predictions regarding what would happen to patients if they were treated differently. Similarly, the data available to train a prediction model may differ from the data seen during practical deployment: In the sepsis and medical imaging examples above, for instance, the training data contained examples where patients had already begun treatment, which are not realistically available during deployment. We provide simple mathematical examples to build more technical intuition for these points in Section 1.2.

In general, the usefulness and reliability of our predictions are *based on assumptions* which relate our historical data to the context in which we expect our predictions to be used. A typical example of an assumption in machine learning, violated in the settings above, is that data is independent and identically distributed (iid), i.e., that our future data is drawn from the same distribution as our training data. The core thrust of this thesis is that we need to be more careful about our assumptions, more rigorous in checking them where possible, and otherwise attempt to develop predictions whose reliability is more robust to violations of our assumptions.

1. **Developing reliable predictions (our focus)**: Before deployment, how do we assess and improve the reliability of our predictions?
   - How can we detect violations of the assumptions that are required for our predictions to be reliable? (Chapters 2 and 3)
   - How can we develop predictions that are reliable under weaker assumptions? (Chapters 4 and 5)
   - How can we assess the sensitivity of our predictions to violations of assumptions? (Chapters 6 and 7)
2. Deployment-time safeguards: How do we prevent inaccurate predictions from causing harm?
3. Ongoing monitoring: How do we detect when our predictions become less accurate over time?
4. Adaptation: How do we adapt our predictions to new contexts or patient populations with limited data?

Figure 1-1: *A (necessarily incomplete) list of broader questions in developing safe and effective machine learning systems. The focus of this thesis is primarily on the first question. Here we use the term "predictions" broadly to include any data-driven recommendation that influences decision-making. Whether those recommendations are disseminated through academic publications (e.g., observational studies that make claims regarding treatment effectiveness) or though computer systems at the point-of-care (e.g., early-detection alerts embedded in the medical record), we want to ensure that our predictions are accurate.*

Our focus in this thesis: How do we make reliable predictions? These broad challenges raise a number of technical questions, some of which are summarized in Figure 1-1. Our focus in this thesis is on proactively assessing and improving the reliability of our predictions before we deploy them broadly, as opposed to e.g., monitoring models in deployment or otherwise developing safeguards to mitigate the impact of unreliable predictions.

Aside: Broader considerations in developing reliable predictions Machine learning and statistical techniques are increasingly used to inform decision-making in sensitive areas, from healthcare to criminal justice and loan approvals. In these contexts, there is a wealth of literature on the potential pitfalls of using machine learning to inform decisions. Without hoping to cover all of the relevant considerations here, we sketch some broader questions of reliability, to set context for our contributions.

Relevant considerations go beyond whether or not our predictions are accurate in the first place. For instance, consider the question of predicting whether or not a particular drug or treatment is effective: In addition to considering the accuracy of these predictions, we might be concerned with issues of fairness or discrimination — for instance, are our predictions systematically wrong for individuals of certain racial groups? Moreover, if our predictions directly inform decisions (e.g., suppose that an insurance company denies coverage due to predicted ineffectiveness), do individuals who are impacted have access to appropriate recourse to both understand why those decisions were made, and contest them if they are based on faulty information? When constructing our predictions, are we ensuring that the privacy of individuals is maintained? We refer interested readers to the recently released Blueprint for an AI Bill of Rights, released by the White House Office of Science Technology and Policy (OSTP, 2022) for more context along these lines.

Nonetheless, in this thesis we will focus on the narrower question of whether or not our predictions are accurate in the first place, as a necessary (but not sufficient) condition for the effective use of predictions in decision-making.

1.2 Background: Causality and Dataset Shift

On a technical level, many of our contributions in this thesis revolve around related ideas from causal inference (Pearl, 2009; Imbens and Rubin, 2015; Hernan and Robbins, 2019; Peters et al., 2017) and dataset shift (Quiñonero-Candela et al., 2009). We begin by introducing two stylized examples in Section 1.2.1 of using data to develop predictions of patient outcomes under treatment and to develop predictions of whether or not a patient has a particular disease. In both cases, our predictions are reliable if future data is similar to our historical data. In Section 1.2.2 we demonstrate how things can go wrong if this assumption is violated, and in Section 1.2.3 we describe alternative assumptions, based on knowledge of the causal data generating process, that allow us to make more reliable predictions in each setting.

1.2.1 Simple motivating examples

To start building more technical intuition for the contributions of this thesis, we will start by introducing some basic notation and stylized examples. Let X denote the variables that we use to make predictions, let Y be the variable we are trying to predict (e.g., an outcome under treatment, or a whether or not a patient has some disease) where $f(X)$ is some function takes X as input, and returns a prediction $\hat{Y}$.

A typical assumption made in standard machine learning tasks is that the distribution of X, Y is the same between our *training* distribution $P_{\text{tr}}(X, Y)$, from which we sample the data used to train $f(X)$, and the *test* distribution $P_{\text{te}}(X, Y)$, from which we sample the data used to evaluate $f(X)$.

Assumption 1.1. The training and test distributions are the same, i.e.,

$$P_{\text{tr}}(X, Y) = P_{\text{te}}(X, Y) \tag{1.1}$$

When Assumption 1.1 holds, the performance (e.g., accuracy) of our predictions on the training distribution will be (roughly speaking) similar to the performance of our predictions on the test distribution. When Assumption 1.1 does *not* hold, such that the test data is distributed differently than the training data, this scenario is often referred to as *dataset shift* or *distribution shift* (Quiñonero-Candela et al., 2009). We introduce two stylized motivating examples here, to illustrate where Assumption 1.1 breaks down, and the ways in which considering the *causal data generating process* can help us to make more reliable predictions.

Example 1.1 (Assessing treatment effectiveness)**.** Suppose that we are trying to assess whether or not some treatment is effective at increasing survival rates among patients. If we could predict what would happen for future patients, if they were to receive the treatment, then we could use such predictions to guide treatment decisions. In Table 1.1 we lay out some illustrative data, where we observe that in historical data, treated patients have worse outcomes than patients who go untreated.

Table 1.1: *Data for Example 1.1, where patients in the control group have higher survival rates than patients in the treated group.*

	Rate of Survival	% of Population
Control	74%	50%
Treatment	36%	50%

Example 1.2 (Diagnosing disease)**.** Suppose that we are interested in predicting the likelihood that patients have a given disease, where we only have access to a single laboratory test, which is not ordered for all patients. In Table 1.2 we lay out some illustrative data, noting a strong correlation between disease prevalence and whether or not a laboratory test has been ordered.[2] Among patients without laboratory tests, the prevalence of disease is 30%, while the overall prevalence of disease is 55%.

Table 1.2: *Distribution of outcomes among patients who do or do not have laboratory tests ordered, in Example 1.2, along with the proportion of patients who have laboratory tests ordered. The overall prevalence of disease is 55%.*

	Rate of Disease	% of Population
Test ordered	80%	50%
No test ordered	30%	50%

If we view both examples as prediction tasks, a naive prediction model would rely entirely on these observed correlations in data — an estimated likelihood of survival under treatment of 36% in Example 1.1, and an estimated likelihood of having disease

[2] It has been observed empirically in medical data that the presence or absence of a laboratory test is itself highly predictive of disease (Agniel et al., 2018), and we credit Subbaswamy et al. (2019) for first introducing us to this type of example in causal graphical form.

(if no lab test is ordered) of 30% in Example 1.1.[3]

In both of these examples, if future data resembles our historical data, then there is no problem from a prediction perspective. But for different reasons, in both cases, these predictions might not be reliable, due to violations of this assumption.

1.2.2 Complications due to violations of assumptions

We now describe some simple complications that could arise in both examples, highlighting the importance of understanding the underlying data generating process.

Example 1.1 (Continued)**.** Survival rates in the control group are 74%, versus 36% in the treatment group, for an overall survival rate of 55%. If we treat these numbers as predictions of survival rates *if we were to treat all patients the same way*, what might happen? In this example, if clinicians stopped treating patients entirely, we would see the survival rate drop from 55% to 50%. If clinicians choose to treat all patients, we would see the survival rate increase from 55% to 60%. Our predicted chance of survival without treatment (74%) and with treatment (36%) was misleading, but why?

Table 1.3: *Data for Example 1.1, where patients in the control group have higher survival rates than patients in the treated group. When we stratify by patient complexity (e.g., how sick the patient is to start with), we observe that the treatment is associated with higher survival rates in both groups of patients. This phenomenon is often referred to as "Simpson's Paradox" in introductions to causal inference (see Section 6.1 of Pearl (2009)), and used to illustrate the fact that correlation is not necessarily equal to causation.*

	Overall	Complicated		Normal	
	% Survival	% Survival	% of pop.	% Survival	% of pop.
Control	74%	20%	5%	80%	45%
Treatment	36%	30%	45%	90%	5%

In Table 1.3, we break out performance by patient complexity (C), and observe that within each group of patients, the treatment (T) improves survival rates (Y).

[3] For simplicity, we ignore the task of prediction when a lab test is available, as this would require us to specify the distribution of e.g., lab values given disease state. Focusing on predictions without a laboratory test is sufficient to illustrate our point here.

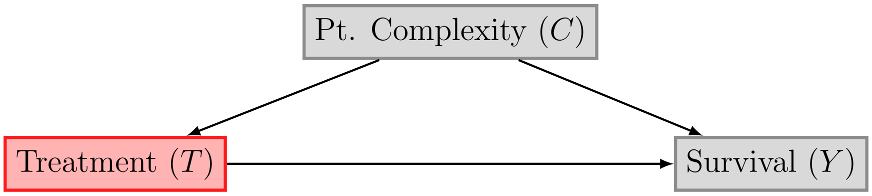


Figure 1-2: *Causal graph for Example 1.1. A causal graph lays out the causal relationships between variables: In this example, patient complexity influences the decision to provide treatment, as sicker patients are more likely to be treated. Similarly, patient complexity influences the outcome of survival, as more complex patients have lower survival rates. In this example, patient complexity is a "confounder", a variable we must adjust for in order to draw valid causal conclusions about the influence of treatment on survival rates.*

In Figure 1-2 we provide an illustrative *causal graph* which describes the underlying data-generating process, and reveals the problem: In our training distribution, the probability of treatment $P_{\text{tr}}(T \mid C)$ depends on patient complexity, where more complex patients are more likely to receive treatment. If we alter the treatment policy (e.g., by treating all patients the same way), this change would imply a violation of Assumption 1.1, the assumption that $P_{\text{tr}} = P_{\text{te}}$.

Example 1.2 (Continued)**.** Disease prevalence in the group of patients who do not receive laboratory tests is 30%. If we use this statistic as our predicted probability of disease for these patients going forward, what might happen? Suppose we apply our predictions in a different hospital with the same distribution of patients, and observe that our predictions are now systematically miss-calibrated: Patients who do not receive laboratory tests now have a 55% prevalence of disease. We show the data from the second hypothetical hospital in Table 1.4. The mix of patients is the same in both hospitals (i.e., the overall disease rate is the same), so what might have happened?

We show the corresponding causal graph Figure 1-3, which illustrates the problem. The difference between the two hospitals is that they have different *laboratory testing policies*. In the first, the probability of ordering a test $P_{\text{tr}}(O \mid Y)$ depends on disease (e.g., due to symptoms of disease), while in the second, laboratory tests are ordered randomly. As a result, the distributions differ, violating Assumption 1.1.

Table 1.4: *Data for Example 1.2, where we imagine deploying our predictions in a new hospital with the same prevalence of disease (55%) but where patients are randomly selected to receive laboratory tests.*

	Original Hospital		New Hospital	
	% Disease	% of pop.	% Disease	% of pop.
Test ordered	80%	50%	55%	50%
No test ordered	30%	50%	55%	50%

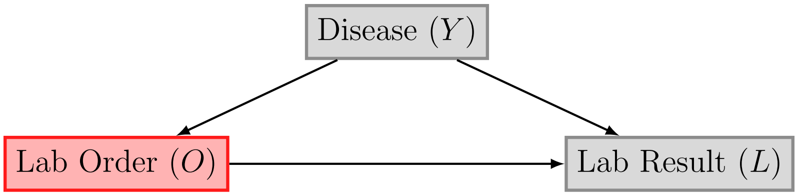


Figure 1-3: *Causal graph for Example 1.2. The choice to order a laboratory test is correlated with whether or not someone has disease, perhaps due to unrecorded symptoms. Patients who receive laboratory tests have a higher pre-test probability of having disease (see Table 1.2).*

Notably, this is not the only way in which these hospitals could differ in their testing policies. A particularly adversarial change would be for clinicians to only order tests for healthy patients, and never order tests for sick patients, in which case the "correct" prediction would be 100% likelihood of disease for those not tested.

1.2.3 Using causal knowledge to develop more reliable conclusions

In both examples, knowledge of the causal data generating process, e.g., the causal graphs given in Figures 1-2 and 1-3, can be used to help us develop more reliable predictions. However, knowledge of the causal data generating process is itself an assumption: In both cases, we replace our assumption that $P_{\text{tr}} = P_{\text{te}}$ with a more realistic assumption on the process that generates the data.

A central aim of causal reasoning is to provide answers to "what if" questions, where understanding the causal data generating process allows us to reason about how different real-world scenarios would imply different distributions over the data: For instance, how outcomes would change under different treatment policies in Example 1.1, or how disease rates would change for untested patients in Example 1.2.

Example 1.1 (Continued)**.** Suppose that the causal graph in Figure 1-2 is valid, e.g., there are no other "confounding" variables that influence both the treatment decision and the outcome. Under this graph, we can estimate the average outcome if we were to treat all patients by using *do-calculus* to construct an adjusted estimate (Pearl, 2009)

$$\mathbb{E}[Y \mid do(T=1)] = \sum_{c} \mathbb{E}[Y \mid T=1, C=c] P(C=c) \tag{1.2}$$

Example 1.2 (Continued)**.** In Example 1.2, our primary goal is not to choose a treatment T, but merely to predict the outcome Y with high accuracy. However, the failure of Assumption 1.1 means that we cannot simply try to minimize error on the training distribution as a route to minimizing error on the test distribution. However, we can use knowledge of the causal data generating process to conclude that changes in clinical practice would be reflected in changes to the conditional distribution $P(O \mid Y)$, keeping constant the other factors. For instance, if Figure 1-2 holds, and if we knew the laboratory testing policy at a new hospital $P_{\text{te}}(O \mid Y)$, then we could estimate the prevalence of disease in the untested population using a simple adjustment

$$\mathbb{E}_{te}[Y \mid O=0] = \frac{P_{\text{te}}(O=0 \mid Y=1) P_{\text{tr}}(Y=1)}{\sum_y P_{\text{te}}(O=0 \mid Y=y) P_{\text{tr}}(Y=y)} \tag{1.3}$$

However, this adjustment requires us to know the laboratory testing policy at a new location. When this policy is unknown, we may prefer to develop a model whose performance is "robust" under different possible policies.

Conclusion: In this thesis, we distinguish between improving the reliability of causal inference, and improving the robustness of prediction models. In both cases, understanding the data generating process will be essential, and adopting a causal perspective gives us tools to do so across both types of problems. However, the particular problems that we address will vary subtly across the two types of applications. When considering causal questions, our focus will be more on *detecting violations of our (causal) assumptions* and *developing predictions that are reliable under weaker* *assumptions* (the first and second question in Figure 1-1). When considering prediction problems, we focus on how causal frameworks can be used to *develop prediction models that have reliable performance even when dataset shift occurs (and Assumption 1.1 fails to hold)* and *assess the sensitivity* of our prediction models to plausible changes in distribution (the second and third question in Figure 1-1).

1.3 Structure and overview of this thesis

In this thesis, we attempt to provide a partial answer to the question:

> *How can we attempt to rigorously stress-test our models, and the conclusions we draw from them? For models and conclusions that do not pass these tests, what can we do to make them more robust?*

We approach this task in two parts, first focusing on improving the reliability of causal conclusions, and then focusing on improving the reliability and robustness of prediction models. Several chapters in this thesis first appeared as published papers in conferences, though this thesis does not cover all of the work that I've done during my PhD. In Table 1.5 we give an overview of the papers included in this thesis, and the chapters that correspond to those papers.

Table 1.5: *Publications and preprints written during my PhD, and their appearance (or lack thereof) in this thesis. Co-first-authorship denoted by *.*

Publication	Venue	Inclusion in Thesis
(Oberst and Sontag, 2019)	ICML 2019	Chapter 2
(Oberst et al., 2020)	AISTATS 2020	Chapter 3
(Hussain et al., 2022)*	NeurIPS 2022	Chapter 4
(Oberst et al., 2021a)	ICML 2021	Chapter 5
(Thams et al., 2022)*	NeurIPS 2022	Chapter 6
(Hussain et al., 2023)	AISTATS 2023	Not included
(Oberst et al., 2022)	Preprint, presented at ACIC 2022	Not included
(Lim et al., 2021)*	NeurIPS 2021	Not included
(Ji et al., 2021)*	AMIA Informatics Summit 2021	Not included
(Boominathan et al., 2020)	KDD 2020	Not included
(Kanjilal et al., 2020)	Science Translational Medicine 2020	Not included

1.4 Part I: Reliable causal inference and policy evaluation

1.4.1 Overview of our perspective

Learning and evaluating new treatment policies from retrospective (or "observational") data requires causal assumptions, such as full observation of confounding factors. These assumptions typically form the core way that "domain knowledge" is injected into the process of causal reasoning, in a process known as *identification*, shown in Figure 1-4. The reliability of our conclusions depends fundamentally on whether or not these (typically untestable) assumptions hold in practice.

In this part of the thesis, we introduce methods for attempting to find flaws in causal models, by incorporating other forms of domain knowledge outside of standard causal assumptions.

In Chapter 2, previously published as our Masters thesis (Oberst, 2019), we introduce a method for helping clinicians review and contest the causal claims of models for sequential decision-making. We demonstrate the utility of this approach for finding flaws in published work on sepsis treatment recommendation. However, this approach is necessarily exploratory and hypothesis-generating in nature. In Chapter 3 and Chap-

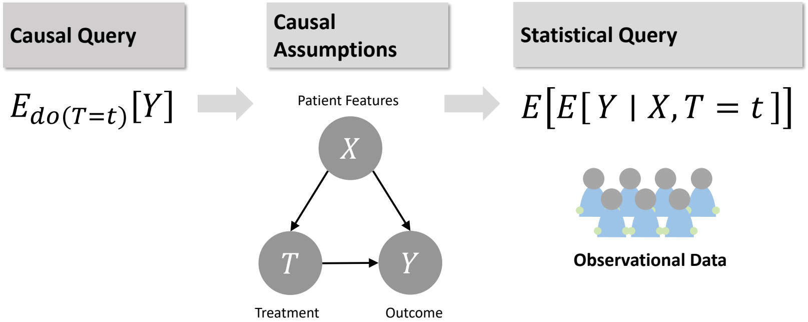
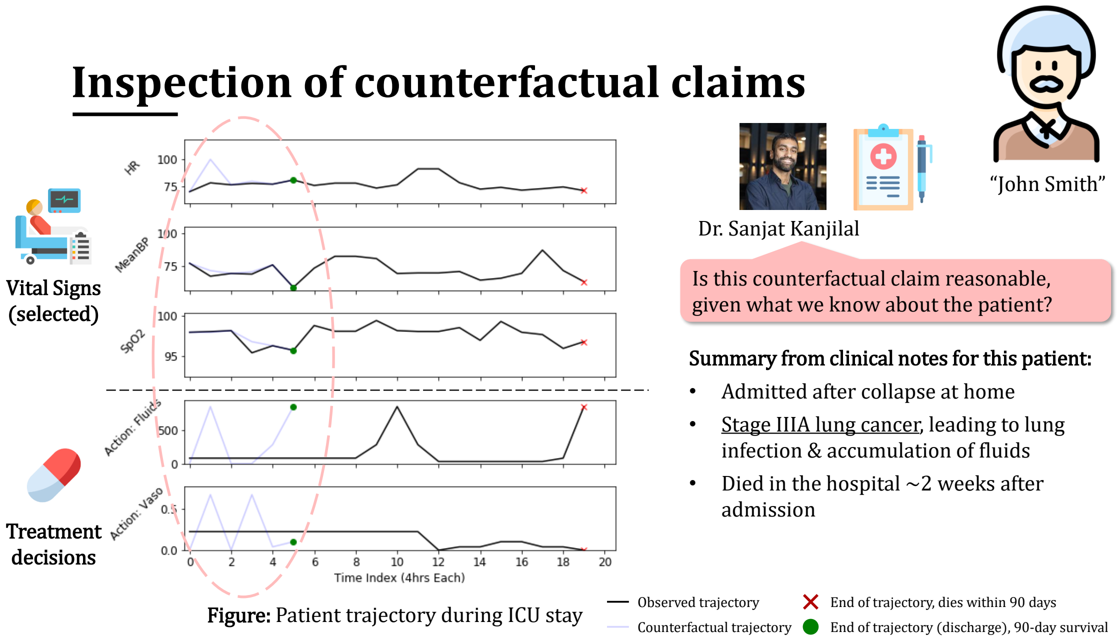


Figure 1-4: *The typical process of performing causal inference: Starting from a causal query (e.g., "what would the average outcome be if all patients were treated?"), a set of causal assumptions (e.g., on the causal relationships between variables, expressed as a causal graph) allows for translation of this causal query into a statistical query that can be estimated from observational data. The reliability of our conclusions depends in large part on whether or not these causal assumptions hold in practice, but these assumptions are not typically testable. In this part of the thesis, we ask "what else can we do to sanity-check our conclusions"?*

ter 4 we provide two alternative perspectives on searching for flaws: In Chapter 3 we give a method, using interpretable boolean rule sets, for characterizing violations of one of the few testable assumptions in causal inference, the assumption of overlap between treated and control populations. In Chapter 4, we give a method for incorporating experimental data (e.g., from a randomized trial), which yields valid conclusions under less stringent assumptions, targeting scenarios where experimental data does not cover the population of interest. This scenario commonly occurs when considering off-label use of drugs, or when attempting to generalize the results of randomized trials to populations not originally eligible for the trial.

1.4.2 Chapter 2: Counterfactual Policy Introspection

How should clinicians evaluate the claim that a new treatment policy will improve outcomes? In this chapter, we develop a technique to help clinicians validate probabilis-



Inspection of counterfactual claims

Is this counterfactual claim reasonable, given what we know about the patient?

Summary from clinical notes for this patient:

- Admitted after collapse at home
- Stage IIIA lung cancer, leading to lung infection & accumulation of fluids
- Died in the hospital ~2 weeks after admission

Figure: Patient trajectory during ICU stay

Figure 1-5: *Counterfactual Policy Introspection. Here we show a real patient from the MIMIC-III dataset (Johnson et al., 2016), evaluating the counterfactual claims of a probabilistic model developed in (Komorowski et al., 2018), highlighting a patient who the model claims would have lived under an alternative treatment policy. Extracting these kind of counterfactual claims is a primary technical contribution of Oberst and Sontag (2019). In black is the actual trajectory of this patient, and in light blue we see a counterfactual claim for how the vital signs (and eventual outcome) would have differed under alternative treatment, taking into account what actually happened to this patient. These claims are not clinically plausible, suggesting flaws in both the original model, and the model-based claim that 95% of patients would survive under the new policy. These counterfactual claims, under mild conditions, represent an attribution of the claimed improvement in outcomes to specific historical patients like this one.*

tic causal models used in sequential decision-making problems (e.g., in model-based reinforcement learning for sepsis treatment). Our approach *decomposes aggregate claims* made by such models (e.g., "80% of hypothetical future patients would survive under the new policy") into *counterfactual claims* on specific historical patients (e.g., "patient X would have survived under the new policy").

To uncover flaws in the clinical reasoning of the model, these counterfactuals can be reviewed by clinicians alongside the full medical record for those patients. An example is shown in Figure 1-4. We use this approach to uncover flaws in a highly-cited paper, the "AI Clinician" (Komorowski et al., 2018): Alongside an infectious disease clinician from Mass General Brigham, we review patients who died in reality, but who the model claimed *would have lived* under its recommended policy. Implausible counterfactual claims were often clearly attributable to confounding factors not included in the model (e.g., terminal cancer), but present in clinical notes. Such insights are directly relevant for improving model design (e.g., by extracting additional features to include as potential confounders).

The main technical innovation is to derive counterfactual claims from any existing model of discrete dynamics. This presents a problem: Counterfactual simulation requires specification of a structural causal model (SCM), but there are many SCMs that are consistent with the original model. Here, any SCM will produce a valid decomposition, but some decompositions are more interpretable than others. To this end, we introduce a condition called "counterfactual stability" that imposes common-sense restrictions on counterfactuals, and introduce a novel SCM that satisfies this condition. This condition generalizes the monotonicity condition of Pearl (1999) from binary to categorical variables (necessary for inferring counterfactual dynamics in models with discrete states). Conditions like these encode the intuition that, e.g., if we observe an increase in blood pressure in the absence of treatment, then we should also see a counterfactual increase in blood pressure if given a blood-pressure-increasing medication. We prove that naive extensions of SCMs that satisfy monotonicity (for binary outcomes) do not satisfy counterfactual stability (for categorical outcomes). To resolve this mismatch, we propose a SCM based on the Gumbel-Max trick, and prove that it does satisfy counterfactual stability. Notably, the research discussed in this chapter has spurred interest from other research groups, leading to follow-up research on alternative counterfactual restrictions (Lorberbom et al., 2021) and use cases for counterfactual simulation (Corvelo Benz and Gomez Rodriguez, 2022).

However, this approach is fundamentally exploratory in nature, requiring the review of individual patients and their counterfactual predictions, and is best understood as a "sanity checking" procedure. In the remaining sections, we discuss alternative approaches which either test those assumptions that are directly testable, or use experimental data to further improve the credibility of causal conclusions drawn from observational data.

1.4.3 Chapter 3: Characterization of Overlap in Observational Studies

How can clinicians tell if the conclusions of a causal analysis apply to a particular patient? A necessary (and testable) condition is that similar patients were observed receiving both the treatment and control. In this chapter, we give an algorithm (OverRule) for creating interpretable descriptions of the well-represented population, which could then be published alongside a retrospective study. The overall goal is demonstrated in Figure 1-6.

In particular, we demonstrate that the problem could be reduced to repeated Neyman-Pearson classification with Boolean rule sets. This method was developed with a clinical collaborator from Beth Israel Deaconess Medical Center, inspired by applications in estimating the effect of post-surgical opioid prescriptions on future misuse, using health insurance claims data. The resulting output was evaluated in user studies with a small group of clinicians, and found to represent plausible clinical patterns. For instance, large opioid doses are rarely prescribed for C-section surgeries, and hence we cannot reliably infer causal effects of large vs. small doses in this population.

1.4.4 Chapter 4: Falsification before Extrapolation in Casual Effect Estimation

Experimental data (i.e., from a clinical trial) is often small-scale and narrow in scope. For instance, Phase 3 clinical trials for COVID vaccines did not originally include pregnant women (Dagan et al., 2021). To assess causal effects in these unrepresented populations, we often turn to observational data. In this chapter, we demonstrate that the experimental data is still useful, despite not covering the population of interest.

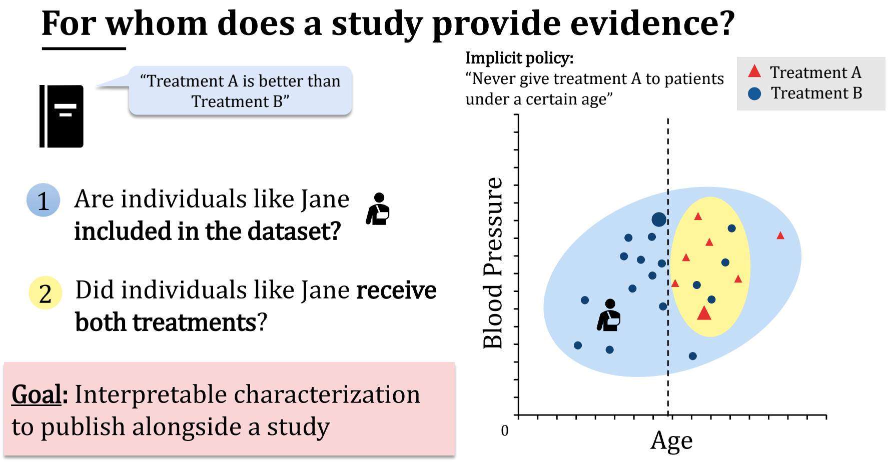

For whom does a study provide evidence?
"Treatment A is better than Treatment B"
1 Are individuals like Jane included in the dataset?
2 Did individuals like Jane receive both treatments?
Goal: Interpretable characterization to publish alongside a study
Implicit policy:
"Never give treatment A to patients under a certain age"
▲ Treatment A
● Treatment B
Blood Pressure
0
Age

Figure 1-6: *When considering a particular patient "Jane Doe", when can we conclude whether or not the results of a particular observational study are applicable? A necessary condition is that patients like Jane are not only included in the study itself, but have some positive probability of being included in both the treatment and control groups.*

In particular, we develop a method that can be applied when multiple observational studies cover both the population of interest and the experimental subpopulation. This scenario arises in two different contexts: First, when there are multiple observational datasets being used to perform a causal analysis. This first setting typically arises in the context of network studies, where data is separately analyzed at different hospital sites for reasons of data privacy, and the results communicated back to a central analyst for aggregation. Second, the multiple studies could correspond to different analyses of the same dataset under different sets of causal assumptions, e.g., controlling for different sets of potential confounding factors.

The core idea is to use the experimental data to test for potential bias, shown in Figure 1-7, and then conservatively aggregate estimates across observational studies that pass this test. This is a form of meta-analysis (the analysis of multiple studies) that comes with guarantees under weaker assumptions than standard meta-analysis assumptions. Instead of requiring that all studies are unbiased (e.g., free of confounding), this approach only requires that at least one observational study is unbiased, a relaxation of standard assumptions in meta-analysis that require all studies to be unbiased.

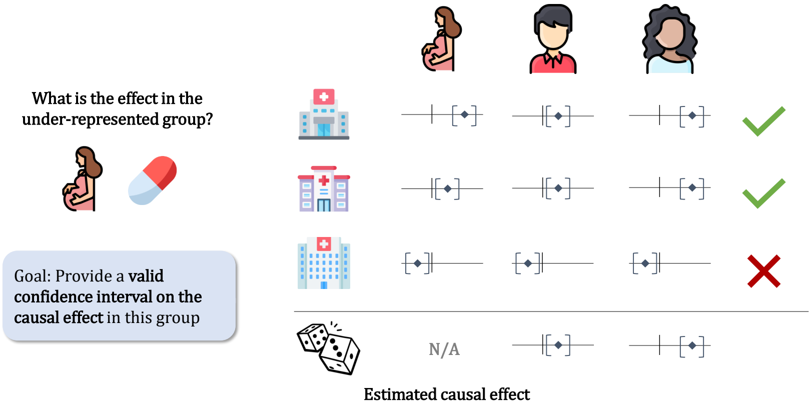

What is the effect in the under-represented group?
Goal: Provide a valid confidence interval on the causal effect in this group
N/A
Estimated causal effect

Figure 1-7: *Combining evidence from multiple observational studies: The first stage of our approach takes advantage of overlapping populations between observational studies and randomized trials (the center and right-most subgroups) to assess whether or not the effect estimates from the observational studies are comparable to those found in the randomized trial. The second stage of our approach (not shown here) conservatively combines the studies that pass this test, to construct confidence intervals on the causal effect that are valid so long as (at least) one of the original observational studies provides valid confidence intervals.*

1.5 Part II: Robust prediction via causal knowledge

1.5.1 Overview of our perspective

Predictive models can fail due to unreliable correlations that change across hospitals or patient populations. In this part of the thesis, I present methods for techniques for *anticipating and avoiding these failures in advance.* We focus on the *proactive* setting, where we only have access to data from the training distribution. In this setting, partial causal knowledge allows us to reason about performance of predictive models in unseen future scenarios.

Building reliable but effective models requires trade-offs. For instance, many predictive models in healthcare rely on "operational" signals, data that reflects decisions made in routine clinical practice. Example 1.2 is a simple example of this kind of signal, corresponding to whether or not a laboratory test is ordered. Changes in clinical practice could impact the correlation between these features and disease. A drastic approach to learning a reliable model in Example 1.2 would discard lab-related features altogether, at the cost of lower predictive performance. The methods I discuss in this thesis allow for *more principled trade-offs* between reliability and effectiveness, by considering *worst-case performance under plausible changes.*

One way to formalize this problem is that we have data from the training distribution, but we care about performance on an unknown distribution Q that likely differs from the training distribution P. An idealized objective is to learn a model that minimizes

$$\mathbb{E}_Q[\ell(f(X), Y)] \tag{1.4}$$

However, we do not have access to Q. Instead, we have limited information about Q, which can be expressed in several forms: We focus on the case where we only have data from P, but we have assumptions that restrict the form of Q. For instance, consider the laboratory testing example, where we have it that

$$Q(Y, O, L) = P(Y)Q(O \mid Y)P(L \mid O, Y) \tag{1.5}$$

Here, we've restricted Q to only differ from P along in the conditional distribution $Q(O \mid Y)$, but have left that distribution unrestricted so far. It is straightforward to show that $Q(Y \mid O, L) \neq P(Y \mid O, L)$ under this shift. Given these assumptions, the analyst in this running example may ask

> *Should I use laboratory testing in my model, given that the correlation between ordering a test and the presence of disease is potentially unstable?*

That is, the analyst has at least one simple choice to make, when trying to build a model with "robust" performance: They can train a model $f(O, L) \approx P(Y = 1 \mid O, L)$ using standard techniques, or they can use only "stable" correlations, which in this case corresponds to simply predicting using the base rate $P(Y = 1) = Q(Y = 1)$, which does not change. There are other choices that the analyst could make in this simple case. They could also, for instance, re-weight the training data to break the existing correlation between Y and O (Subbaswamy et al., 2019). We focus on the two ideas above for simplicity.

The correct modelling choice relies on not just specifying *what* can change (i.e., $Q(O \mid Y)$), but also specifying *how much* it can change, and translating that knowledge into a quantitative comparison between modelling choices. We adopt a *worst-case* perspective here, specifying a set $\mathcal{Q}$ of possible distributions, and considering the worst-case performance of our models under that set of distributions.

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell(f(X), Y)] \tag{1.6}$$

where $\ell(f(X), Y)$ is some loss function. Choosing a realistic set $\mathcal{Q}$ is essential to making effective trade-offs between in-distribution accuracy and reliability. Informally, if $\mathcal{Q}$ consists of P and distributions close to it, the minimizer of the worst-case loss may be very close to the model that simply minimizes $\mathbb{E}_P[\ell(f(X), Y)]$. On the other hand, if $\mathcal{Q}$ includes distributions far away from P, then the minimizer of the worst-case loss may perform very poorly on P itself.

In Figure 1-8, we give an illustration taken from Chapter 6, where we increase and decrease $P(O \mid Y)$, observing the impact on model performance. In this case, the drastic approach of throwing away laboratory testing information in fact yields better performance if testing rates go to zero: For most reasonable changes in testing rates, however, the model $f(O, L)$ is superior.[4]

This example also illustrates the role of causality in mapping a real-world change (here, a change in laboratory testing policies) to a specification of which factors that shift. Under the causal graph given in Figure 1-3, $P(O \mid Y)$ is the only factor that would change under a new policy. We could consider other, "non-causal" changes, but their interpretation is less clear. For instance, if we simply considered a change in $P(O)$, with $P(Y, L \mid O)$ kept fixed, then this would lead to a change in the marginal

[4] The details of this example, e.g., the full data-generating process, which differs slightly from the one presented in Example 1.2 can be found in Appendix D.6.1.

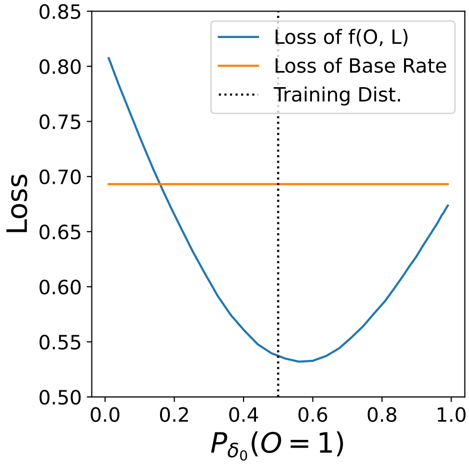


Figure 1-8: *Performance of two models in the lab testing example, as the marginal testing rate changes.*

rates of disease Y in the new distribution.

This simple example demonstrates a more broadly applicable insight, which extends readily to more complex machine learning tasks. For instance, in many computer vision tasks, there is a wide range of meta-data available, and a variety of approaches that either pre-process the data, or alter the training loss, to learn models that selectively ignore certain correlations in the data. This is the analog, in our setting, of learning a model that does not incorporate laboratory testing.

With this in mind, we tackle three challenges in this part of the thesis: First, how do we translate domain knowledge into more precise characterizations of shift? Second, how do we learn models that are adapted to those shifts? Third, how do we incorporate additional information, in the form of data from the target distribution, to further improve performance?

In Chapter 5 we discuss work on distributional robustness in linear models that covers all three of these goals, but under somewhat restrictive generative assumptions. Meanwhile, in Chapter 6 we discuss work that focuses on these first two points in more general settings, showing how to estimate the worst-case loss for a fixed model over a precisely-controlled set of possible distributions, which can in turn be used to choose among different modelling approaches. Finally, in Chapter 7, we conclude with a brief vignette about applying these ideas to issues of model design in language-vision models, using the techniques in Chapter 6 to understand the worst-case performance of different prompting strategies for developing zero-shot classifiers.

Before introducing our work in more detail, we review two lines of existing thought in the literature, and how our work fits in. We contrast the approach we take in this part of the thesis with two strains of work in the literature. First, we discuss causality-motivated approaches to learning robust prediction models. Here, our core premise is that these approaches are not always applicable in the scenarios we consider, and that even when they can be applied, they are overly conservative. Second, we discuss related work in distributional robustness, where objectives such as Equation (1.6) are commonly considered. Here, our core premise is that many existing methods for defining the set $\mathcal{Q}$ are difficult to interpret, and do not allow for incorporating fine-grained knowledge of plausible shifts.

Causality-motivated methods for learning robust models: Several approaches seek to learn models that perform well under arbitrarily large causal interventions (which result in arbitrary changes in selected conditional distributions). Several approaches proactively specify shifting mechanisms/conditional distributions, and then seek to learn predictors that have good performance under arbitrarily large changes in these mechanisms (Subbaswamy et al., 2019; Veitch et al., 2021; Makar et al., 2022; Puli et al., 2022). Other approaches use auxiliary information, such as environments (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2019) or identity indicators (Heinze-Deml and Meinshausen, 2021) to learn models that rely on invariant conditional distributions. For instance, invariant risk minimization (IRM) and related approaches seek a predictor Φ such that $\mathbb{E}(Y \mid \Phi(X))$ is invariant across a set of discrete environments (Arjovsky et al., 2019; Xie et al., 2020; Krueger et al., 2020; Bellot and van der Schaar, 2020). However, recent work has pointed to the theoretical and practical difficulty of learning invariant predictors in the sense of IRM (Rosenfeld and Risteski, 2020; Kamath et al., 2021; Guo et al., 2021), in part due to the fact that recovering a truly invariant model, even in linear settings, requires a large number of environments. Generalization in non-linear settings requires sufficient overlap between environments and strong restrictions on the model class (e.g., Christiansen et al., 2020). Finally, even when it is possible to successfully apply these approaches, their worst-case optimality is often restricted to cases where the shifts are arbitrarily large.

When the causal interventions (i.e., changes in causal mechanisms) are bounded (i.e., not arbitrary), then these approaches are not necessarily optimal. Closest to our work in motivation is prior work on robustness to bounded shift interventions in linear causal models (Rothenhäusler et al., 2021), which we build upon in Oberst et al. (2021a) (Chapter 5). Moreover, Thams et al. (2022) (Chapter 6) can be seen as extending the ideas of Rothenhäusler et al. (2021) to general non-linear causal models, starting with the task of evaluating the worst-case loss itself.

Distributionally robust optimization/evaluation with divergence measures: Distributionally robust optimization (DRO) seeks to learn models that minimize objectives of the form of Equation (1.6) (Duchi and Namkoong, 2021; Duchi et al., 2020a; Sagawa et al., 2020). The major difference between our work and prior work lies in the definition of the set of plausible future distributions $\mathcal{Q}$, often called an "uncertainty set" in the optimization literature, where the goal is to specify a set that captures expected shifts, without being overly conservative.

Shifts in $P(X, Y)$: A conservative approach is to include all joint distributions $P(X, Y)$ within a certain neighborhood of the training distribution. Many coherent risk measures can be written as a worst-case loss of this form. For instance, the Entropic Value-at-Risk (EVaR), with confidence level $1 - \alpha$, corresponds to the worst-case loss over a set of distributions $\mathcal{P} = \{P \ll P_0 : D_{KL}(P \| P_0) \leq -\ln \alpha\}$, where P_0 is the original distribution (Ahmadi-Javid, 2012). Similarly, the Conditional Value-at-Risk (CVaR) with parameter α can be seen as the worst-case loss over an uncertainty set obtained from a limiting f-divergence (see Example 3 of Duchi and Namkoong (2021)), including all α-fractions of the original distribution. These measures are appealing, in that they are straightforward to compute, but can be very conservative. Indeed, such measures often reduce to only considering the distribution of the loss itself. CVaR, for instance, is equivalent to sorting the training examples by their loss, and taking the average loss of the top α-fraction.

Shifts in $P(X)$ alone: Partially due to this overly-conservative behavior, there has been a line of work incorporating additional restrictions on the allowable shift (i.e., adding more assumptions). For instance, Duchi et al. (2020a) considers learning predictive models that optimize a worst-case loss similar to CVaR (a "worst-case subpopulation shift"), but where only $P(X)$ is allowed to change, and $P(Y \mid X)$ is assumed to be constant. However, many real-world shifts do not fit this framework: In Example 1.2, for instance, both $P(X)$ and $P(Y \mid X)$ are changing, where $X = (O, L)$, as a result of a shift in $P(O \mid Y)$.

Shifts in a conditional distribution: The work we introduce in Chapter 6 is closest to Subbaswamy et al. (2021), who consider evaluating the loss under worst-case changes in a conditional distribution, but while we consider parametric shifts, they estimates the loss under worst-case $(1 - \alpha)$ conditional subpopulation shifts. However, it is not obvious how to choose an appropriate level of α: in some settings, seemingly plausible values of α correspond to entirely implausible shifts. In Chapter 6, we give a simple lab-testing example, where the worst-case 20% subpopulation is one where healthy patients are always tested, and sick patients never tested.

In contrast to these methods, the approaches we outline in this section use explicit parametric perturbations to define shifts, as opposed to distributional distances or subpopulations. We now discuss our contributions in each chapter in more technical detail.

1.5.2 Chapter 5: Regularizing towards Causal Invariance: Linear Models with Proxies

In this chapter we give methods for learning linear models with minimal worst-case performance over a set of distributions. In particular, we consider a set of distributions that arise due to causal interventions on an "anchor" variable in an underlying linear

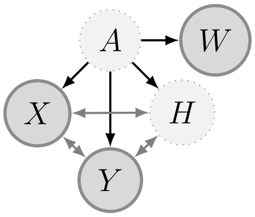


Figure 1-9: *Linear SCM, where variables are linear functions (plus additive noise) of their parents. A is assumed to be observed in prior work, while W is a proxy for the unobserved A in ours.*

structural causal model (SCM). Here, we extend prior work (Rothenhäusler et al., 2021) to allow for specifying both asymmetric changes in distribution (tailored to existing domain knowledge) and changes in distribution that influence unobserved variables.

We use a theoretical model from prior work (Rothenhäusler et al., 2021), given in Figure 1-9, where $A \in \mathbb{R}^{d_A}$ represents a variable whose distribution may change, and which has some (unknown) causal relationship to X, Y, and potentially other hidden variables H. In the simplest case, A may encode discrete environments, but can more generally encode other factors of variation, allowing us to consider distributionally robust objectives of the form

$$\min_{\gamma} \sup_{\mathbb{P} \in C_A(\Omega)} \mathbb{E}_{\mathbb{P}}[(Y - \gamma^\top X)^2], \quad \text{where } C_A(\Omega) \coloneqq \{\mathbb{P}(X, Y, do(A \coloneqq \omega)) : \omega \in \Omega\} \quad (1.7)$$

where the set of possible distributions is defined by interventions on A within some uncertainty set Ω. Rothenhäusler et al. (2021) consider sets $\Omega_\lambda = \{\omega \in \mathbb{R}^{d_A} : \omega\omega^\top \preceq (1+\lambda)\mathbb{E}[AA^\top]\}$, where $\lambda \geq -1$ is a hyperparameter, and the intervention is constrained by a rescaling of the covariance of A. They demonstrate that this objective is equivalent to the following, which can be optimized using the training data *when A is observed*

$$\sup_{\mathbb{P} \in C_A(\Omega_\lambda)} \mathbb{E}_{\mathbb{P}}\left[(Y - \gamma^\top X)^2\right] = \mathbb{E}\left[(Y - \gamma^\top X)^2\right] + \lambda \mathbb{E}_A\left[(\mathbb{E}[Y - \gamma^\top X \mid A])^2\right]. \quad (1.8)$$

However, this leaves several challenges unresolved, which we address in this chapter.

Identifying the worst-case loss with unobservable factors A: Suppose that A represents

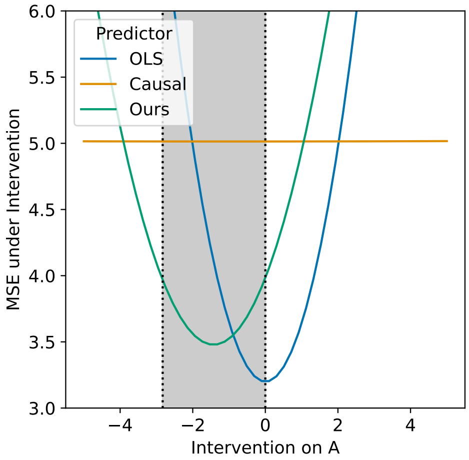


Figure 1-10: *Correctly specifying the robustness set allows for a better trade-off between accuracy and robustness: Here, we plot the MSE (Y-axis) of different models under interventions on A (X-axis). Here, the grey region indicates a user-specified set of plausible interventions on A in the target distribution, and the green line denotes performance of a model trained to minimize worst-case MSE over this set. We show the performance of a standard OLS predictor (blue line) and the invariant casual predictor (orange line) for reference.*

social determinants of health (e.g., income, housing security, etc., which influence both X, Y), and we wish to learn a model that will perform well across different hospitals. Here, we are unlikely to observe A directly, but may instead only have noisy proxies of A (shown as W in Figure 1-9). For instance, we may have self-reported data on income and housing status, as well as third-party data, but lack reliable ground-truth measurements. Here, we demonstrate theoretically that noise in these proxies reduces our worst-case guarantees, if only a single proxy is available. However, we demonstrate that two conditionally independent proxies can be used to recover guarantees as if A were observed, by constructing an objective that is equivalent (in the limit of infinite data) to Equation (1.8), but where proxies W, Z are used in place of A.

Further restricting shifts using domain knowledge: In the same motivating example (A representing socioeconomic factors), we may have additional domain knowledge about plausible shifts in A, such as knowledge that a target hospital generally has lower income patients, rather than higher-income patients. Correctly specifying the robustness set allows for a better trade-off between accuracy and robustness, as shown in Figure 1-10, where we plot the MSE (Y-axis) of different models under interventions on A (X-axis). Here, the grey region indicates plausible interventions on A in the target distribution, and the green line denotes performance of a model trained to minimize worst-case MSE over this set. We show the performance of a standard OLS predictor (blue line) and the invariant casual predictor (orange line) for reference. These types of asymmetric constraints on the uncertainty set can be represented by ellipsoidal constraints on ω in Equation (1.7), and we give a method for learning models that minimize a worst-case loss over these alternative sets of distributions. This type of customized constraint is also relevant for adapting models to new settings: Given the mean and covariance of a single proxy in the test distribution, one can learn a model (using the training data) with optimal performance on the test domain.

In this work, we assume that causal relationships in data can be well-approximated by linear models with additive noise, but this assumption is not generally realistic in many real-world applications. With that in mind, we view this research as most valuable as a tool for building intuition that is useful in further work in more general models (see e.g., Chapter 6). Indeed, linear SCMs are often used to motivate causality-based methods for robustness: For instance, Invariant Risk Minimization (IRM), introduced by Arjovsky et al. (2019), is a popular benchmark used in distributional generalization, whose main theoretical guarantees[5] are given under the assumptions of a linear SCM that relates latent variables to the outcome Y. Similar models are used in follow-up theoretical work to demonstrate some of the shortcomings of approaches like IRM (Rosenfeld and Risteski, 2020).

In the next chapter, we go beyond linear models, to enable the use of similar ideas in modern applications of machine learning such as computer vision.

[5] See Theorem 9 of Arjovsky et al. (2019)

1.5.3 Chapter 6: Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

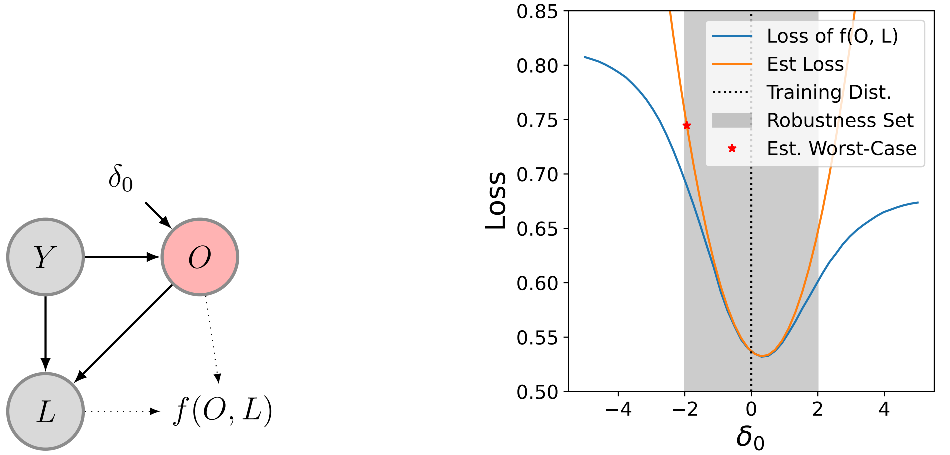
In this chapter, our goal is to *proactively* understand the sensitivity of a predictive model to distribution shift, using only data from the training distribution. The core goal is to estimate the worst-case loss of a given model $f(X)$ under a set of *plausible* future distributions. Our two main contributions are to (a) define a flexible framework for specifying a set of plausible, interpretable, and bounded shifts, and (b) an approximation-based approach for finding worst-case shifts from within that set, which empirically out-performs naive approaches like optimizing a re-weighted objective.

Prior work considers similar worst-case losses over uncertainty sets defined by e.g., f-divergence balls around the training distribution (Duchi and Namkoong, 2021; Duchi et al., 2020a; Subbaswamy et al., 2021). However, these approaches limit our ability to both interpret the resulting distributions and refine the set of allowable changes. For instance, consider a machine learning model that uses X-ray images or laboratory tests to predict disease: We might ask "*how would our performance change, if we tested fewer patients?*" or more broadly "*what changes to testing policy would cause the largest drops in predictive performance?*"

This type of question can be answered clearly via our approach, which we illustrate in the context of (Figure 1-11a), where disease (Y) is predicted using a binary indicator for whether or not a lab test has been ordered (O) and the resulting lab value, if available (L). First, for any factorization of the joint distribution $P(L, O, Y)$, the user specifies a set of factors that can change, e.g., choosing $P(O \mid Y)$ to capture a change in laboratory testing policies.

$$\begin{aligned} & P(L, O, Y) = P(L \mid O, Y) P(O \mid Y) P(Y) \\ & \quad \implies P_\delta(L, O, Y) = P(L \mid O, Y) P_\delta(O \mid Y) P(Y). \end{aligned}$$

Our approach allows for shifts in any number of factors, as long as each changing factor



(a) *Predicting disease* (Y) *using whether a lab test is ordered* (O) *and the lab result if ordered* (L).

(b) *Loss under perturbations* $s(Y;\delta) = \delta_0$ *(blue) alongside quadratic approximation (orange).*

Figure 1-11

is exponential family. Second, the user specifies a function $s(Y;\delta)$, parameterized by δ, to describe the shift, which enters as an additive change in the natural parameters of the exponential family

$$P_\delta(O \mid Y) = \text{sigmoid}(\eta(Y) + s(Y;\delta)),$$

where here, $\eta(Y)$ denotes the original conditional log-odds. In this example, a uniform increase or decrease in testing can be modelled as $s(Y;\delta) = \delta_0$, shown in Figure 1-11b. Finally, the user specifies a set of constraints on δ, e.g., quadratic constraints $\|\delta\|_2 < \lambda$, and we seek to estimate a worst-case loss

$$\sup_{\|\delta\| \leq \lambda} \mathbb{E}_{P_\delta}[\ell(f(X), Y)]. \tag{1.9}$$

In this general approach, each P_δ always shares support with the training distribution, enabling the use of importance sampling to estimate the expected loss. However, reweighting approaches can suffer from high variance in estimation, and finding the worst-case δ involves solving a non-convex optimization problem.

We therefore propose an alternative method, deriving a second-order Taylor approximation to the expected loss under shift (orange line in Figure 1-11b), whose approximation error can be bounded. For quadratic constraints on δ, this yields a tractable optimization objective, a non-convex, quadratically constrained quadratic program (QCQP) which can be solved efficiently (Conn et al., 2000). Using a GAN-based simulation derived from the CelebA dataset, we compare this proposed approach to a purely re-weighting based approach, and find that (for moderately sized shifts) the Taylor approximation approach tends to find better solutions to Equation (1.9), with more reliable estimates of the resulting loss.

1.5.4 Chapter 7: Auditing and Prompt Design for Large Language-Image Models

In this chapter, we illustrate the application of the approach given in Chapter 6 for probing the robustness of models to structured shifts in distribution. In particular, we consider CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), a self-supervised model of image-text pairs, which has demonstrated remarkable zero-shot performance on a variety of computer vision benchmarks. For instance, it outperforms a fully supervised linear classifier (fit on ResNet-50 features) on ImageNet. To perform zero-shot classification, CLIP performs matching between a set of fixed prompts (short strings of text), and a given image. To perform zero-shot classification, it typically suffices to use prompts of the form “a photo of a **label**.” (one for each label in the dataset), and choose the prompt with the highest similarity to the image. We refer to the selection of this set of strings as the problem of “prompt design”.

We probe the robustness of CLIP to structured shifts in distribution, and explore the impact of prompt design on the robustness of the resulting zero-shot classifier. We consider structured shifts in distribution when image attributes are available, with the presumed causal structure shown in Figure 1-12.

Unlike the other chapters in this thesis, the work presented here consists primarily of exploratory work, working through the challenges that can arise when applying the

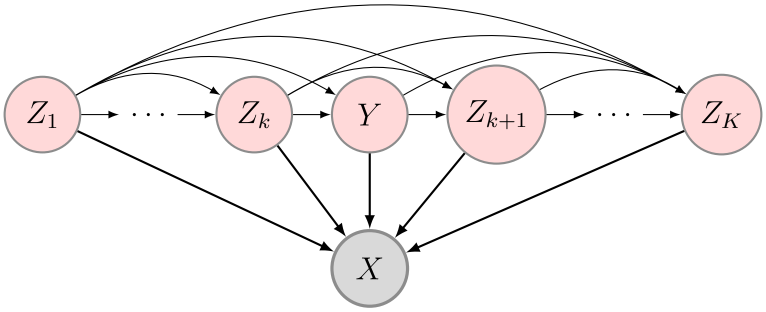


Figure 1-12: *An image X is a function of binary attributes Z and label Y. Some components of Z cause Y, while others are caused by Y, and all of their distributions are subject to change.*

method developed in Chapter 6 to the problem of model design (in this case, prompt design) on a real imaging dataset.

With that in mind, in this chapter we discuss some of those challenges, and potential solutions. The first challenge is the relative complexity of interpreting the shifts themselves, particularly in the absence of a clear causal structure over all variables. The second is the correct performance metric: While worst-case performance is intuitive in some ways, it is lacking in other ways, particularly when the magnitude of the change is not clear a-priori.

Part I

Reliable causal inference and policy evaluation

Chapter 2

Counterfactual Policy Introspection using Structural Causal Models

This chapter was previously published as my M.S Thesis (Oberst, 2019).

2.1 Introduction

2.1.1 Reinforcement Learning in Healthcare: A Challenging Task

There is a long tradition of using data to improve healthcare and public health, from randomized trials to test the efficacy of new drugs, post-market surveillance for adverse drug interactions, and the practice of epidemiology more broadly, e.g., the use of observational studies to understand the public health impact of everything from cigarettes to air pollution. Over the past decade in the United States, there has also been an ever-expanding amount of raw healthcare data, driven by the rapid adoption of electronic medical records (EMRs). As the available data has expanded, so have the ambitions of some segments of the research community, fuelled by the hope that larger and richer datasets can lead to breakthroughs in personalized medicine.

With that in mind, there has been a growing interest in the application of machine learning to healthcare, not only for diagnostic purposes (e.g., image processing in radiology and pathology), but also for learning better treatment policies, tailored to individual patients. This requires solving two closely related subproblems: First, how to learn a policy from observational (that is, retrospective) data, and second, how to evaluate it.

For sequential decision-making settings in healthcare, where a *dynamic treatment policy*[1] is required, several recent papers have used techniques from reinforcement learning (RL) to try and learn optimal policies for treating everything from sepsis (Raghu et al., 2017, 2018; Komorowski et al., 2018; Peng et al., 2018) to HIV (Parbhoo et al., 2017) and epilepsy (Guez et al., 2008). This is a challenging task, in ways that are quite different from modern success stories in reinforcement learning, such as achieving super-human performance at board games (Silver et al., 2018). The latter is a task that can be perfectly simulated, allowing for the (massive-scale) exploration and direct evaluation of different policies in a deterministic setting. In contrast, medicine is a stochastic, partially observable environment where direct experimentation by an algorithm would not be tolerable. As a result, we cannot simply try many policies and see if they work, but need to infer how a new policy would perform, using data collected under an older, different policy. In the RL literature, this is known as *off-policy evaluation*.

Of course, researchers in RL are not the first to have encountered this challenge. The evaluation of dynamic treatment policies (using observational data) is a well-studied causal inference problem in epidemiology and biostatistics, which is generally addressed with the application of g-methods, first introduced by Robins (1986). Lodi et al. (2016) and Zhang et al. (2018) are two recent examples, using g-methods to evaluate HIV treatment and anemia management strategies respectively. The techniques used in RL to evaluate novel treatment policies have much in common with these techniques, such as modelling the environment directly or re-weighting the observed data, as discussed in Section 2.2.

[1] A dynamic treatment policy is one which takes intermediate outcomes into account, like stopping a medical treatment when a patient has an adverse reaction

Quantitative evaluation is nonetheless fraught with difficulties that no mathematical method can address without making assumptions. For instance, if important variables are not measured (such as confounding variables, discussed in Section 2.2.1), then quantitative evaluation can give misleading results. These and other challenges, such as small effective sample sizes and miss-specification of reward, are discussed at length in Gottesman et al. (2019a).

Finally, a wealth of data exists in settings (e.g., EMRs, mobile health) that are not curated by any means, and are certainly not designed primarily for research purposes. This complicates matters further, and stands in contrast to research done with curated data registries, such as the US Renal Data System, used in Zhang et al. (2018), or sequentially randomized trials, such as the Strategic Timing of AntiRetroviral Treatment (START) trial, analyzed in Lodi et al. (2016).

2.1.2 Motivation: Debugging Policies and Models

Quantitative evaluation of policies can therefore be misleading for any number of reasons: There may exist unmeasured confounding in the dataset, the reward function (that is, the objective to be optimized) may be poorly specified, or there may not exist sufficient samples to evaluate policies that diverge too much from existing practice. Creating more robust methods for off-policy evaluation is an area of active research (Gottesman et al., 2019b; Liu et al., 2018; Kallus and Zhou, 2018a), but a fundamental uncertainty remains.

Moreover, it may be difficult to inspect a policy directly, to determine whether or not it seems reasonable: In contrast to the epidemiological studies mentioned earlier (Zhang et al., 2018; Lodi et al., 2016) which pre-specify a dynamic policy to evaluate based on domain knowledge, it is not always clear what a reinforcement-learned policy is doing. In Raghu et al. (2017), for instance, the policy is parameterized by a neural network, and in Komorowski et al. (2018), the policy associates an action with each of 750 patient state clusters derived via k-means clustering.

With that in mind, consider the following hypothetical: Suppose that you have the power to change medical practice, and are given a complex policy which is claimed (e.g., due to off-policy evaluation) to perform far better than existing clinical guidelines. How might you proceed? Given the challenges of retrospective evaluation, you might want to test the policy prospectively, perhaps using a randomized trial. But before you did that, you would want to better *understand* the policy, before investing a large amount of time and money in a gold-standard evaluation. In essence, you may wish to search for 'bugs' in the policy (like a tendency to take dangerous actions), or the model used to generate it (like the omission of a critical input), and iterate until you are confident that the policy has learned something reasonable.

There are a variety of ways you could do this, even if the policy is too complex to be interpretable directly. For instance, a physician might randomly select some real patients, pull up their full medical record, and compare the actions taken by the doctors to the recommendations of the learned policy, to see if they seem reasonable. Jeter et al. (2019) perform such an analysis in their critique of Komorowski et al. (2018), highlighting a sepsis patient where the learned policy makes a counter-intuitive decision to withhold treatment during a critical hypotensive episode. However, manual inspection of randomly selected trajectories may be inefficient, and difficult to interpret without more information: If we are to discover new insights about treatment, shouldn't there be *some* disagreement with existing practice?

This poses two problems: First, how do you surface the 'rationale' of a policy? In an ideal world, we could elicit a justification for each action. We refer to this as the challenge of *policy introspection.* Second, supposing that you could elicit these justifications *en masse* across all trajectories, how would you select the most interesting case examples for manual inspection?

2.1.3 Counterfactual Policy Introspection

In this chapter, we give a procedure that uses *counterfactual* trajectories to address both of these questions, and refer to this procedure as *counterfactual policy introspection*. Given a policy and a learned model of the environment, we provide a post-hoc method to generate counterfactual trajectories for each observed (or 'factual') trajectory, which attempt to describe what the model expects would have happened, in hindsight, if that policy had been used. We note that this is most useful in applications that already require the learning of a model of the environment, such as in model-based reinforcement learning. We can then compare counterfactual trajectories with observed trajectories, potentially with additional side-information (e.g., chart review in the case of a patient) so that domain experts can "sanity-check" a policy and the model used to learn it. In a way that we make precise in Section 2.3.2, if these counterfactuals are obviously wrong, then it provides evidence that the learned model of the environment is flawed.

Thus, our end-to-end procedure for 'debugging' models and policies is as follows, illustrated in Figure 2-1: First, once we have counterfactual trajectories for each observed trajectory, we can highlight episodes where there are surprisingly large differences between the factual and counterfactual outcomes. Second, we can then perform manual examination of the observed and counterfactual trajectories, to identify disagreements between the learned policy and existing practice, and to try and understand the rationale for them. Critically, because these are real patients, we can also go look for additional information to 'kick the tires' of the counterfactual conclusions. Finally, we can use our findings to iterate on the model and policy. For instance, looking at the medical record may suggest new variables to include in our model of the environment, at which point we can repeat the process again.

We stress that these counterfactuals are conceptually distinct from the simulation of new trajectories using a learned model of the environment. In particular, we don't want to know what the model believes might *generally* occur under a different policy: We want to know what would have been different in a *specific* trajectory. In Figure 2-2

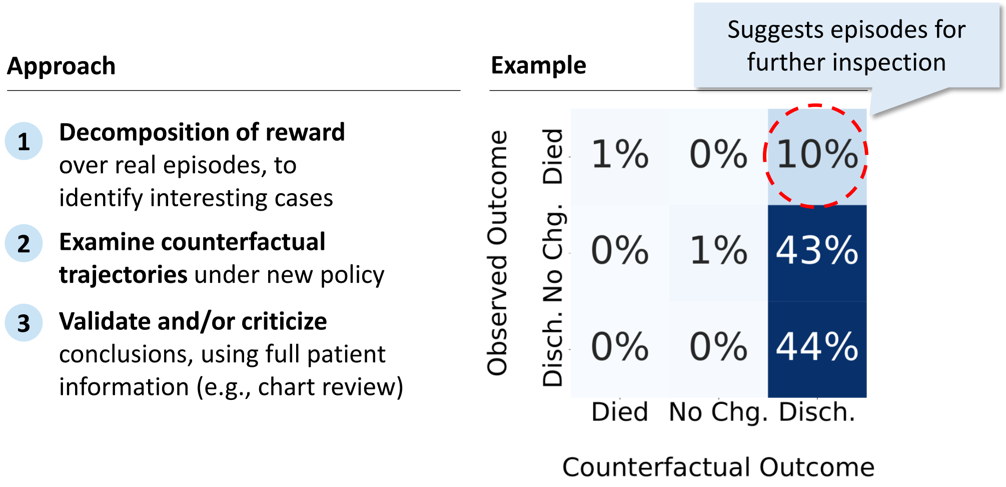

Approach
1
Decomposition of reward over real episodes, to identify interesting cases
2
Examine counterfactual trajectories under new policy
3
Validate and/or criticize conclusions, using full patient information (e.g., chart review)
Example
Suggests episodes for further inspection
Observed Outcome
Died
No Chg.
Disch.
1%
0%
10%
0%
1%
43%
0%
0%
44%
Died
No Chg.
Disch.
Counterfactual Outcome

Figure 2-1: *Conceptual overview of our approach: First, counterfactual trajectories are generated for all observed trajectories, and are then used to guide manual inspection. The figure on the right is taken from a synthetic example of sepsis management in Section 2.6.2, and highlights patients who died, but who would have allegedly lived in the counterfactual.*

we give a conceptual example of this distinction, in line with the medical use case described above. In this example, we imagine an observed trajectory where the patient had a rare, adverse reaction to an antibiotic. In a model-based simulation (or 'roll-out'), what might occur? Since the reaction is rare, then a model-based simulation might reasonably predict the most common outcome for patients *in general* (that the infection is cleared). Naturally, this does not satisfy our intuition for what would have happened to this specific patient (we already know!), but a model-based simulation is not designed to satisfy this intuition. A counterfactual trajectory, on the other hand, is designed to take into account what actually occurred to this patient, in a way that will be made precise in Section 2.2.3.

Moreover, counterfactual trajectories incorporate strictly more information about the observed trajectory, and thus exhibit less variance than a freshly simulated trajectory from a model. This is illustrated in a toy 2D grid-world setting in Figure 2-3, where the counterfactual trajectories in the left-hand figure (in blue) overlap perfectly with the observed trajectory (in black) when the actions are identical, and exhibit little

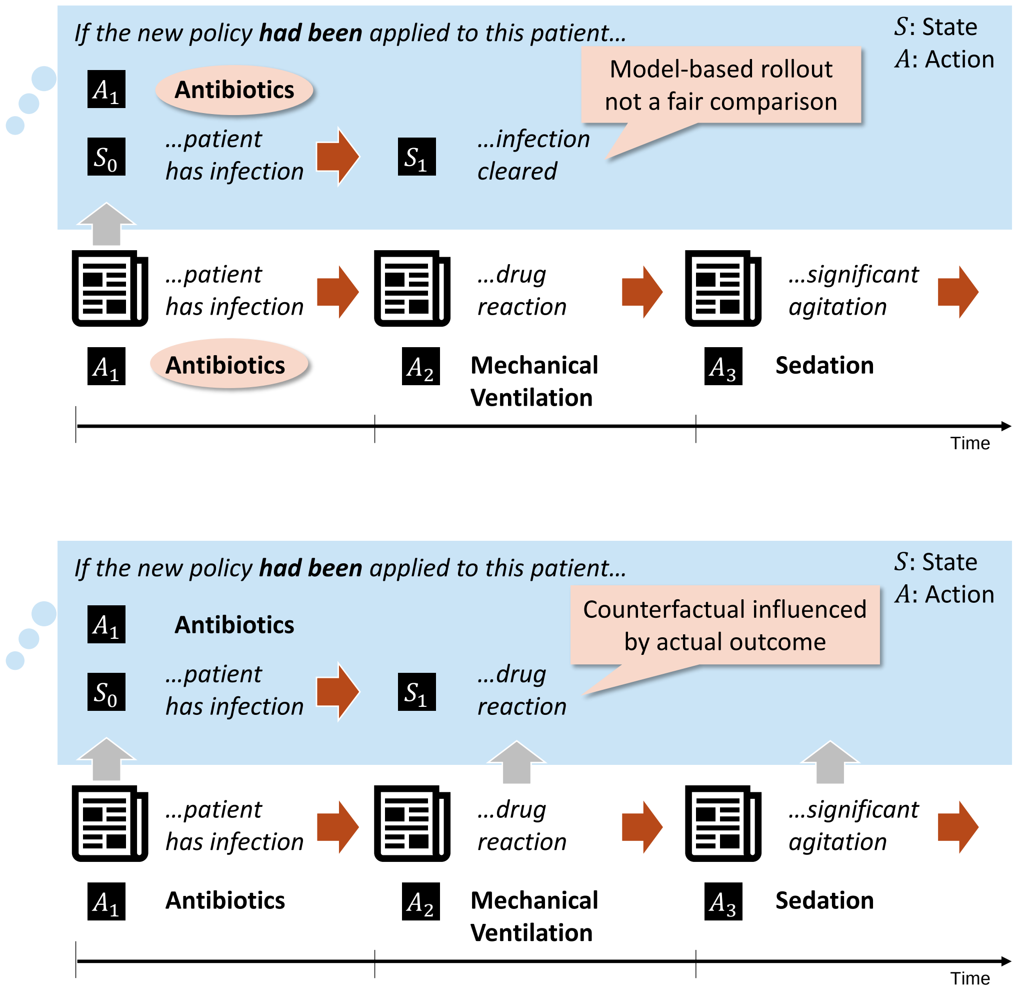

If the new policy **had been** applied to this patient...
S: State
A: Action
Model-based rollout not a fair comparison
A_1 Antibiotics
S_0 ...patient has infection
S_1 ...infection cleared
...patient has infection
...drug reaction
...significant agitation
A_1 Antibiotics
A_2 Mechanical Ventilation
A_3 Sedation
Time
If the new policy **had been** applied to this patient...
S: State
A: Action
Counterfactual influenced by actual outcome
A_1 Antibiotics
S_0 ...patient has infection
S_1 ...drug reaction
...patient has infection
...drug reaction
...significant agitation
A_1 Antibiotics
A_2 Mechanical Ventilation
A_3 Sedation
Time

Figure 2-2: *In this example, we imagine an observed trajectory where the patient had a rare, adverse reaction to an antibiotic. In a model-based roll-out, even if the trajectory is started in the same state, with the same initial action, it is unlikely that all model-based roll-outs will include this adverse event. Thus, the model-based roll-out is harder to critique: Perhaps the model is correct, and this patient just got unlucky. A counterfactual trajectory, on the other hand, is designed to isolate differences which are due to differences in actions.*

variability even after actions diverge. This is in contrast to the simulated trajectories in the right-hand figure (in red), which borrow no information from the observed trajectory, and thus are different from the beginning, even under identical actions. This example is discussed in far more depth in Section 2.6.1.

Returning to our motivating example of evaluating a complex treatment policy, it is

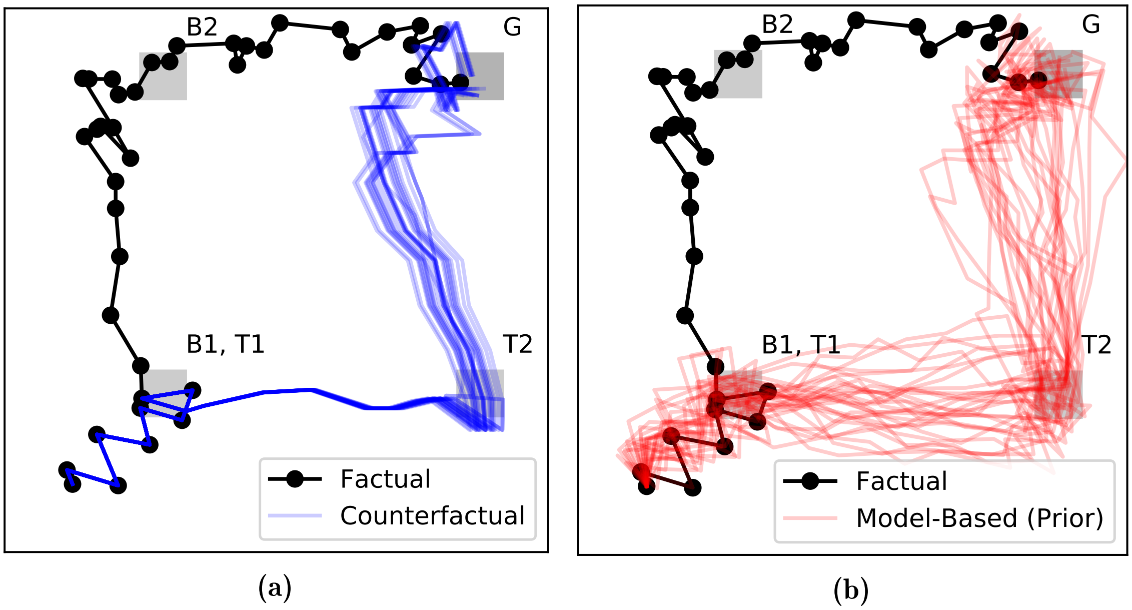


Figure 2-3: *A visual example of how counterfactuals isolate differences that are due solely to divergence in actions from the factual, taken from Section 2.6.1. The black line represents an observed trajectory, whereas the blue and red lines represent counterfactual trajectories and model-based simulations, respectively*

worth repeating that **these counterfactuals may be obviously wrong**, especially if we go to the medical record and use additional side information to check it against our intuition. This is a feature, not a bug, of our approach: In a setting where the model used for counterfactual evaluation is the same model that was used to train the policy, this can be used to confirm that suspicious actions (e.g., withholding treatment) are based on a faulty model of the world, versus a real insight into the best treatment.[2] In a model-based simulation, by contrast, this is difficult to ascertain: Was the model wrong, or was this patient just one of the unlucky ones?

However, towards generating these counterfactual trajectories, we have to deal with a fundamental issue of non-identifiability: As we show in Section 2.4.1, even with an infinite amount of interventional data, there are multiple structural causal models (as introduced in Section 2.2.3) which are consistent with with the data we observe, but which suggest different distributions of counterfactual outcomes on an individual level. This is not a new problem, and a common assumption in the binary setting

[2] We make this intuition precise in Section 2.3.2

to identify counterfactuals is the *monotonicity* condition (Pearl, 2000). However, to our knowledge, there is no analogous condition for the categorical case, as would be required to generate counterfactuals in discrete state-space models of the environment.

This motivates our main *theoretical* contribution, which is two-fold. First, we introduce a general condition of *counterfactual stability* for structural causal models (SCMs) with categorical variables and prove that this condition implies the monotonicity condition in the case of binary categories. Second, we introduce the *Gumbel-Max SCM*, based on the Gumbel-Max trick for sampling from discrete distributions, and demonstrate that it satisfies the counterfactual stability condition. We note that any discrete probability distribution can be sampled using a Gumbel-Max SCM; As a result, drawing counterfactual trajectories can be done in a post-hoc fashion, given any probabilistic model of dynamics with discrete states. To conclude, we restate our main contributions, which are as follows:

1. **Using Counterfactuals for Policy Introspection and Model-Checking:** Our main conceptual contribution is the procedure described above, using counterfactual trajectories as a tool for introspection of learned policies and models. Additionally, we build on the theoretical results of (Buesing et al., 2019) in Section 2.3.2 to note that the expected counterfactual reward over all factual episodes (if the SCM is correctly specified) is in fact equal to the expected reward using freshly simulated trajectories. In this way, if counterfactual conclusions are incorrect on their face, it casts suspicion on the learned model of dynamics used in the first place, and any quantitative estimate of reward (as derived through e.g., the parametric g-formula, discussed in Section 2.2.1) that it yields.

2. **Counterfactual Stability and Gumbel-Max SCMs:** Our main theoretical contribution is twofold: First, we introduce the property of *counterfactual stability* for SCMs with categorical variables, and prove that this condition implies the monotonicity condition (Pearl, 2000) in the case of binary categories. Second, we introduce the Gumbel-Max SCM, a general SCM for categorical variables which we prove to satisfy the counterfactual stability condition. We note that any

discrete probability distribution can be sampled using a Gumbel-Max SCM; As a result, drawing counterfactual trajectories can be done in a post-hoc fashion, given any probabilistic model of dynamics with discrete states.

3. **Application to a Real-World Setting**: In addition to a series of synthetic examples, we replicate the work of Komorowski et al. (2018) in learning a policy for sepsis management using EMR data. We apply counterfactual policy introspection with the assistance of a domain expert (in this case, a clinician), including the review of specific counterfactual trajectories using the full medical record as side information.

2.1.4 Structure of this chapter

- **Background (Section 2.2):** We review the interrelated problems of learning and evaluating a dynamic policy, drawing connections between the literature on causal inference and model-based reinforcement learning. We also review the concepts necessary for generating counterfactuals, such as structural causal models. We draw a distinction between counterfactual and interventional distributions, and highlight both the inherent non-identifiability of counterfactuals, as well as the monotonicity assumption used to identify them in the binary case.

- **Counterfactual Decomposition of Reward (Section 2.3)**: We begin by demonstrating how common causal models assumed in the RL literature (MDPs and POMDPS) can be cast as structural causal models. We further discuss the connection between counterfactual estimates of rewards and notions like CATE and ITE in the causal inference literature. We conclude by building on the theoretical results of (Buesing et al., 2019) in Section 2.3.2 to note that the expected counterfactual reward over all factual episodes (if the SCM is correctly specified) is in fact equal to the expected reward using freshly simulated trajectories.

- **Gumbel-Max SCMs for Categorical Variables (Section 2.4)**: With the motivation from Section 2.3 in mind, in this section we introduce our core theoretical contributions. First, we introduce the property of categorical stability as a categorical analog of the montonicity assumption. Then, we introduce and motivate the Gumbel-Max SCM by proving that it satisfies this property. We also highlight connections to the discrete choice literature, which are useful for building intuition around the counterfactual stability condition.

- **SCMs with Additive Noise for Continuous Variables (Section 2.5)**: In this brief section, we highlight some possible approaches for developing general SCMs for continuous variables, by examining common continuous state-space models in RL and giving an SCM which is consistent with their formulation.

- **Illustrative Applications with Synthetic Data (Section 2.6)**: To build intuition, we demonstrate the use of counterfactual trajectories in two idealized environments: A 2D grid-world and an illustrative simulator of sepsis. The former builds intuition for how counterfactual inference works in SCMs, while the latter demonstrates our proposed use of counterfactuals for policy introspection.

- **Real-Data Case Study: Sepsis Management (Section 2.7)**: In this section, we replicate the work of Komorowski et al. (2018) using real EMR data to learn a policy of sepsis management, and we apply our proposed methodology to perform introspection of the resulting policy. Most notably, we use the full medical record and the help of a clinician to examine counterfactuals for a particular trajectory, and discuss our insights from this exercise in Section 2.7.4.

2.2 Background

In this section, we lay out the necessary background for the later sections. Broadly speaking, we start by discussing the central problem of learning how to act from data. This is intrinsically a *causal* question: We would like to claim that if we acted in a particular way, this would bring about a particular outcome. Thus, in Section 2.2.1, we discuss some basic principles of causal inference, starting with the simplest case of estimating the effect of a binary action from interventional data (as in a randomized control trial), before moving on to techniques used to estimate the effect of dynamic treatment regimes from observational data. We highlight in particular some general classes of methods: Those which model the causal relationships directly, those which rely on re-weighting the data, and those which combine the two approaches.

With this background in hand, we turn to the problem of *learning* a policy from data, and highlight methods used in the reinforcement learning (RL) community for doing so in Section 2.2.2. We draw an explicit connection to the literature on dynamic treatment regimes, noting that RL methods can be viewed as assuming a particular causal graph with a certain Markov structure. With this assumption in mind, we discuss a basic method for learning an optimal policy, known as Policy Iteration, which falls under the general class of RL methods which are 'model-based', in that they assume access to a model of the environment. We then discuss two approaches in the RL literature for evaluating policies that are different from the one that generated the data, a problem known as *off-policy evaluation*: The first of these methods, known as model-based off-policy evaluation (MB-OPE) bears some similarity to the g-formula used in the literature on evaluating dynamic treatment regimes. The second method is a re-weighting method, which is similar to inverse propensity (IP) weighting methods, another set of g-methods.

Finally, we introduce the notion of counterfactuals in Section 2.2.3, where we formalize the distinction between *interventional* questions, like 'what *will happen* if I apply policy X', and *counterfactual* questions, like 'what *would have happened* if I had applied policy X, given that I applied policy Y and observed outcome Z'. To do so, we introduce the mathematical framework of structural causal models, and highlight the challenges inherent in estimating counterfactuals, which are by definition never observed. We note that this is a different (and strictly more challenging) problem than the usual causal inference question, because it deals with individual-level counterfactuals (analogous to the individual treatment effect), instead of population-level causal effects (analogous to the conditional average treatment effect).

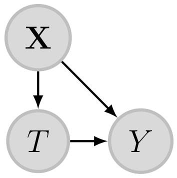


Figure 2-4: *Causal graph corresponding to the motivating example of a binary treatment and binary outcome*

We refer the reader to several reference on the above topics for more detail, in lieu of attempting to reproduce the entirety of these fields within the confines of this chapter. In particular, we recommend Hernan and Robbins (2019) for an overview of causal inference with dynamic treatment regimes, and Pearl (2009); Peters et al. (2017) for an overview of causal graphs and structural causal models. For a general overview of reinforcement learning, we recommend Sutton and Barto (2017).

2.2.1 Causal Inference from Observational Data

Motivating Example: Binary Treatments

Suppose that we want to evaluate the causal effect of a binary action, such as taking an antibiotic, on a binary outcome, such as whether or not an infection is cleared. Let $T \in \{0, 1\}$ represent the action (whether or not we gave the treatment), and let $Y \in \{0, 1\}$ represent the outcome. Suppose we also have access to covariates / features X which describe potential confounding factors, so-called because they influence both the treatment decision and the outcome. For any given individual, we can use Y_1 and Y_0 to represent the *potential outcomes* (Morgan and Winship, 2014) under the treatment and control respectively, of which we only observe one of the two, e.g., $Y = Y_1T + Y_0(1 - T)$. We can also denote this set-up using a *causal graph*, a directed acyclic graph (DAG) which encodes the causal relationships between random variables (Pearl, 2009). In this case, the corresponding DAG is given in Figure 2-4, with arrows that represent the causal relationships between variables.

In this example, we might be interested in the *average treatment effect* (ATE), which can be denoted by

$$\tau = \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)],$$

where the $do(\cdot)$ operator is used to indicate an intervention. The $do(\cdot)$ operator is reviewed in (Pearl, 2009), and is accompanied by the rules of *do-calculus*, which give us a set of conditions which specify when (and how) it is possible to obtain causal relationships, like $\mathbb{P}(Y|do(T = t))$, from observed conditional relations like $\mathbb{P}(Y|T = t)$. Intuitively, the ATE corresponds to the expected difference in outcome between two policies, where we treat everyone $\mathbb{E}[Y|do(T = 1)]$ or we treat no one $E[Y|do(T = 0)]$. In the simplest case, if the treatment assignment is randomized such that $\mathbb{P}(T|X) = \mathbb{P}(T)$, then we have the equivalence $\mathbb{E}[Y|do(T = t)] = \mathbb{E}[Y|T = t]$. For instance, in an ideal randomized control trial with full compliance, we could estimate the causal effect by simply looking at the difference in outcome between the treatment and control groups.

Dealing with Observational Data

It should be noted that causal inference requires assumptions, which are often not empirically verifiable. For instance, if treatment assignment is not randomized, as is typical for observational data, a common approach is to first make the assumption of *no unmeasured confounding*: That is, we assume that we observe, through X, all of the variables which impact both the treatment and the outcome. We refer the reader to a variety of references (Hernan and Robbins, 2019; Pearl, 2009; Morgan and Winship, 2014; Imbens and Rubin, 2015) for a more comprehensive treatment of the topic, but we will briefly highlight three broad approaches, which have analogs in the reinforcement learning literature.

- First, we can model the conditional relationships directly, by estimating $\mathbb{P}(Y|X, T)$, which is equivalent to $\mathbb{P}(Y|X, do(T))$ under the assumption of no unmeasured confounding, and use this to calculate $\mathbb{P}(Y|do(T)) = \int \mathbb{P}(Y|X, T)\mathbb{P}(X)dx$ by marginalizing over X. This is known as *standardization* in epidemiology.

- Second, we can re-weight the data to create a *psuedo-population* that approximates the results of a randomized trial. For instance, we might use an estimate of the treatment probability $\mathbb{P}(T|X)$, known as the propensity score, and use this to re-weight our observations (Rosenbaum and Rubin, 1983b), or stratify into sub-populations with similar propensity (Rubin and Rosenbaum, 1984). The more general form of this approach (discussed below) is known as *inverse probability (IP) weighting* in epidemiology.

- Finally, we can combine the two approaches above to develop *doubly-robust* estimators (Bang and Robins, 2005), which provide asymptotically correct estimates if we can correctly estimate either $\mathbb{P}(Y|X,T)$ or $\mathbb{P}(T|X)$.

ATE, CATE, and ITE

So far, we have implicitly focused on a very simple decision-making problem, by focusing on the estimation of the ATE. In effect, this corresponds to evaluating the difference in the expected outcome between two policies: 'Treat everyone' and 'treat no one'. In the notation of potential outcomes, introduced in Section 2.2.1, the ATE corresponds to the quantity

$$\tau = \mathbb{E}[Y_1 - Y_0]$$

We can refine this further by investigating the *conditional average treatment effect* (CATE), which conditions on a specific subpopulation X, and can be denoted by the quantity

$$\tau_x = \mathbb{E}[Y_1 - Y_0|X]$$

In the causal graph given in Figure 2-4, this can (in principle) be estimated directly using regression models $\hat{f}(X,T) \approx \mathbb{E}[Y|X,T]$ since $P(Y|X, do(T)) = P(Y|X,T)$ in this case. How does this relate to learning a policy? In this simple setting, learning a policy follows naturally from evaluating the effect of the binary treatment. For instance, once we have learned the CATE, we can devise a policy which treats each patient (with covariates X) based on the sign of the estimated CATE $\hat{\tau}_x$.

Note that there is a conceptual distinction between the CATE and what we will refer to as the *individual treatment effect* (ITE), which is simply the difference in potential outcomes, denoted for an individual j by

$$\tau_{ite}^{(j)} = Y_1^{(j)} - Y_0^{(j)}$$

Unlike the ATE and CATE, this represents a statement about a specific individual, versus an expectation over a population. This can be a source of confusion when it comes to the use of counterfactual language: It is not uncommon to estimate the CATE and refer to this as a *counterfactual* or to refer to the CATE as the ITE (see Shalit et al. (2016) and discussion in Appendix B of Liu et al. (2018)).

Note that in this chapter, we will reserve the language of counterfactuals and counterfactual inference to refer to individual-level quantities, like $Y_0^{(j)}, Y_1^{(j)}$.

Extension to Dynamic Treatment Policies

Many of the methods which were originally developed for the simple setting described above do not work (when applied naively) to the setting where we wish to evaluate a dynamic treatment. In this setting, our initial action may have some intermediate effect which influences our choice of later actions, and so on. Robins (1986) introduced a class of general methods for adjustment in this setting, which are referred to *g-methods* in the dynamic treatment regime literature. Among these, we highlight two methods which are analogs to those discussed previously:

- First, the *g-computation algorithm formula*, typically referred to as the g-formula, is a generalization of the standardization approach given in Section 2.2.1. Simply put, the conceptual approach is to estimate the outcome under a specific policy by *simulating* from a model of the overall environment. The g-formula is widely used in epidemiology, where it is referred to as the parametric g-formula when it involves fitting a parametric model of the environment. For instance, Lodi et al. (2016) use this approach to evaluate a policy for HIV treatment, and Zhang et al.

(2018) use it to evaluate a strategy for anemia management.

- Second, the class of inverse probability (IP) weighting methods, which generalize the re-weighting methods discussed previously, such as propensity score re-weighting (Rosenbaum and Rubin, 1983b). See (Hernan and Robbins, 2019) for a more in-depth discussion, including the combination of IP weighting methods with marginal structural models.

2.2.2 Model-Based Reinforcement Learning

With all of this in mind, we shift gears to a different set of literature, namely that of reinforcement learning (RL). In contrast to the above sections, where our focus was on *evaluating* a policy based on observational data, reinforcement learning has its roots in trying to *learn* a policy efficiently, when given the ability to experiment freely in an environment. We cannot hope to summarize all the extant techniques that exist for learning and evaluation in RL, but instead highlight those which are relevant for future chapters, as well as for understanding where our approach fits in.

Seen in relationship to the literature on dynamic treatment regimes, the reinforcement learning literature tends to assume a particular type of causal graph, a Markov Decision Process (MDP), which we describe in Section 2.2.2. While this assumption is shared across techniques used to learn a policy, there is a further distinction between methods which are *model-based*, which rely on learning to model the MDP, versus those that are 'model-free', in the sense that they do not model the environment directly. The techniques discussed in this chapter require a model of the environment, and thus we will focus our discussion in Section 2.2.2 on a simple model-based approach to learning a policy, known as Policy Iteration.

Finally, we discuss two broad types of evaluation, which have connections to the two classes of evaluation methods discussed in the previous section: First, model-based off-policy evaluation (MB-OPE), which can be seen as a specific instance of simulation via the g-formula, and importance re-weighting methods such as weighted importance sampling (WIS), which can be seen as instances of the inverse probability weighting approach described earlier.

Markov Decision Processes (MDPs and POMDPs)

The reinforcement learning literature tends to assume an underlying model of the world which can be represented as having a *Markov* structure, meaning that the state of the world in the future is independent of the past, given the present (observable) state. This leads to a representation which is known as a *Markov Decision Process* (MDP). This can be relaxed by assuming that there exists an underlying Markov structure, but we may not observe it, in which case it is considered a *partially observable Markov Decision Process* (POMDP). In this section we describe these general models, as a prelude to discussing their role in both learning and evaluation.

We follow the description of Finite Markov Decision Processes (MDPs) given in Sutton and Barto (2017), to which we refer the reader for more information. In this setting, the decision-maker (or *agent*) interacts with an *environment* at each discrete time step. The decision maker is presented with a state $S \in \mathcal{S}$, and chooses an action $A \in \mathcal{A}$, which result in a new state $S' \in \mathcal{S}$ as well as a quantitative reward $R \in \mathcal{R}$, and the process continues until an absorbing state is reached, or until a fixed time (known as a fixed-horizon MDP). These states, actions, and rewards are typically indexed by time, and follow the conditional probability distribution (CPD) that governs the MDP, and which is referred to (in this work) as the *dynamics* of the process:

$$\mathbb{P}(S_{t+1}, R_t | S_t, A_t) \tag{2.1}$$

Note that the CPD in Equation (2.1) is Markov in the sense that the next state / reward only depend on the previous state and action, hence the moniker of a Markov Decision Process. Furthermore, this CPD is often assumed to be invariant to the time index, in which case we refer to this as a *homogenous* MDP. Finally, when the state space $\mathcal{S}$ has finite cardinality, we refer to this as a finite MDP.

The goal of the decision-maker at time t is typically to maximize the discounted expected reward over the future states. This is typically denoted as follows[3]

$$G_t := \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2.2}$$

In Equation (2.2), the discount factor $0 \leq \gamma \leq 1$ determines the degree to which future rewards are less valuable than immediate rewards, and this notation can be used to cover episodes which have a finite horizon or terminal states, using the assumption that after the horizon or a terminal state is reached, the subsequent rewards are all zero.

Thus, the goal of the decision-maker is to choose a policy π which maximizes the expected reward. This policy can either be deterministic, in which case $\pi : \mathcal{S} \to \mathcal{A}$ maps states to actions, or stochastic, in which case $\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ gives a probability density or mass function over the set of possible actions for each state, such that $\sum_{a \in \mathcal{A}} \pi(s, a) = 1, \forall s \in \mathcal{S}$. With a slight abuse of notation, we will sometimes write $\pi(a|s)$ in place of $\pi(s, a)$ to convey the fact that it describes a conditional probability distribution over actions.

An extension of this framework is to consider a *partially observable* MDP (POMDP), in which we distinguish between the true state S_t and the observation O_t at each time step, with the assumption that the true state S_t is unobserved. In this case, the generative model is augmented with the CPD $\mathbb{P}(O_t|S_t)$. In the case of a POMDP, the policy may depend on the entire *history* up to time point t, which is denoted as $H_t := \{O_1, A_1, R_1, \ldots, O_{t-1}, A_{t-1}, R_{t-1}, O_t\}$, such that the policy is given by $\pi(a|h)$, with $h \in \mathcal{H}$ informing the action taken.

A *trajectory* or *episode*, denoted τ, is the full sequence of states, actions, and rewards, up to the terminal state or horizon. For a MDP, given a probability distribution over initial states $\mathbb{P}(S_1)$ and policy $\pi(a|s)$, the probability of any given trajectory

[3] See equation 3.8 from Sutton and Barto (2017)

$\tau = \{S_1, A_1, R_1, \ldots, S_T, A_T, R_T\}$ is given by

$$p(\tau) = \mathbb{P}(S_1) \prod_{k=2}^{T} \pi(A_{k-1}|S_{k-1}) \mathbb{P}(S_k, R_k|A_{k-1}, S_{k-1}) \tag{2.3}$$

With an analogous factorization in the case of a POMDP. Because this distribution depends on the policy π, we denote this distribution over τ by $p^\pi(\tau)$, and for any quantity which is a function of the trajectory (e.g., the total reward G), we will write $\mathbb{E}_\pi(\cdot)$ to denote the expected value over trajectories drawn from $p^\pi(\tau)$.

Policy Iteration Algorithm

There are a variety of techniques used to find an optimal policy in the case of a finite MDP, but for our purposes it will be sufficient to discuss the techniques used in (Komorowski et al., 2018), which use straightforward iterative optimization techniques that depend on knowledge of the MDP, which can be estimated from data.

First, we need to introduce the concept of the *value function* for each state, which is defined with respect to a policy π by[4]

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] && (2.4) \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] && (2.5) \end{aligned}$$

With this in hand, the *policy evaluation* problem is to estimate the value function for a given policy. Equation (2.5) defines a fixed point, and the following iterative update rule is known to converge to true value function

$$v_\pi^{(k+1)}(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi^{(k)}(s')\right], \tag{2.6}$$

where $v^{(k)}$ is the value function at the k-th iteration. Initializing a random value function and applying these updates until some desired tolerance is known as the

[4] See Equation 4.4 from Sutton and Barto (2017)

iterative policy evaluation algorithm.

Using this technique for evaluating a policy as a subroutine, the *policy iteration* algorithm improves the policy at each step, using the update rule given by

$$\pi'(a|s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] \tag{2.7}$$

To summarize, policy improvement starts with a random (deterministic) policy and a randomly initialized value function, then alternates between policy evaluation and policy improvement, until it finds a stable policy. For more detail, we refer the reader to Chapters 4.1–4.3 of Sutton and Barto (2017).

Off-Policy Evaluation (OPE)

In the RL literature, it is commonly assumed that we are able to learn from experience. That is, we can experiment with different policies until we find a policy that maximizes our expected reward. From the perspective of healthcare applications, this is analogous to assuming that we can freely run our own randomized experiments as we go along. Evaluation in this setting (the *on-policy* setting) is conceptually straightforward, similar to a randomized trial.

In this chapter, we deal with the setting where this type of experimentation is not possible, e.g., for ethical and practical reasons, and we are restricted to using observational data. This type of setting is referred to in the RL literature as *off-policy batch RL*, to reflect that the policy used to generate the data (the *'behavior' policy*) is different from the policy we wish to evaluate (the *'target'* or *'evaluation' policy*) and the fact that our dataset is restricted to a fixed batch of data.

Here we discuss two methods for off-policy evaluation, which have connections to the classes of evaluation methods discussed in Section 2.2.1:

- Model-based off-policy evaluation (MB-OPE) involves learning a parametric model of an underlying MDP, and then using this to estimate the value of a policy (see e.g., Chow et al. (2015); Hanna et al. (2017)), and can thus be seen as a specific instance of simulation via the g-formula.

- Importance sampling (IS) (Rubinstein, 1981) is the foundation for a series of techniques, such as weighted importance sampling (see e.g., Precup et al. (2000)). As discussed below, these are similar to IP weighting methods.
- There exist several methods for combining these approaches, whether to generate doubly robust estimates of performance (Jiang and Li, 2016; Bibaut et al., 2019; Farajtabar et al., 2018), or using a mixture of IS and MB estimates (Thomas and Brunskill, 2016; Gottesman et al., 2019a).

We take a moment here to describe the form of a basic IS estimator, as well as weighted importance sampling (WIS), as they will be relevant for our later experimental work replicating Komorowski et al. (2018). In general, importance sampling and related approaches (IP weighting, inverse propensity weighting) take advantage of the following relationship, where p, q are two different distributions

$$\mathbb{E}_p[Y] = \int y \cdot p(y) dy = \int y \cdot \frac{p(y)}{q(y)} q(y) = \mathbb{E}_q \left[\frac{p(y)}{q(y)} Y \right]$$

This is the same basic theory that underlies all the IP weighting methods discussed so far.[5] Thus, given samples of a random variable from a distribution q, we can approximate the expectation under the distribution p using the weights $p(y_i)/q(y_i)$ for each y_i, and taking a sample average $\mathbb{E}_p[Y] \approx n^{-1} \sum y_i \cdot p(y_i)/q(y_i)$

In an RL context, we want to estimate the expected reward of an evaluation policy π_e, given data sampled from an MDP under a behavior policy π_b. In this case the importance ratio is straightforward. Examining the probability of any given trajectory, given in Equation 2.3, we note that all the terms cancel in the importance sampling ratio, except for those which involve the policy. Thus, the importance sampling ratio

[5] Note that this relationship is only well-defined if $p(y) > 0 \implies q(y) > 0$. This condition goes by various names depending on the field: In probability theory, it is referred to as *absolute continuity*. In the context of inverse propensity weighting, it is referred to as *overlap* or *positivity*. In the context of reinforcement learning, it is referred to as *coverage*.

is given by the following, where we use $\rho_{1:T}$ to denote the importance sampling ratio over T time steps

$$\rho_{1:T} = \prod_{i=1}^{T} \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}$$

Using importance sampling, we can get an unbiased and consistent estimator of the reward under the evaluation policy using $\mathbb{E}_{\pi_e}[G] \approx n^{-1} \sum_i \rho^{(i)} G^{(i)}$, where we drop the subscript on ρ, use the superscript to indicate observed trajectories, and write G as the total discounted reward. However, in practice the IS estimator can exhibit high variance, especially if some actions are rare under the behavior policy (such that $1/\pi_b(a_t|s_t)$ is very large).

Weighted importance sampling is an alternative estimator which exhibits much lower variance, albeit at the cost of introducing some bias.[6] The weighted importance sampling estimator performs a weighted average instead of a simple average, and is given by

$$\frac{\sum_i \rho^{(i)} \cdot G^{(i)}}{\sum_i \rho^{(i)}},$$

It is important to note that all variants of importance sampling are subject to the same assumptions as any other causal analysis. That is, we typically need to estimate the behavior policy from data, and if there is some unmeasured confounding factor which cause our estimates of the behavior policy π_b to be incorrect, then our IS or WIS estimates will also be incorrect. This well-known fact is demonstrated in our synthetic experiments in Section 2.6.2.

2.2.3 Structural Causal Models and Counterfactuals

When we discussed binary treatments in Section 2.2.1 we discussed potential outcomes Y_1, Y_0. In that setting, we observe one of these, but the other is unknown, representing the theoretical counterfactual outcome. In many applications of causal inference, we

[6] Weighted importance sampling is still consistent, in the sense that it converges to the correct value in the infinite data limit, with the bias asymptotically approaching zero.

wish to estimate some general effect of an intervention, such as the conditional average treatment effect $\mathbb{E}[Y_1 - Y_0|X]$ (e.g., Schulam and Saria, 2017; Johansson et al., 2016), because this represent general knowledge about interventions that we can apply to future patients. But we do not particularly care about e.g., estimating Y_0 given Y_1 for a particular patient that we have already treated, because we cannot go back in time and take a different action.

In a sense that we will make precise in Section 2.2.3, the CATE is a property of the *interventional* distribution of Y, describing how Y changes in response to interventions on other variables (in this case, T). However, we would like to go a step beyond this, as described in Section 2.1.3. We would like to take into account *what actually happened* to get a more precise estimate of *what would have happened* had a different action (or set of actions) been taken. This is a *counterfactual* question. In essence, we want to estimate something that is conceptually akin to the individual treatment effect $Y_1 - Y_0$, rather than just the CATE.

To do so, we need to introduce the mathematical formalism of structural causal models, which give a well-defined answer to these questions. In Section 2.2.3 we introduce the general framework, in Section 2.2.3 we formalize the conceptual distinction between interventional and counterfactual distributions, and in Sections 2.2.3-2.2.3 we discuss the fundamental challenge of non-identifiability, as well as some assumptions that make identification possible in the binary case.

Structural Causal Models (SCMs)

As promised, we review the concept of *structural causal models*, and encourage the reader to refer to Pearl (2009) (Section 7.1) and Peters et al. (2017) for more details. A word regarding notation: As a general rule throughout, we refer to a random variable with a capital letter (e.g., X), the value it obtains as a lowercase letter (e.g., $X = x$), and a set of random variables with boldface font (e.g., $\mathbf{X} = \{X_1, \ldots, X_n\}$). Consistent with Peters et al. (2017) and Buesing et al. (2019), we write P_X for the distribution of a variable X, and p_x for the density function.

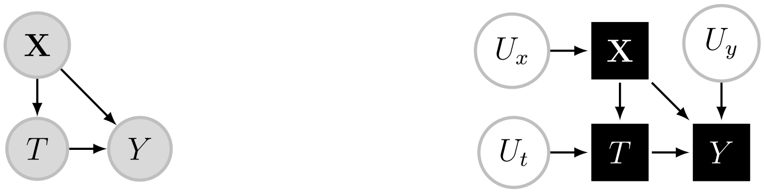


Figure 2-5: *Example translation of a causal graph into the corresponding Structural Causal Model.* ***Left:*** *Causal DAG on an outcome Y, covariates X, and treatment T. Given this graph, we can perform do-calculus (Pearl, 2009) to estimate the impact of interventions such as $\mathbb{E}[Y|X, do(T=1)] - \mathbb{E}[Y|X, do(T=0)]$, known as the Conditional Average Treatment Effect (CATE).* ***Right:*** *All observed random variable are assumed to be generated via structural mechanisms f_x, f_t, f_y via independent latent factors U which cannot be impacted via interventions. Following convention of Buesing et al. (2019), calculated values are given by black boxes (and in this case, are observed), observed variables are given in grey, and unobserved variables are given in white.*

Definition 2.1 (Structural Causal Model (SCM))**.** A structural causal model $\mathcal{M}$ consists of a set of independent random variables $\mathbf{U} = \{U_1, \ldots, U_n\}$ with distribution $P(\mathbf{U})$, a set of functions $\mathbf{F} = \{f_1, \ldots, f_n\}$, and a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ such that $X_i = f_i(\mathbf{PA}_i, U_i), \forall i$, where $\mathbf{PA}_i \subseteq \mathbf{X} \setminus X_i$ is the subset of $\mathbf{X}$ which are parents of X_i in the causal DAG $\mathcal{G}$. As a result, the prior distribution $P(\mathbf{U})$ and functions $\mathbf{F}$ determine the distribution $P_X^{\mathcal{M}}$.

As a motivating example to simplify exposition, we will assume the causal graphs (and corresponding SCM) given in Figure 2-5. An astute reader will recognize this as the same binary setting discussed previously, representing (for example) the effect of a medical treatment T on an outcome Y in the presence of confounding variables $\mathbf{X}$.

Interventional vs. Counterfactual Distributions

The SCM $\mathcal{M}$ defines a complete data-generating processes, which entails the *observational distribution* $P(\mathbf{X}, Y, T)$. It also defines an *interventional distribution*, describing the effect of any possible intervention.

Definition 2.2 (Interventional Distribution)**.** Given an SCM $\mathcal{M}$, an intervention $I = do\left(X_i := \tilde{f}(\tilde{\mathbf{PA}}_i, \tilde{U}_i)\right)$ corresponds to replacing the structural mechanism $f_i(\mathbf{PA}_i, U_i)$ with $\tilde{f}_i(\tilde{\mathbf{PA}}_i, U_i)$. This includes the concept of atomic interventions, where we may write more simply $do(X_i = x)$. The resulting SCM is denoted $\mathcal{M}^I$, and the resulting distribution is denoted $P^{\mathcal{M};I}$.

For instance, suppose that Y corresponds to a favorable binary outcome, such as 5-year survival, and T corresponds to a treatment. Then several quantities of interest in causal effect estimation, including (but not limited to) the ATE and the CATE, are defined by the interventional distribution, which is *forward-looking*, telling us what might be expected to occur if we applied an intervention. However, we can also define the *counterfactual distribution* which is *retrospective*, telling us what might have happened had we acted differently. For instance, we might ask: Having given the drug and observed that $Y = 1$ (survival), what *would have happened* if we had instead withheld the drug? This is formalized in an SCM as follows:

Definition 2.3 (Counterfactual Distribution)**.** Given an SCM $\mathcal{M}$ and an observed assignment $\mathbf{X} = \mathbf{x}$ over any set of observed variables, the counterfactual distribution $P_X^{\mathcal{M}|\mathbf{X}=\mathbf{x};I}$ corresponds to the distribution entailed by the SCM $\mathcal{M}^I$ using the posterior distribution $P(\mathbf{U}|\mathbf{X} = \mathbf{x})$.

Explicitly, given an SCM $\mathcal{M}$, the counterfactual distribution can be estimated by first inferring the posterior over latent variables, e.g., $P(\mathbf{U}|\mathbf{X} = \mathbf{x}, T = 1, Y = 1)$ in our running example, and then passing that distribution through the structural mechanisms in a modified $\mathcal{M}^I$ (e.g., $I = do(T = 0)$) to obtain a counterfactual distribution over any variable[7]. In this way, we make precise the meaning of several terms we will use in this chapter. When we say *counterfactual inference*, we are referring to this process of obtaining a counterfactual distribution. Similarly, we sometimes use the term *counterfactual posterior* to refer to the counterfactual distribution, to reflect the fact that it is simply posterior inference in a particular type of causal model.

[7] This process is called abduction, action, and prediction in Pearl (2009)

Non-Identifiability of Binary SCMs

So, given an SCM $\mathcal{M}$, we can compute an answer to our counterfactual question: Having given the drug and observed that $Y = 1$ (survival), what *would have happened* if we had instead withheld the drug? In the binary case, this corresponds to the *Probability of Necessity* (PN) (Pearl, 2009; Dawid et al., 2015), because it represents the probability that the exposure $T = 1$ was necessary for the outcome.

Intuitively, this is impossible to answer with certainty, even though we may ask ourselves these types of questions frequently in the real world. For instance, in medical malpractice, establishing fault requires just such a counterfactual claim, showing that an injury would not have occurred "but for" the breach in the standard of care (Bal, 2009; Encyclopedia, 2008).

Mathematics matches our intuition in this case: The answer to the question is not identifiable without further assumptions, a general property of counterfactual inference. That is, there are multiple SCMs which are all consistent with the interventional distribution, but which produce different counterfactual estimates of quantities like the Probability of Necessity (Pearl, 2009).

Monotonicity Assumption for Identification of Binary SCMs

Nonetheless, there are plausible (though untestable) assumptions we can make that identify counterfactual distributions. Consider our intuition in the following case: Suppose that a non-smoker develops lung cancer. What would have happened if they had (counterfactually) smoked a pack a day? Our intuition is that, at the very least, it would not have *helped*, and they would have developed the cancer regardless, all else being equal. This type of intuition is formalized mathematically as the *monotonicity assumption* (Pearl, 2000; Tian and Pearl, 2000), and is in fact sufficient to identify the Probability of Necessity and related quantities in epidemiology (Cuellar and Kennedy, 2018; Yamada and Kuroki, 2017).

Definition 2.4 (Monotonicity)**.** A SCM of a binary variable Y is monotonic relative to a binary variable T if and only if it has the following property[8,9]: $\mathbb{E}[Y|do(T=t)] \geq \mathbb{E}[Y|do(T=t')] \implies f_y(t,u) \geq f_y(t',u),\ \forall u$. We can write equivalently that the following event never occurs, in the case where $\mathbb{E}[Y|do(T=1)] \geq \mathbb{E}[Y|do(T=0)]$: $Y_{do(T=1)} = 0 \wedge Y_{do(T=0)} = 1$. Conversely for $\mathbb{E}[Y|do(T=1)] \leq \mathbb{E}[Y|do(T=0)]$, the following event never occurs: $Y_{do(T=1)} = 1 \wedge Y_{do(T=0)} = 0$.

In particular, this assumption restricts the class of possible SCMs to those which all yield equivalent counterfactual distributions over Y. For instance, the following SCM exhibits the monotonicity property, and replicates any interventional distribution where $g(x,t) = \mathbb{E}[Y|X=x, do(T=t)]$:

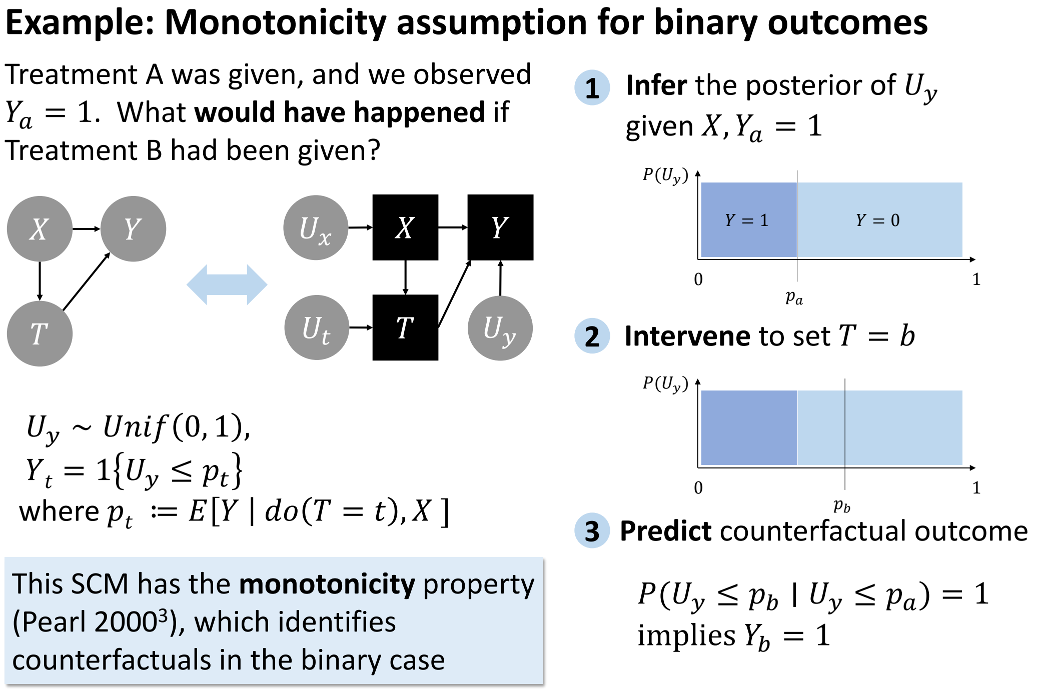
$$Y = \mathbf{1}\left\{U_y \leq g(x,t)\right\}, \quad U \sim \text{Unif}(0,1)$$

In Figure 2-6 we demonstrate how this plays out for a binary treatment and outcome.

There is a wide range of literature in statistics, epidemiology, and machine learning which makes use of this assumption: In epidemiology, it implicitly appears in early work on estimating quantities like the 'relative risk ratio' (Miettinen, 1974), which are often imbued with causal interpretations (Pearl, 2009; Yamada and Kuroki, 2017). Formalizing the assumption of monotonicity, required to correctly impute causal meaning to these quantities, is covered in Balke and Pearl (1994); Pearl (2000); Tian and Pearl (2000). More recent work in epidemiology uses the assumption of monotonicity explicitly, (e.g., to estimate the counterfactual effect of water sanitation in Kenya in Cuellar and Kennedy, 2018), and there has been ample discussion and debate regarding how this reasoning could apply (in principle) to legal cases, such as litigation around the toxic effects of drugs (Dawid et al., 2016). In statistics, monotonicity of treatment with respect to an instrumental is a core assumption of

[8] We could also write this property as conditional on X

[9] This definition differs slightly from the presentation of monotonicity in Pearl (2009), where $f_y(t,u)$ being monotonically increasing in t is given as the property, with the testable implication that $\mathbb{E}[Y|do(T=t)] \geq \mathbb{E}[Y|do(T=t')]$ for $t \geq t'$. Because the direction of monotonicity is only compatible with the corresponding direction of the expected interventional outcomes, we fold this into the definition of monotonicity directly, to align with our later definition of counterfactual stability. Also note that we use the notation $Y_{do(T=t)} \coloneqq f_y(t,u)$ here



Example: Monotonicity assumption for binary outcomes

Treatment A was given, and we observed $Y_a = 1$. What **would have happened** if Treatment B had been given?

$U_y \sim Unif(0,1)$,
$Y_t = 1\{U_y \leq p_t\}$
where $p_t := E[Y \mid do(T = t), X\,]$

This SCM has the **monotonicity** property (Pearl 2000[3]), which identifies counterfactuals in the binary case

1 **Infer** the posterior of U_y given $X, Y_a = 1$

2 **Intervene** to set $T = b$

3 **Predict** counterfactual outcome

$P(U_y \leq p_b \mid U_y \leq p_a) = 1$
implies $Y_b = 1$

Figure 2-6: *Example of a structural causal model which satisfies the monotonicity assumption, and the process of performing counterfactual inference.*

instrumental variable analysis (Imbens and Angrist, 1994). Finally, the monotonicity assumption has been used recently in the machine learning community by Kallus (2019) to classify treatment non-responders.

2.3 Counterfactual Decomposition of Reward

2.3.1 Viewing MDPs and POMDPs as SCMs

In this section we will describe how to reformulate MDPs and POMDPs as structural causal models, retaining their implied interventional distributions while enabling the counterfactual inference procedure described previously. Critically, our results are not limited to MDPs and POMDPs, as any graphical model can be reformulated as a structural causal model. Thus, our results apply more generally wherever e.g., the parametric g-formula is used, but we focus primarily on MDPs and POMDPs in this

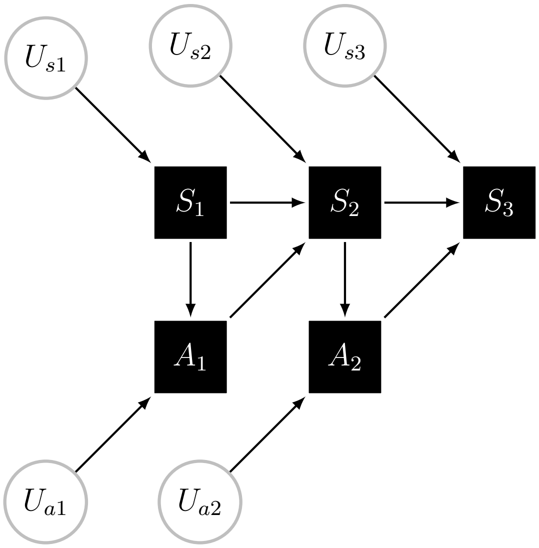


Figure 2-7: *SCM for a MDP, with states S_t and actions A_t, where the action is generated via the mechanism $\pi(U_a, S_t)$, or $\pi(S_t)$ if the policy is deterministic. Rewards are not shown for simplicity. Black squares are functions of their parents in the graph, and are observed, while white circles are unobserved random variables.*

chapter. Note that we will abuse language slightly throughout this thesis, referring to both (a) a structural causal model over all observed variables, as well as (b) the individual mechanisms for each variable (e.g., $S_{t+1} = f_s(s_t, a_t, u_{s_{t+1}})$) as structural causal models.

For a MDP, we can write the states, actions, and rewards as deterministic functions of their parents in the MDP (e.g., for any individual state, these are the previous state and action), as well as an independent exogenous variable. This is shown visually in Figure 2-7. If we are given a deterministic policy to evaluate, then the only SCMs and exogenous variables that we need to consider modelling (for the counterfactual) are those which impact the state transitions (as well as the rewards, if they are not a deterministic function of state). For continuous state-space models, we will need a continuous SCM, as discussed in Section 2.5, and for discrete state-space models (e.g., a finite MDP), we will need a categorical SCM, as discussed in Section 2.4.

Similarly, as noted in Buesing et al. (2019), we can view an episodic Partially Observable Markov Decision Process (POMDP) as an SCM, as shown in Figure 2-8, where S_t corresponds to states, A_t corresponds to actions, O_t corresponds to ob-

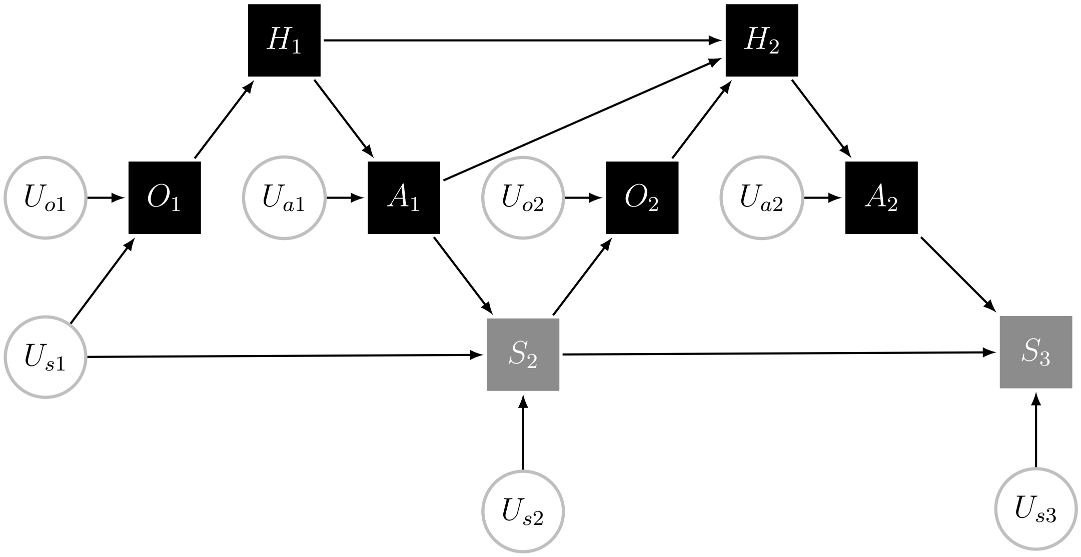


Figure 2-8: *SCM for a POMDP, slightly modified from a similar figure in Buesing et al. (2019), with initial state $U_{s1} = S_1$, states S_t, and histories H_t, where the action is generated via the mechanism $\pi(U_a, H_t)$, or $\pi(H_t)$ if the policy is deterministic. Rewards are captured as part of observed variables O for simplicity. Black and grey squares are functions of their parents in the graph, with black squares being observed and grey squares being unobserved. White circles still represent unobserved variables.*

servable quantities (including reward R_t), H_t contains history up to time t, i.e., $H_t = \{O_1, A_1, \dots A_{t-1}, O_t\}$, and stochastic policies are given by $\pi(a_t|h_t)$.

Thus, the only remaining task required to convert a MDP or POMDP into an SCM is to define the individual mechanisms in such a way that the conditional probability distributions are preserved. This will be discussed in more detail in Sections 2.4-2.5.

For now, we will define some additional notation[10] that will prove useful later, and then discuss in Section 2.3.2 why this reformulation as an SCM is useful for understanding the model-based estimates of reward that a MDP or POMDP might produce. In the context of reinforcement learning with POMDPs, we are typically concerned with estimating the expected reward of a proposed policy $\hat{\pi}$. To formalize notation, a given policy π implies a density over trajectories $\tau \in \mathcal{T} = (S_1, O_1, A_1, \dots, A_{T-1}, S_T, O_T)$, which we denote as $p^\pi(\tau)$, and we let $R(\tau)$ be the total reward of a trajectory τ. For ease of notation, we sometimes write $\mathbb{E}_{\hat{\pi}}$ and $\mathbb{E}_{obs}$ to indicate an expectation taken

[10] We also re-define some notation we used previously, for which we apologize profusely to the reader. From now on τ is a trajectory, not the ATE.

with respect to $\tau \sim p^{\hat{\pi}}$ and $\tau \sim p^{\pi_{obs}}$ respectively, where $\hat{\pi}$ refers to the proposed ('target' or 'evaluation') policy, and π_{obs} to the observed ('behavior') policy.

2.3.2 Counterfactual Decomposition of Reward

Model-Based OPE as CATE Estimation

If we wish to compare the performance of a proposed policy $\hat{\pi}$ and the observed policy π_{obs}, we might compare the difference in expected reward. The expected reward under π_{obs} can be estimated in this case using observed trajectories, without a model of the environment. The difference in expected reward is conceptually similar to the average treatment effect (ATE) of applying the proposed vs observed policy, and we denote it as δ:

$$\delta \coloneqq \mathbb{E}_{\hat{\pi}}[R(\tau)] - \mathbb{E}_{obs}[R(\tau)] \tag{2.8}$$

However, it may be useful to drill down into specific cases: Perhaps there are certain environments, for instance, in which the proposed policy would perform better or worse than the observed policy. One natural decomposition is to condition on the first observed state to estimate a conditional expected reward, e.g.,

$$\delta_o \coloneqq \mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o] - \mathbb{E}_{obs}[R(\tau)|O_1 = o] \tag{2.9}$$

Equation 2.9 corresponds conceptually to CATE estimation, where we condition only on pre-treatment information (in this case, O_1, which occurs before the first action). However, we can go no further than that without a structural causal model, as we need a way to 'condition' on the entire observed trajectory.

Counterfactual OPE as ITE Estimation

Given a structural causal model, we can use information from the entire trajectory to decompose Equation (2.9) further, over actual trajectories that we have observed, to highlight differences between the observed and proposed policy. With an SCM in hand, we can decompose Equation 2.9 further as follows:

Lemma 2.1 (Counterfactual Decomposition of Expected Reward)**.** *Let trajectories τ be drawn from $p^{\pi_{obs}}$. Let $\tau_{\hat{\pi}}$ be a counterfactual trajectory, drawn from our posterior distribution over the exogenous U variables under the new policy $\hat{\pi}$. Note that under the SCM, $\tau_{\hat{\pi}}$ is a deterministic function of the exogenous U variables, so we can write $\tau_{\hat{\pi}}(u)$ to be explicit:*

$$\begin{aligned}&\mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o_1] \\ &= \int_{\tau} p^{\pi_{obs}}(\tau|O_1 = o_1)\mathbb{E}_{u\sim p^{\pi_{obs}}(u|\tau)}[R(\tau_{\hat{\pi}}(u))]d\tau\end{aligned}$$

Proof. This proof is similar to the proof of Lemma 1 from (Buesing et al., 2019), but is spelled out here for the sake of clarity. Recall that the distribution of noise variables U is the same for every intervention / policy. Thus, $p^{\pi_{obs}}(u) = p^{\hat{\pi}}(u) = p(u)$. We will write p' and $\hat{p}$ for $p^{\pi_{obs}}$ and $p^{\hat{\pi}}$ respectively to simplify notation.

Furthermore, recall that all variables are a deterministic function of their parents in the causal DAG implied by the SCM. Most importantly, this means that the trajectory τ is a deterministic function of the policy π and the exogenous variables U. With that in mind, let $\tau_{\hat{\pi}}(u)$ indicate the trajectory τ as a deterministic function of $\hat{\pi}$ and u. We will occasionally use indicator functions to indicate whether or not a deterministic value is compatible with the variables that determine it, e.g., $\mathbf{1}\{\tau|u,\pi\}$ is equivalent to the indicator for $\mathbf{1}\{\tau = \tau_{\pi}(u)\}$. Note that the first observation is independent of the policy, and is just a function of the exogenous U, so we will write $\mathbf{1}\{o_1|u\}$ in that case. For simplicity, we will remove the conditioning on O_1 to start with:

$$\begin{aligned}
&\mathbb{E}_{\hat{p}}[R(\tau)] \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \hat{p}(u) du && (2.10) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot p'(u) du && (2.11) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \left(\int p'(\tau, u) d\tau \right) du && (2.12) \\
&= \int \int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau) \cdot p'(\tau) du d\tau && (2.13) \\
&= \mathbb{E}_{\tau \sim p'} \left[\int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau) du \right] && (2.14) \\
&= \mathbb{E}_{\tau \sim p'} \mathbb{E}_{u \sim p'(u|\tau)} \left[R(\tau_{\hat{\pi}}(u)) \right] && (2.15) \\
&= \int_{\tau} p^{\pi_{obs}}(\tau) \mathbb{E}_{u \sim p'(u|\tau)} \left[R(\tau_{\hat{\pi}}(u)) \right] d\tau && (2.16)
\end{aligned}$$

In step (2.10) we are just using the definition of the expectation under $\hat{p}$, along with the notation $\tau_{\hat{\pi}}(u)$ to indicate that the trajectory is a deterministic function of the exogenous u and the policy $\hat{\pi}$. In step (2.11) we replace $\hat{p}(u)$ with $p'(u)$ because they are equivalent, as noted earlier. In step (2.12) we expand $p'(u)$ over possible trajectories τ arising from the observed policy. In step (2.13) we rearrange terms and swap the order of the integral, and in step (2.14) we rewrite the outer integral as an expectation. In step (2.15) we further condense notation, and then expand in step (2.16) to match the notation in the Lemma. If we introduce the conditioning on

O_1, we see that it is substantively the same.

$$
\begin{aligned}
&\mathbb{E}_{\hat{p}}[R(\tau)|o_1] \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \mathbf{1}\{o_1|u\} \cdot \hat{p}(u)du && (2.17) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \mathbf{1}\{o_1|u\} \cdot p'(u)du && (2.18) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|o_1)du && (2.19) \\
&= \int R(\tau_{\hat{\pi}}(u)) \cdot \left(\int p'(\tau, u|o_1)d\tau\right) du && (2.20) \\
&= \int\int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau) \cdot p'(\tau|o_1)dud\tau && (2.21) \\
&= \int p'(\tau|o_1)\left[\int R(\tau_{\hat{\pi}}(u)) \cdot p'(u|\tau)du\right] d\tau && (2.22) \\
&= \int_\tau p'(\tau|o_1)\mathbb{E}_{u\sim p'(u|\tau)}[R(\tau_{\hat{\pi}}(u))]d\tau && (2.23)
\end{aligned}
$$

The main difference in this case is that is just that we carry the indicator into the prior on U at step (2.19), which we can do because O_1 does not depend on the policy that is applied. Note that Equation (2.23) matches the statement of the Lemma. □

Corollary 2.1 (Counterfactual Decomposition of δ_o).

$$
\begin{aligned}
\delta_o &:= \mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o_1] - \mathbb{E}_{obs}[R(\tau)|O_1 = o_1] \\
&= \int_\tau p^{\pi_{obs}}(\tau|O_1 = o_1)\mathbb{E}_{u\sim p^{\pi_{obs}}(u|\tau)}[R(\tau_{\hat{\pi}}(u)) - R(\tau)]d\tau
\end{aligned}
$$

Proof. By Lemma 2.1, we have it that

$$
\begin{aligned}
\delta_o &:= \mathbb{E}_{\hat{\pi}}[R(\tau)|O_1 = o] - \mathbb{E}_{obs}[R(\tau)|O_1 = o] \\
&= \int_\tau p'(\tau|o_1)\mathbb{E}_{u\sim p'(u|\tau)}[R(\tau_{\hat{\pi}}(u))]d\tau \\
&\quad - \int_\tau p'(\tau|o_1)\mathbb{E}_{u\sim p'(u|\tau)}[R(\tau_{\pi_{obs}}(u))]d\tau \\
&= \int_\tau p^{\pi_{obs}}(\tau|O_1 = o_1)\mathbb{E}_{u\sim p^{\pi_{obs}}(u|\tau)}[R(\tau_{\hat{\pi}}(u)) - R(\tau)]d\tau
\end{aligned}
$$

Note that in the last step, we recognize that $\mathbb{P}_{u\sim p'(u|\tau)}[\tau_{\pi_{obs}}(u) = \tau] = 1$, because the posterior density over u is zero for all u such that $\tau_{\pi_{obs}}(u) \neq \tau$. □

Corollary 2.1 implies that we can decompose the expected difference in reward between the policies into differences on *observed episodes* over counterfactual trajectories, if the SCM is correct. In the context of Buesing et al. (2019), this fact is used to argue that counterfactuals approximate draws from the interventional distribution, since efficient estimation of the latter is their ultimate goal.

In our case this fact serves an additional purpose: It *theoretically motivates the use of counterfactuals as a model-checking tool.* In principle, if the SCM is correct, then the counterfactuals can be used to identify how observed episodes contribute to overall estimates of reward, and thus ground the model-based conclusions in specific counterfactual claims that can be vetted by domain experts. In practice, we consider this decomposition a heuristic, as we do not believe the SCM is necessarily correct. That said, our empirical work in Sections 2.6-2.7 gives anecdotal evidence that this equality holds approximately in some situations when the learned MDP is not correct.

2.4 Gumbel-Max SCMs for Categorical Variables

In the previous chapter, we illustrated how to convert a model of the environment into a structural causal model, as well as the potential benefits of doing so for the purpose of decomposing model-based rewards into counterfactual claims. All that remained was to specify the specific causal mechanisms for each of the variables in the respective SCMs.

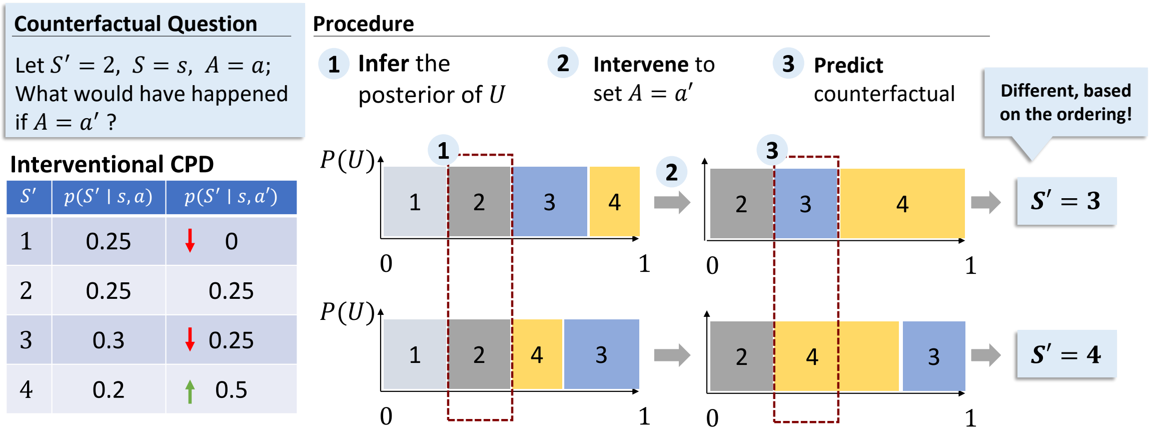
However, it is at this point that we face a non-identifiability issue: Multiple SCMs can all entail the same interventional distribution, but a different set of counterfactual trajectories, and therefore a different decomposition under Lemma 2.1. This motivates the theoretical work of this chapter: We must make our assumptions carefully, as they cannot be tested by data, so it is worth investigating which assumptions are consistent with our causal intuition. We illustrate this non-identifiability (with respect to categorical distributions) in Section 2.4.1. Then we introduce the condition of *counterfactual stability* (in Section 2.4.2) for a discrete distribution on k categories, and show that it is compatible with the monotonicity condition of Pearl (2000) in that it implies the monotonicity assumption when $k = 2$. Then we introduce the Gumbel-Max SCM for discrete variables in Section 2.4.3, and prove that it satisfies the counterfactual stability condition, and in Section 2.4.4 describe an intuitive connection to discrete choice models.

2.4.1 Non-Identifiability of Categorical SCMs

We will first illustrate that the non-identifiability of counterfactual distributions applies to categorical distributions as well. Consider the categorical distribution over k categories, e.g., the transition kernel $P(S'|S = s, A = a)$ over discrete states. Let $p_i \coloneqq P(S' = i|S = s, A = a)$. There are multiple ways that we could sample from this distribution with a structural mechanism f and latent variables U. For instance, we could define an ordering **ord** on the categories, and define k intervals of $[0, 1]$ as $[0, p_{\mathbf{ord}(1)}), [p_{\mathbf{ord}(1)}, \sum_{i=1}^{2} p_{\mathbf{ord}(i)}), \dots, [\sum_{i=1}^{k-1} p_{\mathbf{ord}(i)}, 1]$. Then we could draw $U \sim Unif(0, 1)$, and return the interval that u falls into.

However, different permutations **ord** will yield equivalent interventional distributions but can imply different counterfactual distributions. For instance, consider the following example, shown visually in in Figure 2-9. Let $k = 4$ and $p_1 = p_2 = 0.25, p_3 = 0.3, p_4 =$

0.2 and consider an intervention $A = a'$ which defines a different distribution $p'_1 = 0, p'_2 = 0.25, p'_3 = 0.25, p'_4 = 0.5$. Now consider two permutations, **ord** $= [1, 2, 3, 4]$ and **ord**$' = [1, 2, 4, 3]$, and the counterfactual distribution under a' given that $S' = 2, A = a$. In each case, posterior inference over U implies that $P(U|S' = 2, S = s, A = a) \sim Unif[0.25, 0.5)$. However, under **ord** this implies the counterfactual $S' = 3$, while under **ord**$'$ it implies $S' = 4$.



Counterfactual Question

Let $S' = 2,\ S = s,\ A = a$;
What would have happened if $A = a'$?

Interventional CPD

S'	$p(S' \mid s, a)$	$p(S' \mid s, a')$
1	0.25	↓ 0
2	0.25	0.25
3	0.3	↓ 0.25
4	0.2	↑ 0.5

Figure 2-9: *Example of non-identifiability of categorical counterfactual outcomes. The table on the bottom right illustrates the difference in the conditional probability distribution (the 'interventional' distribution) as a function of actions a versus a'. The procedure is illustrated in the middle, where the two rows represent two possible orderings (**ord** and **ord'**) both of which define a causal mechanism $S' = f(S, A, U)$ with $U \sim Unif(0, 1)$ that replicates the interventional probability distribution. From left to right, we see the application of counterfactual inference: (1) Infer the posterior of U, represented by the red box, (2) intervene to set $A = a'$, and (3) predict the counterfactual by evaluating under the posterior of U. These two SCMs produce different counterfactual outcomes, with the outcome of $S' = 3$ being particularly unintuitive, since the interventional probability was reduced under the shift from a to a'.*

Note that in this example, the mechanism $f_{\textbf{ord}}$ implied a non-intuitive counterfactual outcome: Even though the intervention $A = a'$ *lowered* the probability of $S' = 3$ (relative to the probability under $A = a$) without modifying the probability of $S' = 2$, it led to a delta distribution in the counterfactual posterior on $S' = 3$. Since all choices for **ord** imply the same interventional distribution, there is no way to distinguish between these mechanisms with data.

This motivates the following sections, where we posit a desirable property for categorical

SCMs to possess, and which rules out this result (among others) and is compatible with the notion of monotonicity introduced by Pearl (2000). We then demonstrate that a mechanism based on sampling independent Gumbel variables satisfies this property, which motivates the use of the Gumbel-Max SCM for performing counterfactual inference in this setting.

2.4.2 Counterfactual Stability Property

We now introduce our first contribution, the desired property of *counterfactual stability* for categorical SCMs with k categories, laid out in in Definition 2.5. This property would rule out the non-intuitive counterfactual implications of $f_{\mathbf{ord}}$ in Section 2.4.1. We then demonstrate that this condition implies the monotonicity condition when $k = 2$.

First, with apologies to the reader, we will once again introduce some notation. Denote the interventional probability distribution of a categorical variable Y with k categories as $P^{\mathcal{M};I}(Y) = \mathbf{p}$ under intervention I, and $\mathbf{p}'$ under intervention I', where $\mathbf{p}, \mathbf{p}' \in \Delta^k$, the probability simplex over k categories. To simplify notation for interventional outcomes, we will sometimes denote by Y_I the observed outcome Y under intervention I, and denote by $Y_{I'}$ the counterfactual outcome under intervention I', such that p_i and $P(Y_I = i)$ are both equivalent to $P^{\mathcal{M};I}(Y = i)$, and similarly for I'. For counterfactual outcomes, we will write $P^{\mathcal{M}|Y_I=i;I'}(Y)$ for the counterfactual distribution of Y under intervention I' given that we observed $Y = i$ under the intervention I.

Definition 2.5 (Counterfactual Stability)**.** A SCM of a categorical variable Y satisfies *counterfactual stability* if it has the following property: If we observe $Y_I = i$, then for all $j \neq i$, the condition $\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j}$ implies that $P^{\mathcal{M}|Y_I=i;I'}(Y = j) = 0$. That is, if we observed $Y = i$ under intervention I, then the counterfactual outcome under I' cannot be equal to $Y = j$ unless the multiplicative change in p_i is less than the multiplicative change in p_j.

Corollary 2.2. *If $\mathcal{M}$ is a SCM which satisfies counterfactual stability, then if we* *observe $Y_I = i$, and $\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j}$ holds for all $j \neq i$, then $P^{\mathcal{M}|Y_I=i;I'}(Y = i) = 1$.*

This definition and corollary encode the following intuition about counterfactuals: If we had taken an alternative action that would have *only increased* the probability of $Y = i$, without increasing the likelihood of other outcomes, then the same outcome would have occurred in the counterfactual case. Moreover, in order for the outcome to be different under the counterfactual distribution, the relative likelihood of an alternative outcome must have increased relative to that of the observed outcome. The connection to monotonicity is given in Theorem 2.1, whose proof is deferred to Section 2.4.5.

Theorem 2.1. *Let $Y = f_y(t, u)$ be the SCM for a binary variable Y, where T is also a binary variable. If this SCM satisfies the counterfactual stability property, then it also satisfies the monotonicity property with respect to T.*

2.4.3 Gumbel-Max SCMs Satisfy Counterfactual Stability

Unlike monotonicity with binary outcomes and treatments, the condition of counterfactual stability does not obviously imply any closed-form solution for the counterfactual posterior. Thus, we introduce a specific SCM which satisfies this property, and discuss how to sample from the posterior distribution in a straightforward fashion. We start by recalling the following fact, known as the Gumbel-Max trick (Luce, 1959; Yellott, 1977; Yuille. and L, 2011; Hazan and Jaakkola, 2012; Maddison et al., 2014; Hazan et al., 2016; Maddison et al., 2017):

Definition 2.6 (Gumbel-Max Trick)**.** We can sample from a categorical distribution with k categories as follows, where $\tilde{p}_i$ is the unnormalized probability $P(Y = i)$: First, draw $g_1, \dots, g_k$ from a standard Gumbel, which can be achieved by drawing $u_1, \dots, u_k$ iid from a Unif$(0, 1)$, and assigning $g_i = -\log(-\log u_i)$. Then, set the outcome j by taking $\arg\max_j \log \tilde{p}_j + g_j$.

Clearly, we can perform this for any categorical distribution, e.g., the transition distribution $p_i = P(S' = i|S, A)$; In particular, for any discrete variable Y whose parents in a causal DAG are denoted $\mathbf{X}$, a *Gumbel-Max SCM* assumes the following causal mechanism, where $\mathbf{g} = (g_1, \ldots, g_k)$ are independent Gumbel variables:

$$Y = f_y(\mathbf{x}, \mathbf{g}) \coloneqq \arg\max_j \{\log P(Y = j | \mathbf{X} = \mathbf{x}) + g_j\}$$

Like any mechanism which replicates the conditional distribution under intervention, this mechanism is indistinguishable from any other causal mechanism based on data alone. That said, it does satisfy the property given in Definition 2.5.

Theorem 2.2. *The Gumbel-Max SCM satisfies the counterfactual stability condition.*

The intuition is that, when we consider the counterfactual distribution, the Gumbel variables are fixed. Thus, in order for the argmax (our observed outcome) to change in the counterfactual, the log-likelihood of an alternative outcome must have increased relative to our observed outcome.

We note that posterior inference in the Gumbel-Max SCM is straightforward. Given a Gumbel-Max SCM as defined above, where $Y = \arg\max_j \log p_j + g_j$ and $p_j \coloneqq P(Y_I = j)$, we can draw Monte Carlo samples from the posterior $P(\mathbf{g} | Y_I = i)$ using one of two approaches: First, we can use rejection sampling, drawing samples from the prior $P(\mathbf{g})$ and rejecting those where $i \neq \arg\max_j \log p_j + g_j$. Alternatively, it is known (Maddison et al., 2014; Maddison and Tarlow, 2017) that in the posterior, the maximum value and the argmax of the shifted Gumbel variables $\log p_j + g_j$ are independent, and the maximum value is distributed as a standard Gumbel (in the case of normalized probabilities). Thus, we can sample the maximum value first, and then sample the remaining values from shifted Gumbel distributions that are truncated at this maximum value. Then, for each index j, subtracting off the location parameter $\log p_j$ will give us a sample of g_j. We can then add this sample $\mathbf{g}$ to the log-probabilities under I' (i.e., $\log \mathbf{p}'$) and take the new argmax to get a sample of the counterfactual outcome Y under intervention I'.

2.4.4 Intuition: Connection to Discrete Choice Models

The Gumbel-Max sampling mechanism was initially introduced in the discrete-choice literature (Luce, 1959), where it is used as a generative model for decision-making under utility maximization (Train, 2002; Aguirregabiria and Mira, 2010), where the log probabilities may be assumed to follow some functional form, such as being linear in features. This is motivated by understanding the impact of different characteristics on consumer choices, see (Aguirregabiria and Mira, 2010, Example 1).

We discuss this connection further in this section, but note the contrast with our approach: Whereas the traditional discrete choice literature assumes a particular functional form (e.g., linear in features) for the log probabilities, we decouple this structural mechanism (for estimation of counterfactuals) from the statistical model used to estimate the conditional probability distributions under interventions. We encourage the reader to consult (Train, 2002, e.g., Chapter 2) for more details, but we highlight some relevant pieces of intuition below, with their connection to the counterfactual stability condition.

Discrete choice models that utilize Gumbel noise are known in the econometrics literature as *logit discrete choice models*, and are part of a broader class of discrete-choice models which are derived on the principle of utility maximization, known as *random utility models*. This literature is motivated by consumers as decision-makers, deciding between different discrete alternatives. In the context of modelling state transitions in an MDP, we can make the analogy that the 'decision-maker' is nature, and the choice is the next discrete state. First, we introduce two core assumptions: The concept of random utility maximization, which is introduced as a core assumption of discrete-choice models in (Train, 2002), and the assumption of additive separability.

Random Utility Maximization We assume that the decision-maker acts to optimize utility. In particular, the decision-maker associates some utility U_i with each discrete choice / alternative i, and chooses the alternative i if and only if $U_i > U_j \ \forall i \neq j$. Because U is not observed directly, we treat it as a random variable. We only observe the conditional probability distribution on Y, known as the *conditional choice probability*, given by

$$P(Y = i|X) = \int \mathbf{1}\left\{U_i > U_j, \forall j \neq i\right\} p(U|X) dU \tag{2.24}$$

Additive Separability Without loss of generality, the utility U can be re-written in terms of a deterministic component which depends on observable variables X, and an unobserved component ϵ, so that $U_j = V_j + \epsilon_j$, where V is assumed to be a function of observable variables, and is called the *representative utility.* With that in mind, Equation (2.24) can be rewritten as

$$\begin{aligned} &P(Y = i|X) \\ &= \int \mathbf{1}\left\{V_i(x) + \epsilon_i > V_j(x) + \epsilon_i, \forall j \neq i\right\} p(\epsilon|X) d\epsilon \end{aligned} \tag{2.25}$$

The assumption of *additive separability* states that the unobserved components ϵ are independent of the observed components, i.e., $\epsilon \perp\!\!\!\perp X$. Tying these assumptions back to our proposed notion of counterfactual stability, the implication from a counterfactual perspective is that if we intervene on the variables X, we do not change the values of ϵ as a result of additive separability. Thus, the assumption of random utility maximization implies that if we observe $Y_x = i$, then a necessary condition for substituting j for i is that

$$V(x')_j - V(x)_j > V(x')_i - V(x)_i \tag{2.26}$$

Different choices of discrete-choice models imply different functional forms for V and different distributions on ϵ. In the logit model, the ϵ_i variables are assumed to be drawn iid (over alternatives i) from a Gumbel distribution (also known as a Type 1 Extreme Value distribution). This implies that

$$P(Y = i|X) = \frac{\exp V_i(x)}{\sum_j \exp V_j(x)} \tag{2.27}$$

Because any scaling or shifting of the utility is irrelevant, we can set the scale of V such that $V_i = \log p_i$, consistent with Equation (2.27), and see that Equation (2.26) corresponds to the counterfactual stability condition.

2.4.5 Appendix: Proofs

Theorem 2.1. Let $Y = f_y(t, u)$ be the SCM for a binary variable Y, where T is also a binary variable. If this SCM satisfies the counterfactual stability property, then it also satisfies the monotonicity property with respect to T.

Proof. To simplify notation further, let $p^{t=1} := P(Y = 1|do(T = 1))$, $p^{t=0} := P(Y = 1|do(T = 0))$, and let $Y_t := Y_{do(T=t)}$. Without loss of generality, assume that $p^{t=1} \geq p^{t=0}$.

To show that counterfactual stability implies monotonicity, we want to show that the probability of the event $(Y_1 = 0 \wedge Y_0 = 1)$ is equal to zero. We will do so by proving both cases: First that $P^{\mathcal{M}|Y_0=1;do(T=1)}(Y = 0) = 0$ and second that $P^{\mathcal{M}|Y_1=0;do(T=0)}(Y = 1) = 0$. We can start with the assumption that $p^{t=1} \geq p^{t=0}$ and write:

$$\begin{aligned} p^{t=1} &\geq p^{t=0} \\ \implies p^{t=1}(1 - p^{t=0}) &\geq p^{t=0}(1 - p^{t=1}) \\ \implies \frac{p^{t=1}}{p^{t=0}} &\geq \frac{(1 - p^{t=1})}{(1 - p^{t=0})} \end{aligned}$$

Using the counterfactual stability condition, the last inequality implies that if we observe $Y_0 = 1$, then the counterfactual probability of $Y_1 = 0$ is equal to $P^{\mathcal{M}|Y_0=1;do(T=1)}(Y = 0) = 0$, as desired. For the second case, where we observe $Y_1 = 0$, we can simply manipulate the inequality to see that

$$\frac{(1 - p^{t=0})}{(1 - p^{t=1})} \geq \frac{p^{t=0}}{p^{t=1}}$$

Which yields the conclusion that $P^{\mathcal{M}|Y_1=0;do(T=0)}(Y=1)=0$, as desired, completing the proof. □

Theorem 2.2. The Gumbel-Max SCM satisfies the counterfactual stability condition.

Proof. Recall that we write the shorthand $p_i := P^{\mathcal{M};I}(Y=i)$, and $p'_i := P^{\mathcal{M};I'}(Y=i)$. Suppose that Y is generated from a Gumbel-Max SCM $\mathcal{M}$ under intervention I, and we observe that $Y_I = i$. The Gumbel-Max SCM implies that almost surely:

$$\log p_i + g^{(i)} > \log p_j + g^{(j)} \quad \forall j \neq i \tag{2.28}$$

To demonstrate that the Gumbel-Max SCM satisfies the counterfactual stability condition, we need to demonstrate that $\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j} \implies P^{\mathcal{M}|Y_I=i;I'}(Y=j)=0$ for all $j \neq i$.

We will proceed by proving the contrapositive, that for all $j \neq i$, $P^{\mathcal{M}|Y_I=i;I'}(Y=j) \neq 0 \implies \frac{p'_i}{p_i} < \frac{p'_j}{p_j}$.

Fix some index $j \neq i$. The condition $P^{\mathcal{M}|Y_I=i;I'}(Y=j) \neq 0$ implies that there exist values $g^{(i)}, g^{(j)}$ such that

$$\log p'_i + g^{(i)} < \log p'_j + g^{(j)} \tag{2.29}$$

Because the Gumbel variables $g^{(i)}, g^{(j)}$ are fixed across interventions, this implies there exist values for these variables which satisfy both inequalities (2.28) and (2.29). Thus, we proceed by subtracting inequality (2.28) from inequality (2.29), maintaining the direction of the inequality and cancelling out the Gumbel terms. The rest is straightforward manipulation using the monotonicity of the logarithm.

$$\log p'_i - \log p_i < \log p'_j - \log p_j$$
$$\log(p'_i/p_i) < \log(p'_j/p_j)$$
$$(p'_i/p_i) < (p'_j/p_j)$$

This demonstrates that $P^{\mathcal{M}|Y_I=i;I'}(Y=j) \neq 0 \implies (p'_i/p_i) < (p'_j/p_j)$ as desired, and taking the contrapositive completes the proof. □

2.5 SCMs with Additive Noise for Continuous Variables

In this brief chapter, we collect some thoughts on structural causal models that reflect the conditional probability distribution of continuous random variables. Although this is not the primary focus of this chapter, we include it here for completeness, as a reference for how the conceptual ideas of this thesis (e.g., decomposition of reward and investigation of counterfactual trajectories) can be applied in the continuous setting.

In contrast to the categorical case, we do not have specific non-identifiability examples for continuous SCMs, nor do we have a corresponding assumption, analogous to counterfactual stability, which suggests specific SCMs for this case. However, we note that a common model assumed in this case takes the form of Equation 2.30, where the next state $s_{t+1} \in \mathbb{R}^n$ is assumed to follow a Gaussian distribution conditioned on the previous state $s_t \in \mathbb{R}^n$ and action $a \in \mathcal{A}$, and the mean and covariance are determined by arbitrary functions $\mu_\theta : \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}^n$ and $\Sigma_\theta : \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}^{n\ timesn}$ of the previous state and action. For instance, in Chua et al. (2018), these are parameterized by neural networks, and Σ_θ is a diagonal covariance.

$$\mathbb{P}(s_{t+1} \mid s_t, s_t) = \mathcal{N}(\mu_\theta(s_t, a_t), \Sigma_\theta(s_t, a_t)) \tag{2.30}$$

This particular model can be re-written equivalently as the following SCM with additive noise that is drawn independently at each time step, where we write L_θ as the Cholesky decomposition of Σ_θ such that $L_\theta L_\theta^T = \Sigma_\theta$. In the case where Σ_θ is a diagonal covariance, as in Chua et al. (2018), this is simply the element-wise square root of Σ_θ.

$$s_{t+1} = \mu_\theta(s_t, a_t) + L_\theta(s_t, a_t) \cdot \epsilon_t \tag{2.31}$$

$$\epsilon_t \sim \mathcal{N}(0, I_n) \tag{2.32}$$

In a similar fashion, many models of dynamics used for reinforcement learning in continuous state spaces can be re-formulated as structural causal models with additive noise that follows some known distribution. Moreover, if the only source of stochasticity in the transitions is an additive term which is an invertible function of ϵ_t, as in Equation 2.31, then counterfactual inference is particularly simple, as the exogenous term ϵ_t can be identified exactly from the observable prediction error.

2.6 Illustrative Applications with Synthetic Data

In this chapter, we develop some intuition for how counterfactuals could be used in practice, using some illustrative applications. First, in Section 2.6.1 we use a toy example of a 2D gridworld to illustrate the differences between counterfactual trajectories and model-based trajectories. Then we give an illustrative example of how counterfactuals could be used to 'debug' a policy in Section 2.6.2, using a synthetic environment of sepsis management. We note that all code required to replicate these synthetic experiments will be made available at https://github.com/clinicalml/cf-policy-introspection.

2.6.1 Building Intuition: 2D Gridworld

To illustrate the concepts behind counterfactual trajectories, we start with a simple 2D example, inspired by a similar experimental setup in (Gottesman et al., 2019b).[11] In Section 2.6.1, we describe the simulator setup, and in Section 2.6.1 we demonstrate how counterfactual inference proceeds in this setting. Finally, we show in Section 2.6.1

[11] We thank Omer Gottesman for providing the original code used in his work

how this enables us to decompose differences in reward (between a target and behavior policy) across individual episodes.

As an addendum, in Section 2.6.1 we demonstrate how counterfactuals take maximum advantage of the information present in a trajectory, by making inferences over all sources of variation, not only a single per-trajectory latent variable.

Simulator Setup

In this example, the agent is navigating a 2D domain, with state $s \in [0,1]^2$ and four possible actions $a \in \{[0,0.1],[0.1,0],[0,-0.1],[-0.1,0]\}$ corresponding to the four cardinal directions (north, east, south, and west). The goal of the agent is to reach the *goal region* $G = \{(x,y) : x \in [0.9,1.0], y \in [0.9,1.0]\}$, and the reward is -1 at each time point until the agent enters the goal region, when it receives a reward of $+10$. The dynamics are as follows

$$s_{t+1} = s_t + a_t + w(s_t;\beta) + \epsilon_t$$

Where $\epsilon_t \sim \mathcal{N}(0, I\sigma_\epsilon^2)$ represents time-varying gusts of wind, and $w(s_t;\beta) = [-\beta \cdot y_t, 0]$ is a cross-wind which pushes in either the western or eastern direction, with a magnitude that increases as the agent progresses north. The β parameter is drawn uniquely for each instance from a Gaussian $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$. We will refer to this as the *prior* on β, but we note that this just represents the *population-level* distribution of β, and could be the posterior population distribution after many trajectories have been observed. We call it a prior to distinguish from the counterfactual posterior over the particular β in each trajectory, which we will seek to infer as part of performing counterfactual inference. Thus, the entire generative model is given by the following, where $\pi(s_t)$ is a deterministic policy which we describe in the next section.

$$\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \tag{2.33}$$

$$\epsilon_t \sim \mathcal{N}(0, I\sigma_\epsilon^2) \tag{2.34}$$

$$a_t = \pi(s_t) \tag{2.35}$$

$$s_{t+1} = s_t + a_t + w(s_t; \beta) + \epsilon_t \tag{2.36}$$

We can view the trajectories as arising from the POMDP / Structural Equation Model given in Figure 2-10, where we leave out the rewards for the sake of simplifying exposition.

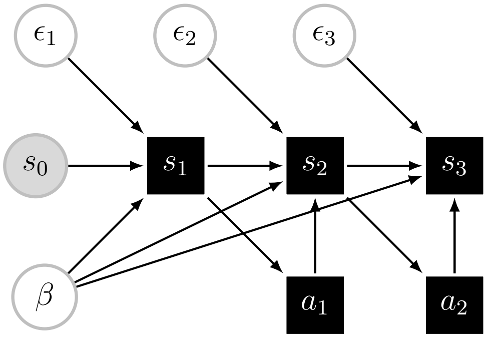
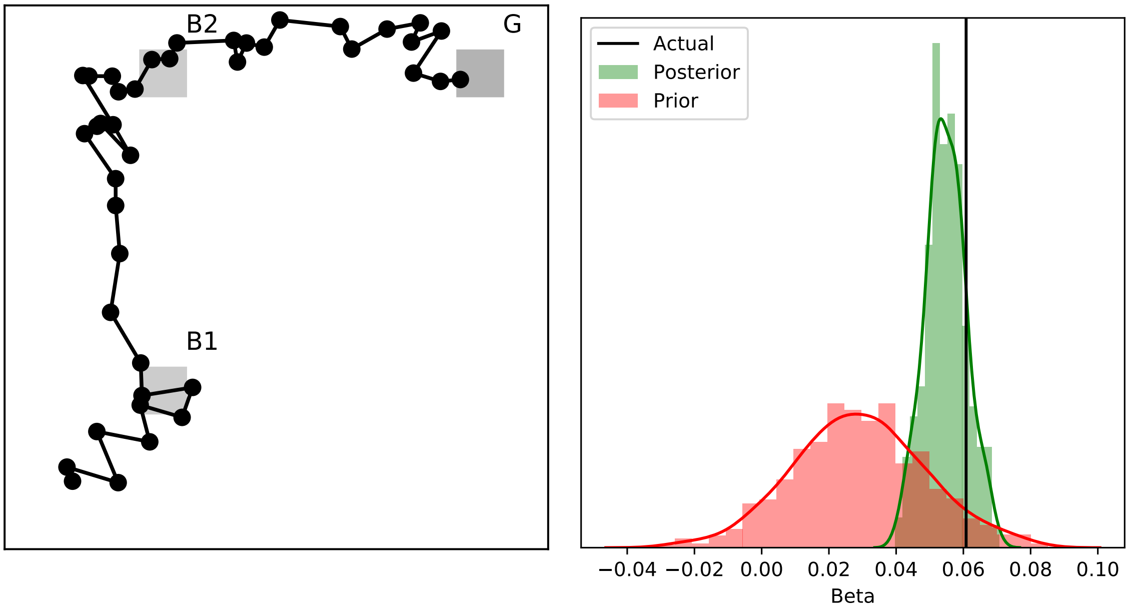


Figure 2-10: *The structural causal model model for our 2D sequences, where each black box is a deterministic function of its parents, and the initial state s_0 is an observed random variable. In practice, all of our sequences start at the same position, so s_0 is a deterministic value.*

Generating Counterfactual Trajectories

In Figure 2-11a we plot a trajectory from this model, which we will use as a running example, where $\sigma_\epsilon = 0.001, \mu_\beta = 0.03, \sigma_\beta = 0.02$. This trajectory follows a myopic behavior policy $\pi_b(s_t)$, which is defined with respect to a series of 'checkpoint' regions that the agent must enter before heading to the goal region in the top right, and at each time point it takes the action which will minimizes the ℓ_2 distance between a naive prediction $s'_{t+1} = s_t + a_t$ and the center of the next region. In this case, the policy π_b seeks to traverse the regions denoted B1, B2 before seeking the region denoted G.



(a) *An observed (factual) trajectory where $\beta = 0.061$, which traverses the regions B1, B2 before seeking the goal region G*

(b) *Prior versus posterior distribution over the value of β for this specific instance*

Figure 2-11: *Factual trajectory and posterior over latent variable β*

The particular draw of β in this case is 0.061.

In this setting, counterfactual inference starts with posterior inference over β, ϵ, which factorizes as

$$p(\beta, \epsilon | \mathbf{x}, \mathbf{y}, \mathbf{a}) = p(\epsilon | \beta, \mathbf{x}, \mathbf{y}, \mathbf{a}) p(\beta | \mathbf{x}, \mathbf{y}, \mathbf{a}).$$

Thus, the first step is posterior inference over β, the results of which are given in Figure 2-11b using MCMC.[12] Note that once we draw a sample of β from the posterior, we can uniquely identify ϵ from Equation 2.36, so the only uncertainty in the counterfactual is due to β.

The advantage of the counterfactual approach is that it allows us to associate a set of counterfactual trajectories with every factual trajectory. This is demonstrated in Figure 2-12a, where we generate counterfactual trajectories under a target policy π_t which seeks to traverse a different set of checkpoints (T1, T2) before heading to the goal region G. We make two notes about the counterfactual trajectories,

[12] We use Pyro (Bingham et al., 2018) to perform MCMC.

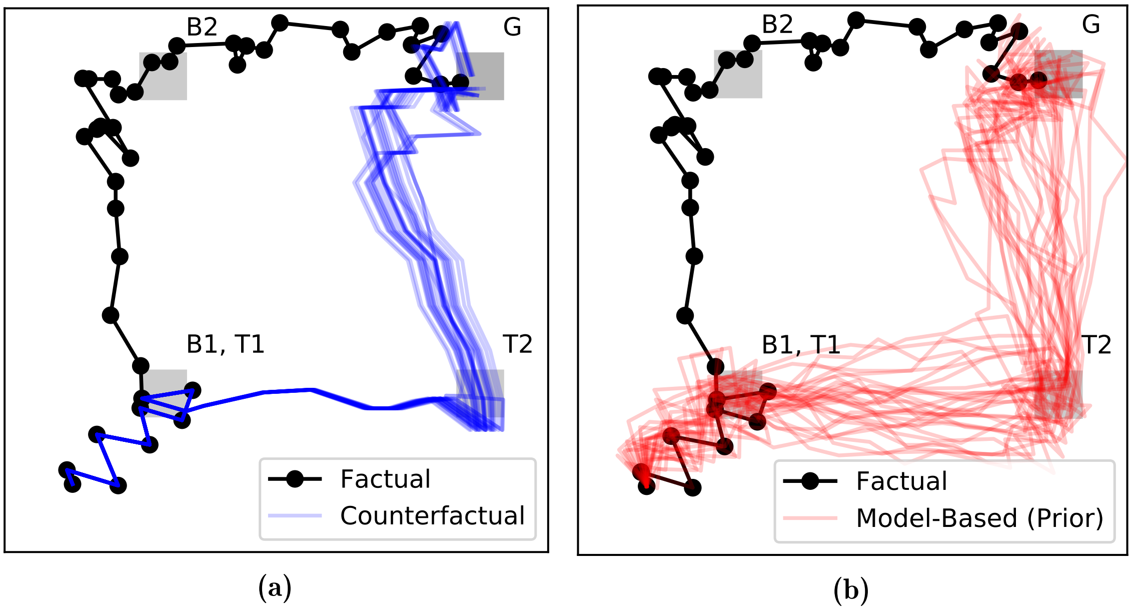


Figure 2-12: *In both cases, the target policy $\pi_t(s_t)$ is used, which seeks to pass through checkpoints T1 and T2 before proceeding to the goal region G. In (a) we see 30 trajectories from the counterfactual posterior, which can be contrasted with (b) where we see 30 trajectories sampled from the generative model given by Equations (2.33-2.36), starting from the same point.*

contrasting them with model-based trajectories in Figure 2-12b (generated using the model given by Equations (2.33-2.36), starting at the same point): First, the counterfactual trajectories are identical to the factual trajectory up until the checkpoint B1/T1, because both policies take the same actions up until that point. Second, the counterfactual trajectories have much less variation because they incorporate all the information from the original trajectory (including both time-dependent and time-independent variation) through the posterior, whereas a model-based roll-out starting from the same point does not.

Decomposition of Reward via Counterfactuals

We can use these counterfactuals to associate with each factual trajectory an expected reward under the counterfactual ‘had the target policy been used instead’, and use this to examine where those differences are projected to be largest. Figure 2-13 demonstrates this over 100 factual trajectories (which follow the behavior policy) and their expected counterfactual reward (under the target policy)[13]. We plot the factual reward observed against the counterfactual reward[14], and shade each point according to the expected value of β under the posterior. This visually demonstrates that the difference in reward is greatest for larger values of β, but does so in a way that can be tied back to individual episodes.

In a real-data application, this type of analysis can be done in an exploratory fashion, to (a) search for trajectories where the difference in reward is estimated to be largest, and (b) examine what differentiates those trajectories from the others.

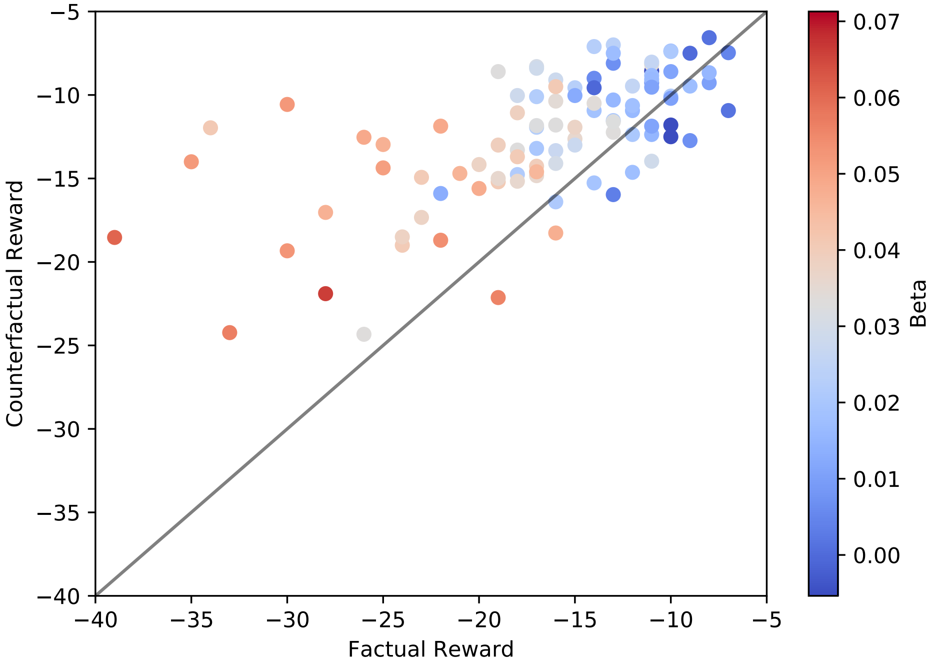


Figure 2-13: *Decomposition of reward*

Addendum: Counterfactual vs. Model-Based Trajectories

In this section, we demonstrate how counterfactuals take maximum advantage of the information present in a trajectory, by making inferences over all sources of variation. In this case, we make inferences over both β and ϵ, and this allows us to draw a

[13] We used 30 counterfactual trajectories for each factual trajectory, in order to compute the expected counterfactual reward.

[14] One point is excluded from the plot, with a factual / counterfactual reward of approximately -60 and -18 respectively.

contrast with two other ways that, conceptually, we could have generated trajectories from the same model.

1. **Model-based roll-out**: Sample a new $\beta \sim p(\beta)$, and then sample a new $\epsilon_t \sim p(\epsilon)$ at each time step. Given a deterministic policy, these random parameters imply a fixed trajectory.

2. **Model-based roll-out (posterior on β)**: Sample $\beta \sim p(\beta|\mathbf{s}, \mathbf{a})$, and then sample a new $\epsilon_t \sim p(\epsilon)$ at each time step.

3. **Counterfactual roll-out**: Sample $\beta, \boldsymbol{\epsilon} \sim p(\beta, \boldsymbol{\epsilon}|\mathbf{s}, \mathbf{a})$, which in this model is equivalent to sampling a value for β from the posterior, and then inferring the unique value of ϵ_t for each time step.

To demonstrate the differences between these approaches, we use two environments: The first environment is the same as the one described above (with $\sigma_\epsilon = 10^{-3}, \mu_\beta = 0.03, \sigma_\beta = 0.02$), and the second has a lower variance over ϵ but a higher prior variance on β (with $\sigma_\epsilon = 10^{-4}, \mu_\beta = 0.03, \sigma_\beta = 0.04$). A single trajectory is sampled from each environment[15], and are given in Figure 2-14, along with the resulting posterior over β.

With these two environments in hand, we can explore the differences between the three approaches given above. This is illustrated in Figure 2-15. In particular, we note the drawbacks of the second approach (generating a posterior over β alone), which has the appealing feature that it does not require a structural causal model with deterministic functions, only a graphical model with some time-independent latent factor β. Intuitively, this approach will face two drawbacks, which are illustrated in Figure 2-15:

- First, it is not guaranteed to replicate the same outcomes if the same actions are taken, violating our intuition for how a counterfactual should behave. This can be seen in Figure 2-15, where the counterfactuals are the only trajectories that exhibit this behavior.

[15] The trajectory from the prior section is used for the first environment, for continuity

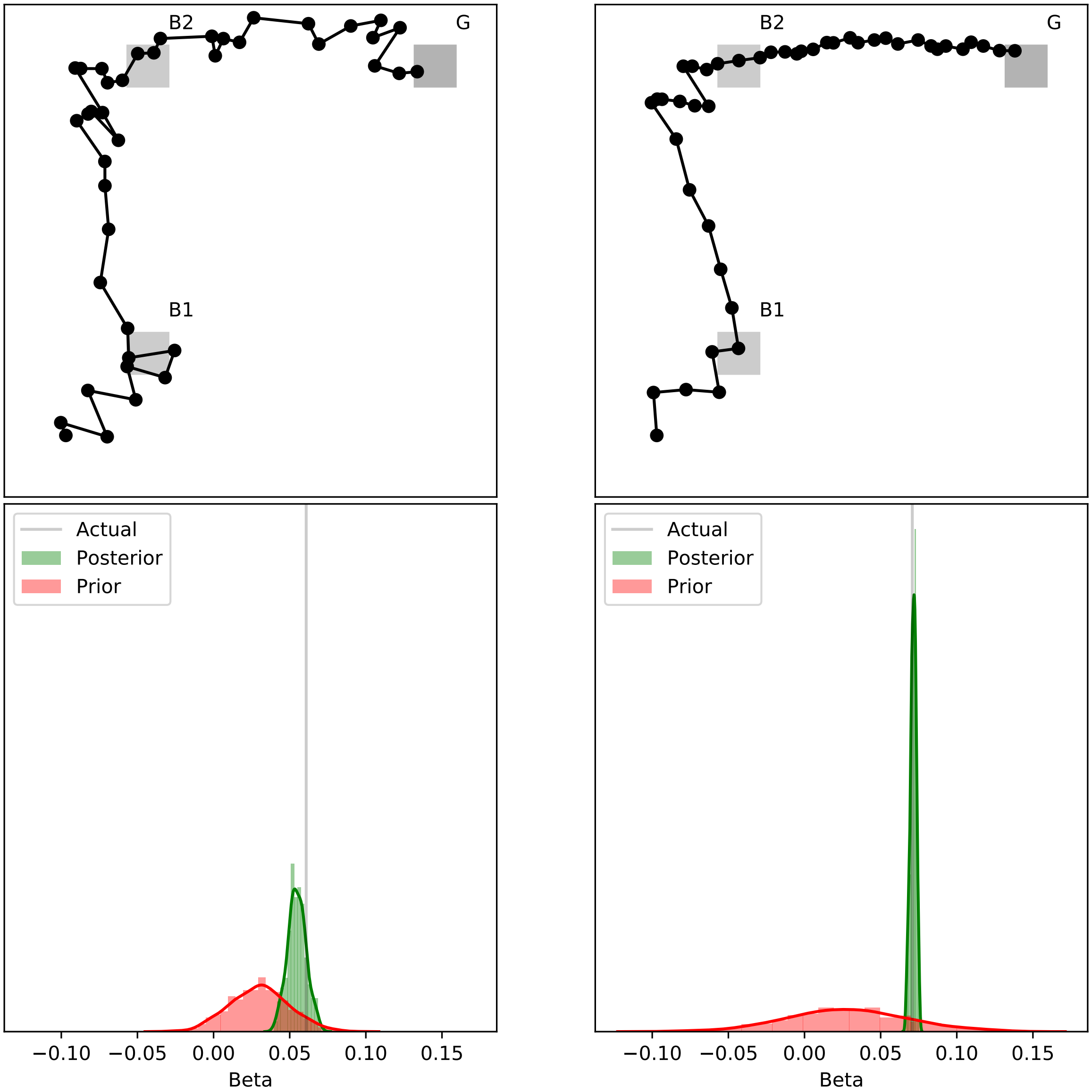


Figure 2-14: *A single trajectory sampled from each of the two environments. The latter environment has a higher prior variance over β, and a lower variance over ϵ. Below the trajectories are the corresponding prior and posterior distributions over β.*

- Second, it will ignore the information provided by ϵ, leading to unnecessary variance in the roll-out. If we have a SCM which is an accurate representation of the environment (as we do in this case), we can reduce the variance substantially by taking this information into account, especially when the variance of ϵ is high. This is also seen in Figure 2-15, where in the top row there is little (visual) difference between the two model-based approaches.

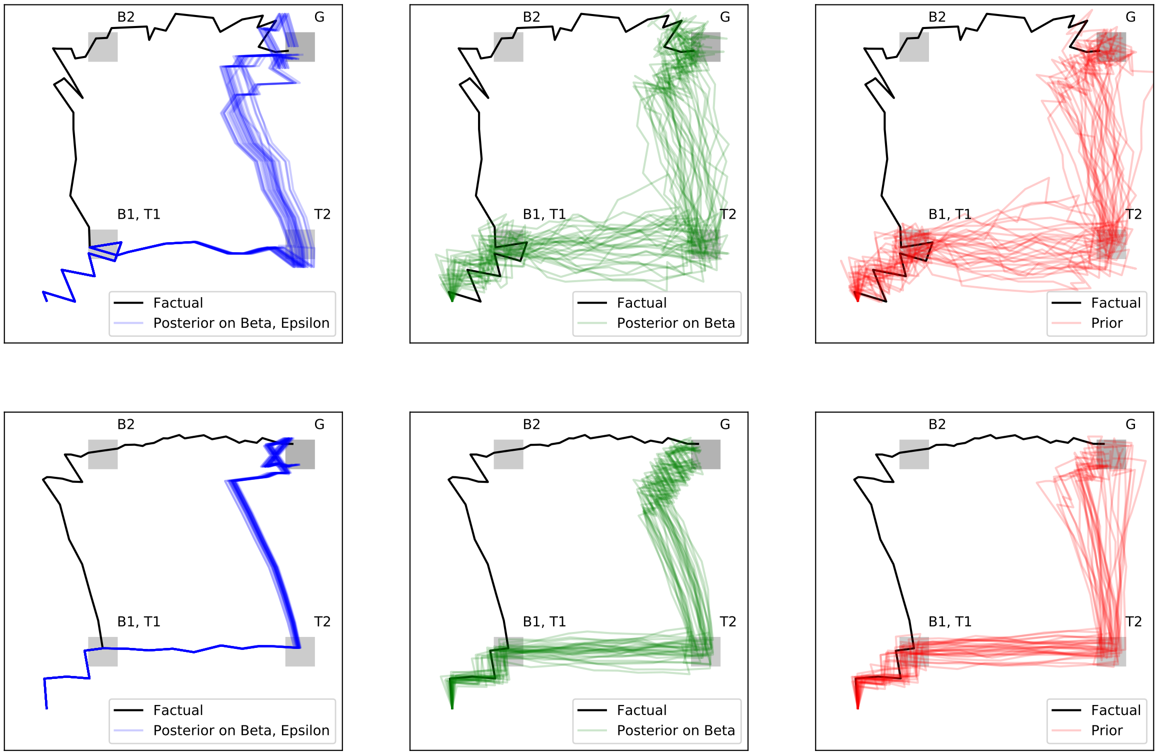


Figure 2-15: *Comparison of the three approaches given above. The first row represents the first environment, where $\sigma_\epsilon = 10^{-3}, \sigma_\beta = 0.02$, and the second row represents the second environment, where $\sigma_\epsilon = 10^{-4}, \sigma_\beta = 0.04$. The black trajectory represents a factual trajectory, and is constant across the columns. From left to right, we have counterfactual trajectories (which use the posterior on β and ϵ), model-based trajectories which only use a posterior on β, and model-based trajectories which use neither, just sampling from the prior. Note that in the first row, using the posterior on β does not reduce the variation as much as it does in the second row, due to the differences in σ_ϵ.*

This concludes our conceptual example, which should drive home the idea that, if we have an accurate SCM of the environment, we can construct counterfactual trajectories which (a) allow us to decompose differences in reward across individual episodes, and (b) are easier to contrast with the original trajectory than other model-based trajectories (e.g., without using a SCM), through modelling all sources of variation in the factual trajectory. In particular, using a SCM allows us to isolate only the differences which are due to the change in policy, keeping all independent sources of variation constant. In the next section, we will take this a step further, and show how counterfactuals can help us 'debug' a policy and model of an environment, even if our SCM is not entirely correct.

2.6.2 Illustrative Example: Sepsis Management

As discussed in Section 2.1, our hope is to provide a method for qualitative introspection and 'debugging' of RL models, in settings where a domain expert could plausibly examine individual trajectories. We give an illustrative example of this use case here, motivated by recent work examining the use of RL algorithms for treating sepsis among intensive-care unit (ICU) patients. In particular, we use a simple simulator of sepsis and "debug" a RL policy that is learned on observed trajectories. This replicates an analysis originally presented in our publication (Oberst and Sontag, 2019).[16]

An analysis like this requires three ingredients: First, we are given *observed trajectories*, but cannot directly interact with the environment[17]. Second, we have access to a *structural causal model* of the environment. In this case, that model is a finite MDP, learned based on observed samples, combined with the assumption of a Gumbel-Max SCM for transition distributions. Finally, we need a *target policy* to evaluate. We refer to the policy which generated the data as the *behavior* policy, to distinguish it from the target policy.

In Sections 2.6.2-2.6.2 we describe our illustrative scenario, in which a target RL policy appears to perform well using off-policy evaluation methods such as weighted importance sampling, when it is actually much worse than the behavior policy. In Sections 2.6.2-2.6.2 we then demonstrate how our method could be used to identify a

[16] We also thank Christina X. Ji and Fredrik D. Johansson for their work on developing an earlier version of the sepsis simulator.

[17] We do not assume access to a simulator; In this example, it is used only for obtaining the initial observed trajectories

promising subset of trajectories for further introspection, and uncover the flaws in the target policy using side information (e.g., chart review of individual patients).

Setup of Illustrative Example

Environment: Our simulator includes four vital signs (heart rate, blood pressure, oxygen concentration, and glucose levels) with discrete states (e.g., low, normal, high), along with three treatment options (antibiotics, vasopressors, and mechanical ventilation), all of which can be applied at each time step. Reward is +1 for discharge of a patient, and -1 for death. Discharge occurs only when all patient vitals are within normal ranges, and all treatments have been stopped. Death occurs if at least three of the vital signs are simultaneously out of the normal range. In addition, a binary variable for diabetes is present with 20% probability, which increases the likelihood of fluctuating glucose levels.

Observed Trajectories: For the purposes of this illustration, the behaviour policy was constructed using Policy Iteration (Sutton and Barto, 2017) with full access to the parameters of the underlying MDP (including diabetes state). This was done deliberately to set up a situation in which the observed policy performs well. To introduce variation, the policy takes a random alternative action w.p. 0.05. Using this policy, we draw 1000 patient trajectories from the simulator, with a maximum of 20 time steps. If neither death nor discharge is observed, the observed reward is zero.

Structural Causal Model: For this illustration, we 'hide' glucose and diabetes state in the observed trajectories; Given this reduced state-space, we learn the parameters of the finite MDP by using empirical counts of transitions and rewards from the 1000 observed trajectories, with death and discharge treated as absorbing states. For state / action pairs that are not observed, we assume that any action leads to death, and confirm that this results in a target policy which never takes an action that has never been observed. For counterfactual evaluation, we make the assumption that the transitions are generated by a Gumbel-Max SCM.

Target Policy: The target policy is learned using Policy Iteration on the parameters of the learned MDP. Because the target policy is learned using a limited number of samples, as well as an incomplete set of variables, it should perform poorly relative to the behavior policy.

Further details of the simulator can be found in the source code, which will be made available at `https://www.github.com/clinicalml/cf-policy-introspection`.

Off-Policy Evaluation Can Be Misleading

First, we demonstrate what might be done to evaluate this target policy without the use of counterfactual tools. In Figure 2-16, we compare the observed reward of the actual trajectories against the estimated reward of the target policy. Using weighted importance sampling on the given trajectories, the target policy appears superior to the behavior policy. We also use the parameters of the learned MDP to perform model-based off-policy evaluation (MB-PE), using the MDP as a generative model to simulate trajectories and their expected reward. Both of these suggest that the target policy is superior to the behavior policy. In reality, the target policy is inferior (as expected by construction), as verified by drawing new samples from the simulator under the target policy. This corresponds conceptually to what would happen if the target policy were deployed "in the wild".

With this in mind, we demonstrate how examining individual counterfactual trajectories gives insight into the target policy. The first step is to apply counterfactual off-policy evaluation (CF-PE) using the same MDP and the Gumbel-Max SCM. This yields similarly optimistic results as MB-PE. However, by pairing counterfactual outcomes with observed outcomes of individual patients, we can investigate *why* the learned MDP concludes (wrongly) that the target policy would be so successful.

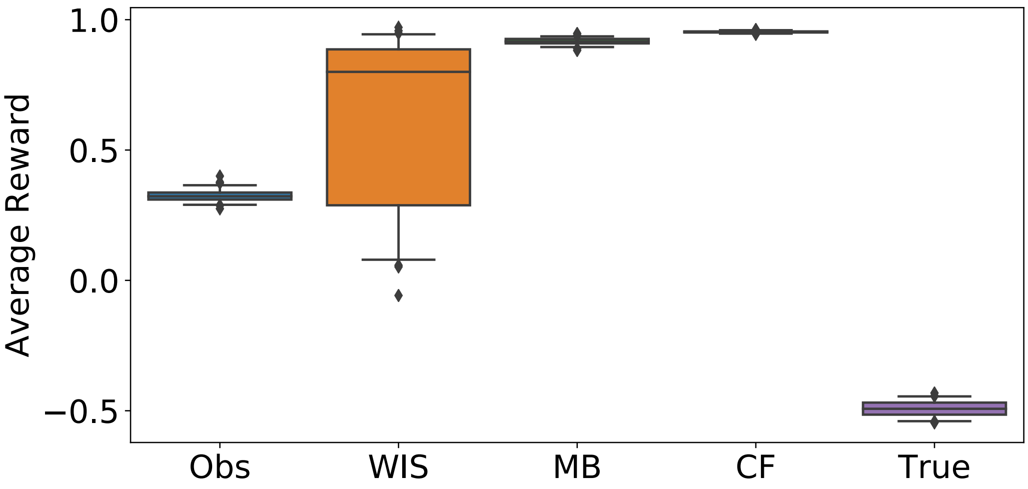


Figure 2-16: *Estimated reward under the target (RL) policy, with 95% uncertainty intervals generated through 100 bootstrapped samples (with replacement) of the same 1000 observed trajectories (for 1-4) and of 1000 new trajectories under the target policy (for 5).* ***(1) Obs****: Observed reward under the behavior policy.* ***(2) WIS****: Estimated reward under the target policy using weighted importance sampling.* ***(3) MB****: Estimated reward using the learned MDP as a generative model.* ***(4) CF****: Estimated reward over counterfactual trajectories (5 per observed trajectory).* ***(5) True****: Observed reward under the target policy, over 1000 newly simulated trajectories.*

Identification of Informative Trajectories

To debug this model (without access to a simulator), we can start by drawing counterfactual trajectories for each individual patient under the target policy. With these in hand, we can assign each individual patient to one of nine categories, based on the most frequently occurring counterfactual outcome (death, no change, or discharge) in Figure 2-17. This highlights individual trajectories for further analysis, as discussed in the next section[18].

[18] We only draw 5 counterfactuals per observed trajectory for illustrative purposes here, but note that standard concentration arguments could be used to quantify how many of these independent draws are required to achieve a desired precision on counterfactual quantities of interest, e.g., the probability of death

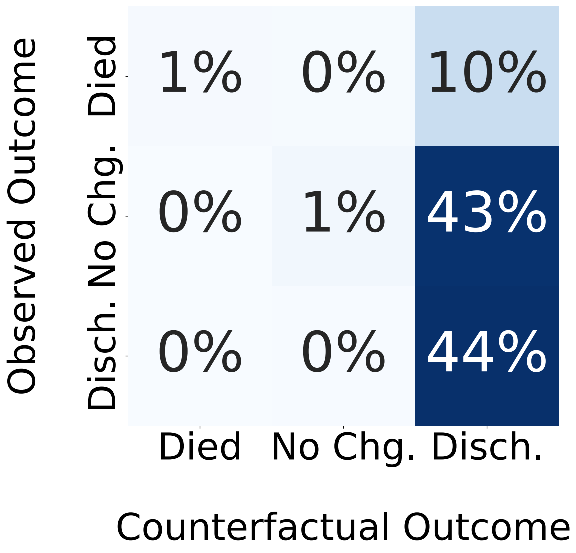


Figure 2-17: *Decomposition of 1000 observed patient trajectories based on observed outcome (Died, no change, and discharged) vs counterfactual outcome under the target policy, using the most common outcome over 5 draws from the counterfactual posterior.*

Insights from Examining Individual Trajectories

Using this decomposition, we can focus on the 10% of observed trajectories where the model concludes that "if the physician had applied the target policy, these patients would have most likely lived".

This is a bold statement, but also one that is plausible for domain experts to investigate (e.g., through chart review of these specific patients), to try and understand the rationale. We illustrate this type of analysis in Figure 2-18, which shows both the observed trajectory and the counterfactual trajectories for a simulated patient.

This example illustrates a dangerous failure mode, where the target policy would have halted treatment despite the glucose vital being dangerously low (e.g., at $t = 5, 7, 8, 11$). Under the learned MDP, the counterfactual optimistically shows a speedy discharge as a result of halting treatment. To understand why, recall that discharge occurs when all four vitals are normal and treatment is stopped. Because diabetes and glucose fluctuations are relatively rare, and because the MDP does not observe either, the model learns that there is a high probability of discharge when the first three vitals are normal, and the action of 'stop all treatments' is applied.

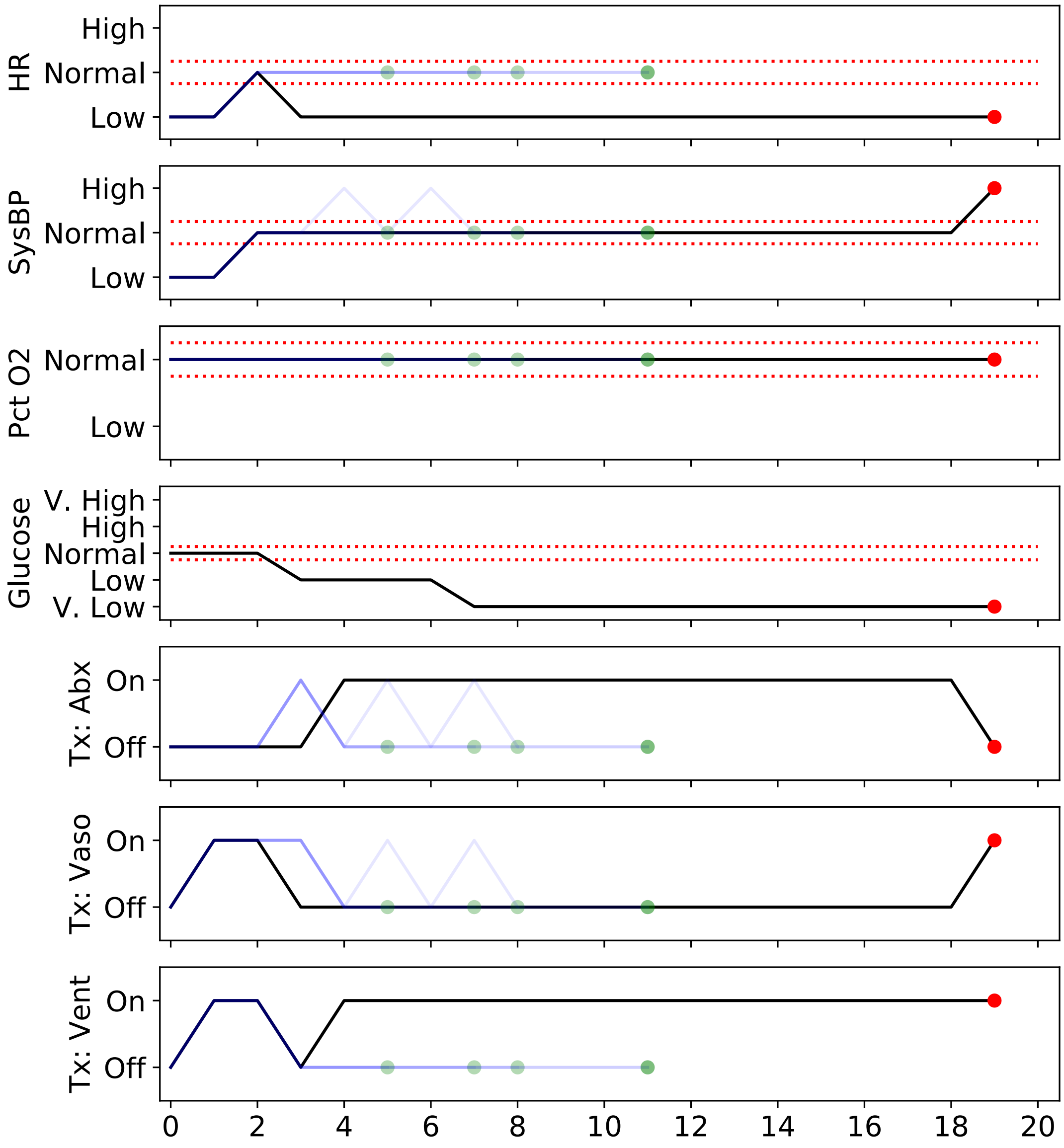


Figure 2-18: *Observed and counterfactual trajectories of a patient. The first four plots show the progression of vital signs, and the last three show the treatment applied. For vital signs, the normal range is indicated by red dotted lines. The black lines show the observed trajectory, which ends in death (signified by the red dot), and the blue lines show five counterfactual trajectories all of which end in discharge, signified by green dots. The glucose vital sign was not included in the model, and hence does not have a counterfactual trajectory. Note how this differs from a newly simulated trajectory of a patient with the same initial state, e.g., all the counterfactual trajectories are identical to the observed trajectory up to a divergence in actions ($t = 2$).*

Addendum: Impact of Hidden State

In the experiments given above, we hide the glucose and diabetes state from the model of dynamics used for the RL policy. In this section we explore the impact of that

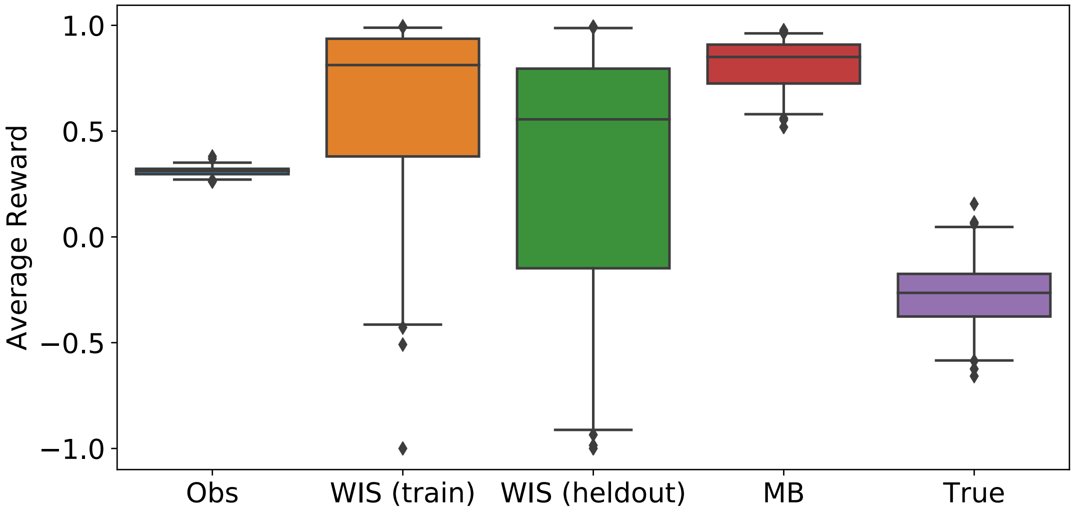


Figure 2-19: *Boxplots show the median and intervals which capture 95% of the 100 evaluations, each time with a newly simulated set of 1000 episodes used for training and 1000 episodes used for the held-out WIS estimator; WIS (train) is used on the training episodes, as in the previous sections, and WIS (held-out) is performed on the held-out set of 1000 episodes*

choice on the off-policy evaluations, as well as on the quality of the RL policy.

To demonstrate, in Figure 2-19, we replicate Figure 2-16, but with some important differences. First, instead of using 100 bootstrapped samples of the original 1000 trajectories, we instead repeat the entire process 100 times, with an independent set of trajectories drawn from the simulator in each case. These uncertainty intervals are wider, reflecting the variation which is not captured by bootstrapping alone. Second, we compare the use of a WIS estimator used on the training data (i.e., the original 1000 episodes used to learn the model of dynamics), with a WIS estimator used on a held-out set of 1000 independent episodes. While the example given in the Section 2.6.2 is meant to conceptually capture what might happen in a single analysis (where only a single set of trajectories is available), Figure 2-19 demonstrates the variability across analyses, including those with access to a large held-out set of trajectories.

Towards understanding the impact of hiding variables from the RL policy, we performed the same experiment again, but giving the RL policy access to the entire state space. The results are shown in Figure 2-20, and the results from both figures are shown in

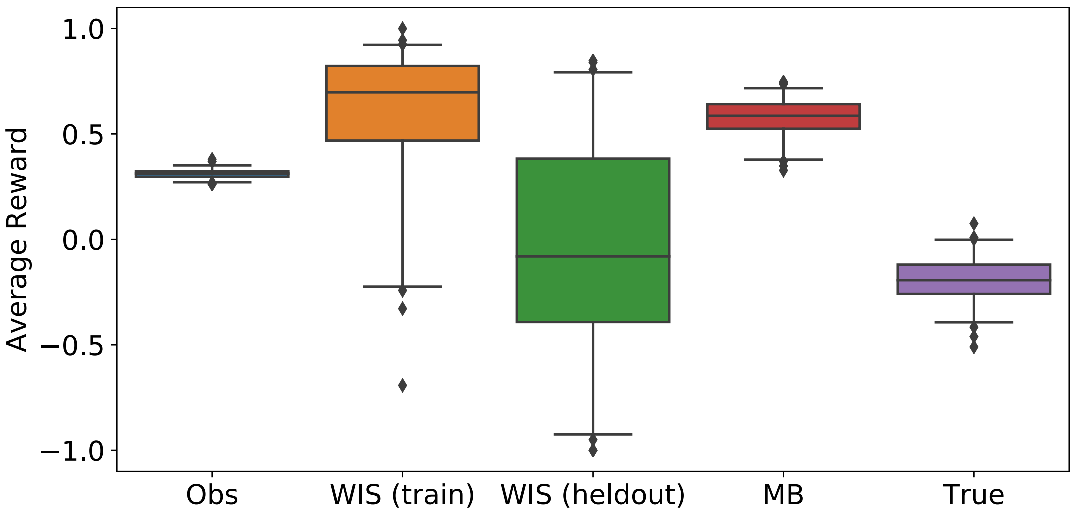


Figure 2-20: *Same setup as Figure 2-19, but allowing the model of dynamics (and the estimated behavior policy) to see the full state*

Table 2.1

Table 2.1: *Performance given as Mean (95% CI) from Figures 2-19- 2-20*

	Hidden state	No hidden state
Observed Reward	0.31 (0.27, 0.35)	0.31 (0.27, 0.35)
WIS (train)	0.61 (-0.42, 0.99)	0.58 (-0.23, 0.92)
WIS (heldout)	0.32 (-0.92, 0.99)	-0.04 (-0.94, 0.80)
MB Estimate	0.81 (0.57, 0.96)	0.58 (0.37, 0.73)
True RL Reward	-0.27 (-0.59, 0.05)	-0.19 (-0.41, 0.00)

There are several reasons why weighted importance sampling, and other off-policy evaluation methods, could fail to capture the true performance of a target policy. These include issues like confounding and small sample sizes, as discussed in (Gottesman et al., 2019a). In this particular synthetic example, all of the following factors may play a role in the above results, but it is difficult to say conclusively how strong each factor is, and how they interact to produce the results: (i) Confounding due to unobserved states, (ii) sample complexity of learning the MDP, which is more pronounced when all state information is observed (144 states vs 1440 states), and (iii) small sample sizes in both the training and held-out datasets.

With that in mind, we believe that building a more comprehensive simulated environment, in which these various factors can be disentangled more precisely, would be a valuable direction for future work. In addition, we believe such an environment would be useful for evaluation of a variety of off-policy techniques beyond the limited set discussed in this chapter e.g., more recently developed methods such as Thomas and Brunskill (2016); Liu et al. (2018).

2.7 Real-Data Case Study: Sepsis Management

In this chapter, we replicate the work of Komorowski et al. (2018), which seeks to learn an optimal policy for treating patients with sepsis in the ICU, using model-based RL techniques based on a finite MDP. We then apply our method of counterfactual policy introspection to the resulting policy and model, with the goal of understanding how well our approach works with a real-world example. We recapitulate a high-level overview of their methodology in Section 2.7.1, while deferring to the original paper for the full details of their setup. Having learned an MDP and corresponding policy following their approach, we perform a similar set of analyses to those we performed in Section 2.6.2: In Section 2.7.2 we estimate the reward using WIS on a held-out test set, and in Section 2.7.3 we decompose the counterfactual reward across trajectories in the test set.

Most notably, we find there are a very small number of patients who the model believes would have died in the counterfactual, and (as such) most of the patients who died in their observed trajectories are projected to have lived under the counterfactual. We select a random trajectory from this latter set for further analysis in Section 2.7.4, and review it alongside the full medical record, with the assistance of a clinician. In short, we find that it recommends actions which are not appropriate for this patient, based on information available in the clinical notes, and it expects unrealistic outcomes in the counterfactual as a result of those actions. We discuss this case in more depth in Section 2.7.4.

Finally, in Section 2.7.5 we discuss some aspects of the original paper that made this analysis challenging, as well as some broader reflections on the exercise as a test-case for understanding where our approach works well, and where it has limitations.

2.7.1 Replicating Komorowski et al. (2018)

The authors of (Komorowski et al., 2018) seek to learn a better policy for treating patients in the ICU with sepsis, as discussed previously in this chapter. In this section, we describe their approach at a high level, as well as our methodology for replicating it. We would like to thank Matthieu Komorowski for his assistance in replicating the original paper.

Data Source There are two sources of data used in Komorowski et al. (2018); First, they use data from the MIMIC-III database (Johnson et al., 2016), which contains de-identified medical records from >50k admissions to critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts. It also contains out-of-hospital mortality information using the Social Security Administration Death Master File. In their work, MIMIC-III is used for model development, and a separate dataset is used for model testing, the eICU Research Institute Database (eRI). We used the MIMIC-III database for both model development and testing, using a held-out test set of patients for evaluation, in part due to the availability of clinical notes.

Data Processing We used MATLAB code supplied by the authors at `https://github.com/matthieukomorowski/AI_Clinician` to process the raw data into the necessary format, which consists of one row of data for each 4-hour block of a patient's ICU stay, with a maximum of 20 rows per patient. We used slightly modified versions of the scripts, which will be made available at `https://github.com/clinicalml/cf-policy-introspection`. The original scripts are

1. `AIClinician_Data_extract_MIMIC3_140219.ipynb` to extract data from the MIMIC-III database

2. `AIClinician_sepsis3_def_160219.m` to create the sepsis cohort itself

3. `AIClinician_MIMIC3_dataset_160219.m` to construct the data table for downstream analysis

Learning and Evaluation We wrote our own python script to replicate the following procedure for selecting the best policy, using the code provided in `AIClinician_core_160219.m` as a guide when details were not clear from the main paper.

1. Center and scale all of the non-binary variables across the entire dataset, using log transformations where appropriate, and converting binary variables into $[-0.5, 0.5]$. For the two action variables (fluids and vasopressors), discretize into 5 bins, with the first reserved for zero treatment, and the remaining 4 based on quantiles over the entire dataset. Hold out 20% of the MIMIC-III data (by patient ID) as a test set.

2. Repeat 500 times, using a different 80/20 train / validation split on the remaining patient IDs:

   (a) Use K-Means clustering on 25% of the data[19] to assign each 4-hour block to one of 750 states

   (b) Use 90-day survival as the reward signal, with $100, -100$ corresponding to survival and death, respectively. This reward is obtained at the end of a trajectory (or after 20 steps, whichever is lower). Create two new absorbing states to reflect these outcomes.

   (c) Use empirical transition counts (state, action → state) to fill in the (three dimensional) transition matrix $P(S'|S, A)$, ignoring any state / action pair with fewer than 5 observations (we will refer to this later as 'truncation'). In the original paper, many of the state / action pairs have no observations, or fewer than 5 observations, so the transition matrix is not fully defined. We resolved this by treating any observed state / action pair as leading to

[19] This was done in the original paper for computational reasons, and we do the same

the ‘death’ absorbing state, towards the stated goal in Komorowski et al. (2018) of preventing the RL policy from taking any action which is rarely or never seen at a certain state. See Section 2.7.5 for more discussion on this point. Rewards are defined with respect to the absorbing state, so this suffices to define the MDP.

(d) Learn a policy from this MDP using Policy Iteration, and evaluate using Weighted Importance Sampling (WIS) on the validation set. In the original paper, the physician policy is estimated on the training set using the empirical transition counts, after truncation (described above), and then softened so that all actions have non-zero probability. The approach to softening could cause some probabilities to be negative, so we use a slightly different approach, described in Section 2.7.5. The RL policy is also softened to an ϵ-greedy policy for the purposes of WIS, where the learned action is taken with 99% probability, and otherwise a random alternative is taken.

(e) Calculate a 95% confidence interval using bootstrapped validation samples, and record the lower bound.

3. Using the k-means clustering, estimated MDP, and the resulting policy which obtained the highest WIS lower bound on the validation set, evaluate on the test set.

2.7.2 Off-Policy Evaluation with WIS

We give the results of our replication in Figure 2-21 and Tables 2.2 and 2.3. First, we note that there is a large variation in WIS performance on the validation set, with an average estimated reward which is lower than that of the behavior policy. Second, the test WIS results (using the ‘best’ policy) are highly variable as well, as revealed through bootstrapping on the 4415 test samples in Table 2.3. This motivates the rest of this section, where we dig further into the counterfactual trajectories to better ‘sanity check’ this policy.

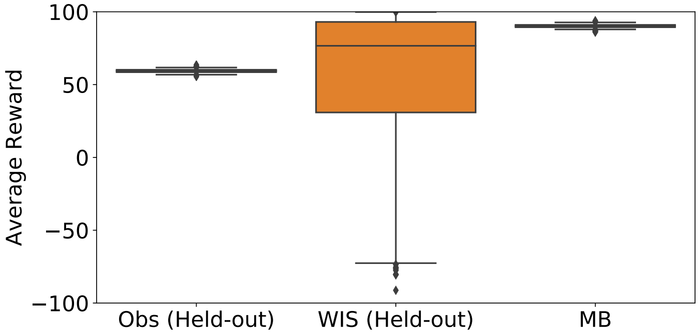


Figure 2-21: *Observed reward of the physician policy (Obs) versus the estimated reward of the learned RL policy using both weighted importance sampling on the validation set (WIS) as well as a model-based (MB) estimate derived from simulating 1000 trajectories, using the learned policy, on the learned MDP. Box-plots show the median and 95% range across 500 iterations. Higher is better.*

Table 2.2: *Results from 500 iterations of the procedure described in Section 2.7.1. Mean, median, and 95% range calculated over all iterations, and 1000 simulated trajectories were used to derive the model-based result, using the same MDP that was used to learn the policy. Higher is better.*

	Mean	Median	95% range
Observed (Validation)	59.33	59.43	(56.81, 61.85)
WIS (Validation)	53.00	76.64	(-73.00, 99.91)
Model-based	90.22	90.20	(87.85, 92.70)

2.7.3 Decomposition with Counterfactuals

First, we draw 5 counterfactual trajectories (under the chosen policy) for each of the test trajectories, using the techniques described in Section 2.4. In Figure 2-22 we take the most common outcome across the counterfactual trajectories to assign each individual to one of six categories, based on their factual outcome of 90-day survival and their counterfactual outcome, which can include 'no outcome' (see Section 2.7.5 for more discussion on this point).

Most notably, we find that *very few patients* have a negative outcome in the counterfactual, and most of the patients who died would have lived. In the next section we investigate this further by selecting a random trajectory from the latter set of patients.

Table 2.3: *Performance of the chosen policy on the held-out test set of 4415 trajectories, using bootstrapping (750 iterations) to estimate the distribution*

	Mean	2.5%	25%	50%	75%	97.5%
Observed	60.28	57.83	59.46	60.32	61.09	62.82
WIS	60.26	-28.42	47.72	69.42	83.50	96.59

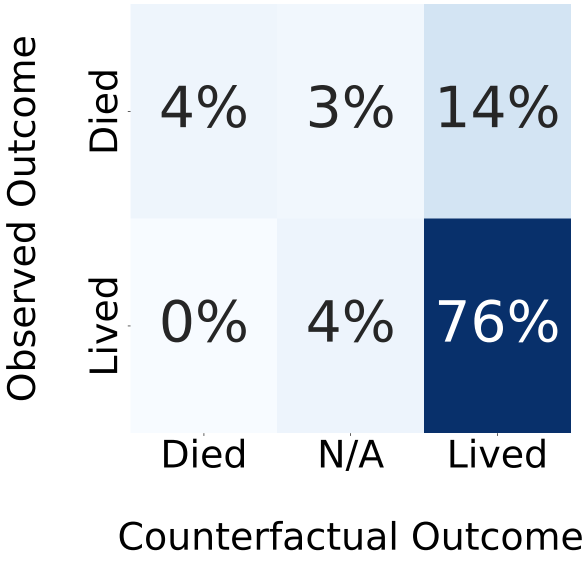


Figure 2-22: *Comparison of outcomes (90-day survival) between the observed and counterfactual trajectories, on the test set. Most notably, under the counterfactual it is estimated that very few patients would have died, and most of the patients who died would have lived. However, 7% of patients have no outcomes in the counterfactuals, due to a nuance discussed in Section 2.7.5*

2.7.4 Inspection of Counterfactuals using the Full Medical Record

As stated many times throughout this chapter, one of the main conceptual advantages of using counterfactuals is that they are conceptually easier to 'disprove', and that faults in the counterfactuals are a (heuristic) indication of faults in the learned model of the environment. In particular, by forcing our model to make counterfactual claims about an actual patient, we can bring additional side-information to bear on scrutinizing the conclusions. To that end, we present an illustrative example in this section, where we review the medical record of a patient alongside their counterfactual trajectories. In particular, we take a randomly selected patient from among those who died but 'would have lived' under all their estimated counterfactual trajectories.

We began by reviewing the clinical notes for this patient (the de-identified notes are available in MIMIC-III) with an infectious disease clinician[20]. A summary of the major takeaways from reviewing those notes:

- *Cause of admission:* This patient was admitted after collapsing, with initial suspicion that this was due to either a respiratory or cardiac failure, and was taken immediately to the cath lab where cardiac causes were ruled out. Chest imaging showed a large amount of fluid around the right lung, and a large mass in the lower right lobe. This was discovered to be State IIIA lung cancer, suggesting the possible etiology of the patients' presentation to be cardiovascular collapse and a post-obstructive pneumonia secondary to compression from the mass.

- *Treatment before and during ICU:* Cardiovascular compromise and inflammation from pneumonia contributed to the build up of a large amount of fluid in the pleural space. Thus, clinicians elected to place a chest tube, which subsequently drained >1L of serous fluid. The patient's clinical status responded rapidly, suggesting the external compression from the fluid was a major contributor to his ICU course. Antibiotics and vasopressors in this setting act as temporizing measures until the definitive intervention of chest tube placement could be performed.

- *Cause of death:* Despite the placement of a chest tube, the underlying problem of a large lung mass leading to cardiovascular compromise remained unaddressed. Given the morbidity of the necessary chemotherapy, it was decided by the providers, the patient and the family that further aggressive intervention would not have been in the patient's interests.

After reviewing the notes, we reviewed the counterfactual trajectories alongside the factual trajectories. We present a condensed output in Figure 2-23, consisting of a few

[20] Dr. Sanjat Kanjilal, MD, MPH, the Associate Medical Director of Clinical Microbiology at Brigham & Women's Hospital. We thank Dr. Kanjilal for all of his help with this work.

important vital signs, and defer the full output to Figures 2-24-2-27.[21] In particular, we make the following observations

- **No basis (in medical record) for proposed actions**: Recall that the patient was in fluid overload due to congestive heart failure and capillary leakage, which were themselves the result of the adjacent lung mass. The optimal approach in this setting is to carefully reduce the cardiac afterload using diuretics and anti-hypertensives, as well as drainage of the pleural effusion. Thus, while vasopressors and fluids are not grossly counter-indicated, they would have the opposite effect — increasing the work of the heart because they increase cardiac afterload, eventually resulting in worsening of the patients clinical status. Thus, while in the early admission period it is not unreasonable to provide vasopressors and fluids to maintain vital signs, there is a clinical trade-off, and there is no support in the medical record for giving *maximum dose of vasopressors* in the early stages, present in several of the counterfactual trajectories.

- **Consequences of proposed actions are not reflected in CF trajectories**: As noted, the alternative policy gives the maximum dose of vasopressers early on. However, the first 12 hours (first 3 time periods) look almost identical in the counterfactuals to the actual trajectory, and do not reflect the expected effect of additional vasopressors on blood pressure and other vital signs. In particular, maximum dose of vasopressors should have resulted in a significant blood pressure response, which is not evident in these counterfactual trajectories.

- **The anticipated outcomes are not credible given medical record**: Most glaringly,

[21] *How to read counterfactual trajectories:* To visualize the counterfactual trajectories, we map the patient state back to the original space of variables. To do so, we used the median of each feature in each cluster (across the entire dataset), though this is not entirely reliable, as can be seen by comparing the black solid lines (the median values for the corresponding state in k-means) with the black dotted lines, which indicate the true values of each variable. This mismatch is discussed further in Section 2.7.5. To read the trajectories, note that the observed trajectory is given in black, and the counterfactuals are given in light blue, with both derived from the medians (for each feature) of their respective states. Black dotted lines indicate the patient state without using k-means clustering. Red crosses and green dots both indicate the end of the trajectory, as well as the outcome, with green indicating 90-day survival and red indicating a lack thereof. Grey circles indicate no outcome in the counterfactual. Red dotted lines indicate the middle 90% across all patients, in the original data prior to k-means.

> the anticipated outcomes (discharge from the ICU and 90-day survival) are not credible given what we know about the patient from their medical record. For instance, the first counterfactual trajectory ends in 8 hours (with subsequent 90-day survival). That stands in contrast to what we know from the medical record, that the death of this patient was due to irreversible lung damage caused by Stage IIIA lung cancer and pneumonia, neither of which would have been resolved by this treatment.

Our review suggests an important possible limitation of the underlying learned MDP and policy. Important features (such as the underlying infection and lung cancer in this case) are not included in the model, but could reasonably impact both the outcome of the patient as well as the treatment decisions of clinicians. This issue also arises in a second trajectory that we randomly sampled (not shown here), in which the clinical notes indicated that the patient died from complications due to pre-existing Hodgkin Lymphoma and treatment in the ER (prior to admission to the ICU) which triggered respiratory failure and irreversible lung injury. The counterfactuals all indicated 90-day survival, contradicting the clinical notes which suggest that by the time the patient entered the ICU, nothing more could be done.

In conclusion, if we are to fully trust a model of dynamics, and the policy that is derived from it, then we would like to see a series of counterfactuals that 'make sense' to a clinician, as a type of explanation and justification for why the RL policy might have performed better than existing practice. As always, it is possible that the structural causal model itself is incorrect in this case, but we present this method as a useful (and simple) heuristic to apply, for generating hypotheses which could be useful for iterating on the model and resulting policy.

2.7.5 Challenges and Lessons Learned

There were a number of challenges in applying our methodology as imagined, some of which are due to idiosyncrasies with the approach in Komorowski et al. (2018).

Specification of outcome The outcome used in the original paper was 90-day mortality after discharge from the hospital, which was treated as an absorbing state. Moreover, for each patient, a maximum of 20 time-steps (of 4 hours each) were allowed, with the outcome always coming at the end of an observed trajectory. Thus, it has the implicit interpretation of 'discharge followed by [survival / death] after 90 days'. However, there is no guarantee that any model-based trajectory (including the counterfactuals) will end within 20 steps, leading to some instances where the counterfactual ends without an observed outcome.

Specification of states First, there are some idiosyncrasies with how state variables were encoded in the original paper. For instance, every variable is included in the k-means clustering, including those which should not fluctuate over the course of an ICU stay (such as gender and age). Second, we observed that our approach to visualization, of using the median value of each feature for each state, has some limitations. In particular, perhaps due to not having a large enough set of discrete states, when we 'impute' the factual trajectory based on the discrete states and compare it to the actual trajectory for those features, they are not always comparable. See Figure 2-28 for an example of this, taken from the same patient as above. This suggests that for our method to be most useful, the MDP should either operate in the original state space or operate in an invertible representation of it.

Estimation of behavior policy Because the behavior policy is derived using empirical counts, and because rare actions are truncated, it leads to an estimated zero-probability of several (observed) state/action pairs (including in the training set). This makes WIS impossible to use, because it relies on each observed action having non-zero probability. The solution to this taken in the original paper was to subtract a small amount from every action that has a non-zero probability, and add it to the other actions evenly. The way it was implemented in the supplied MATLAB code, this could cause some actions to have negative probability, because the amount subtracted was equal across observed actions. We resolved this in two ways: First, we did not implement truncation for the purposes of learning the behavior policy. Second, we softened the policy by instead adding a pseudo-observation of 0.01 to every action/state pair which was never seen, in the empirical counts.

Empirical MDP In the original paper, empirical counts are used to estimate the MDP, but does not result in a valid set of conditional distributions, because some state/action pairs are never observed. This is critical for our approach, because we need the observed trajectories to have non-zero probability under the MDP to calculate a counterfactual. Here we chronicle our efforts to resolve this, as well as explaining our final resolution:

1. As an initial attempt to resolve this, we first introduced the notion that you instantly die if you take an action that had never been taken, and used this to learn the policy (so that it avoids those actions). This approach forced us to re-learn the MDP on the test data for evaluation purposes, so avoid zero-probability trajectories.

2. This proved to be an inadequate solution for running on test data, because it results in a skewed model-based (and thus, counterfactual) estimate of reward; While the policy takes actions that were observed in the training data, they may not be observed in the test data, and by construction of our test MDP led to instant death.

3. Thus, we settled on using a softened MDP for the counterfactual evaluations, based only the training data, where we added a pseudo-observation of 10^{-3} for each transition, did not truncate observations, and did not use the 'instant death' rule. We confirmed (see Table 2.4) that this did not meaningfully impact the model-based estimate of reward under the RL policy, so we took it as a good proxy for the original MDP used to learn the policy.

Table 2.4: *Comparison of MDPs; 1000 model-based trajectories were averaged, and this was done 10 times to give 90% confidence intervals*

Approach	Average Reward	90% interval
Train	89.68	(88.08, 91.11)
Train (Soft)	85.32	(83.74, 86.52)

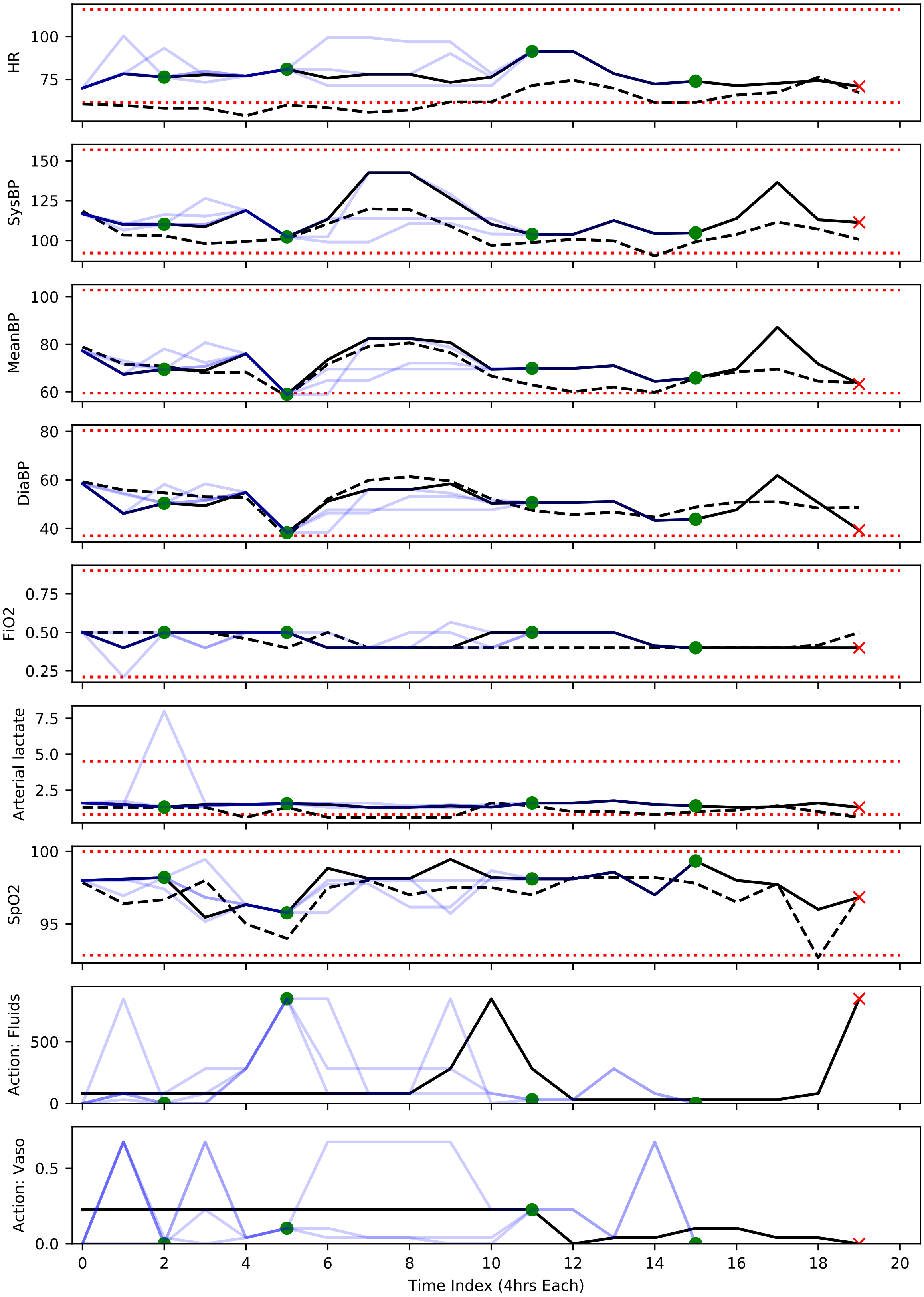


Figure 2-23: *Five counterfactual trajectories (for selected features), two of which end at $t = 15$. See description in the main text for how to read counterfactuals. HR: heart rate. BP: blood pressure. FiO2: fraction of inspired oxygen. SpO2: Peripheral oxygen saturation.*

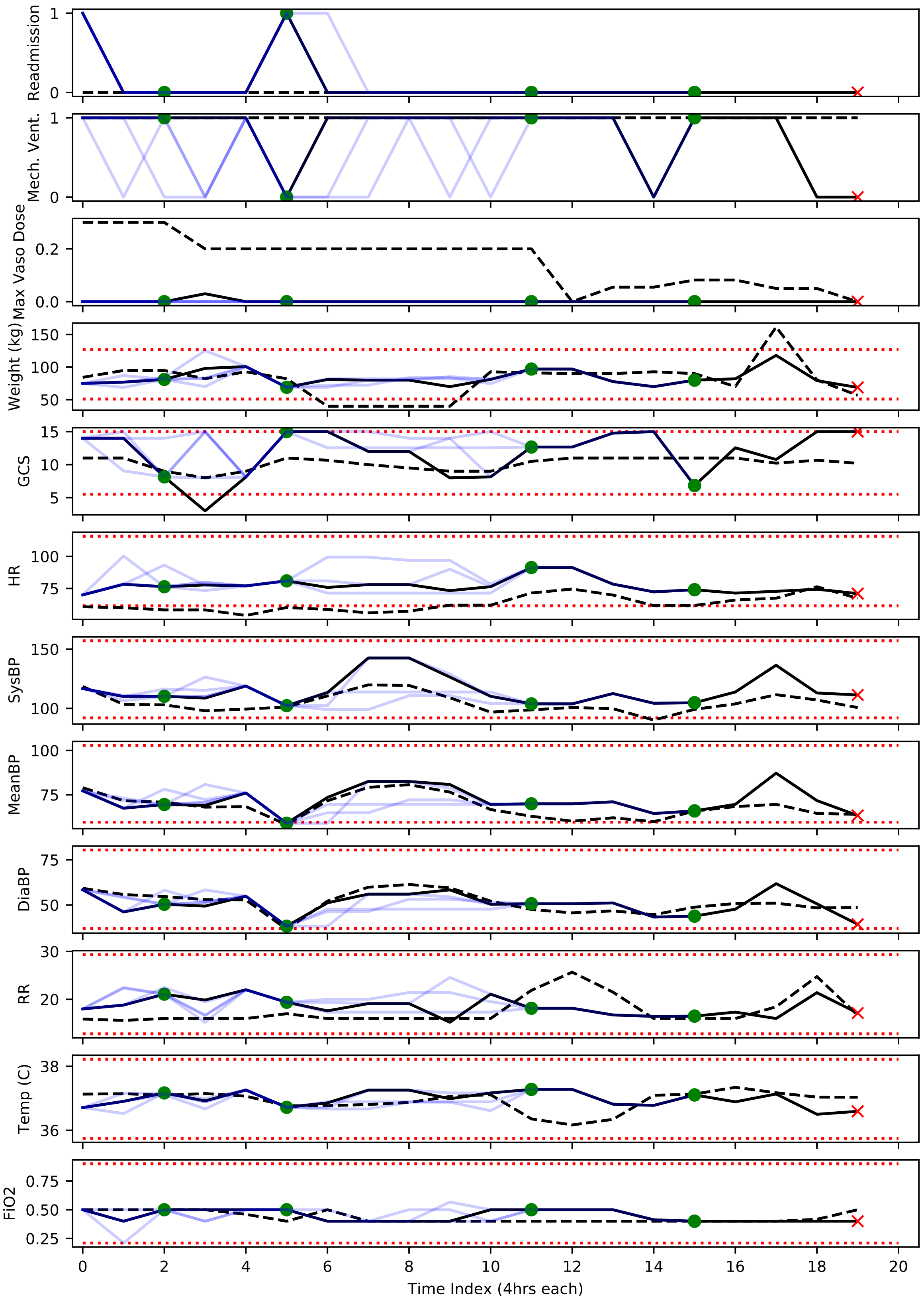


Figure 2-24: *Example Trajectory including all features (Part 1/4). See description in the main text.*

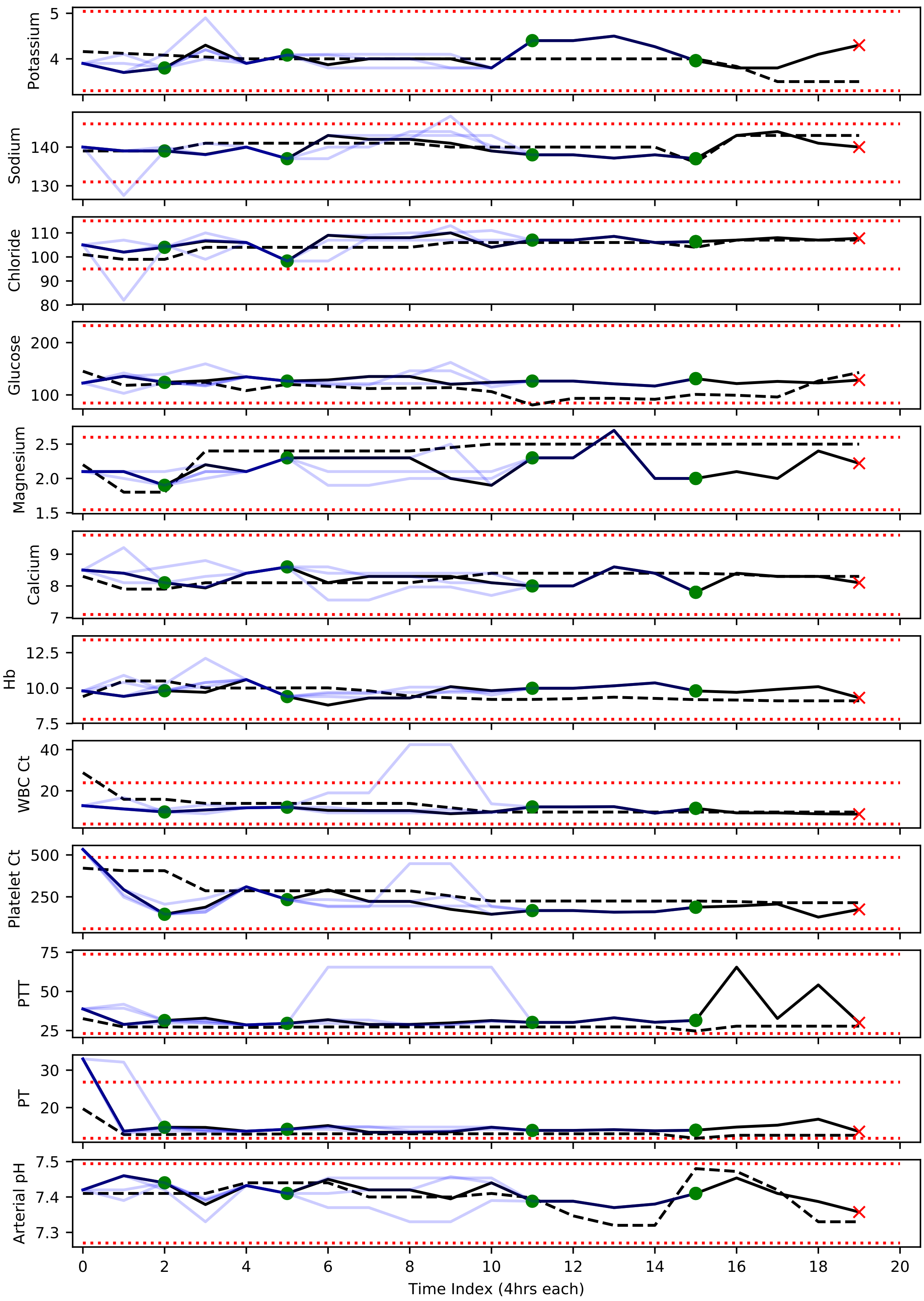


Figure 2-25: *Example Trajectory including all features (Part 2/4). See description in the main text.*

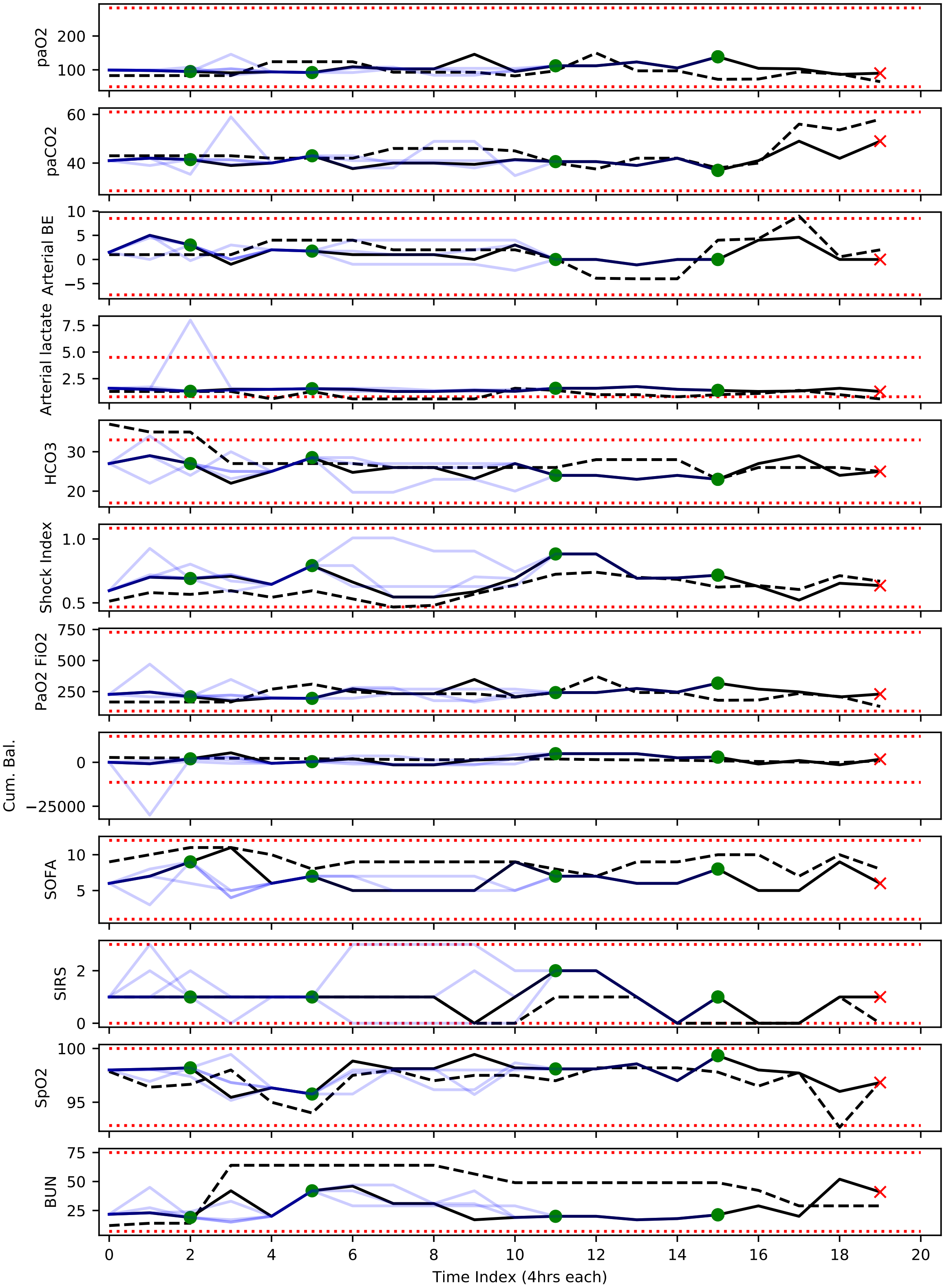


Figure 2-26: *Example Trajectory including all features (Part 3/4). See description in the main text.*

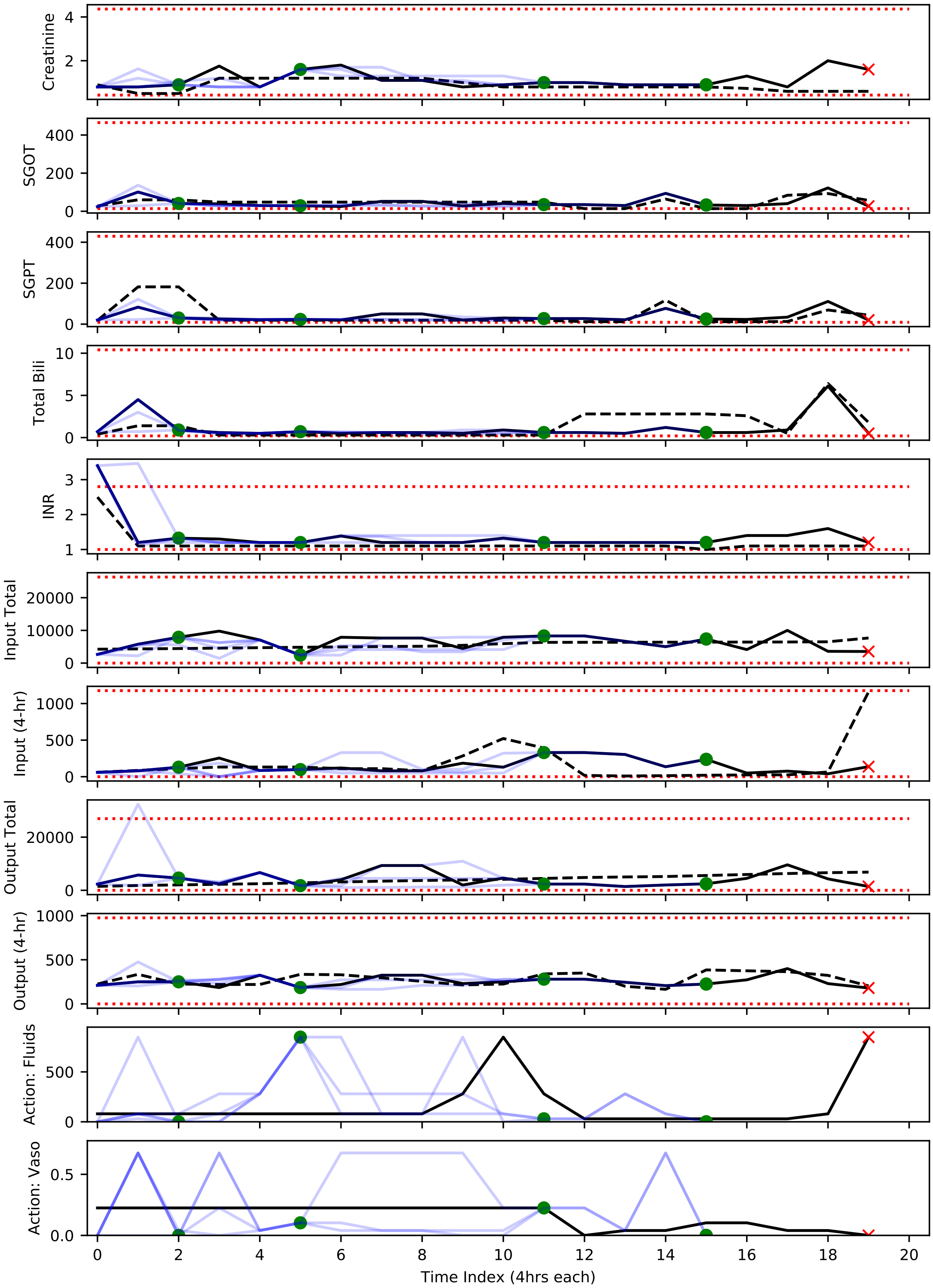


Figure 2-27: *Example Trajectory including all features (Part 4/4). See description in the main text.*

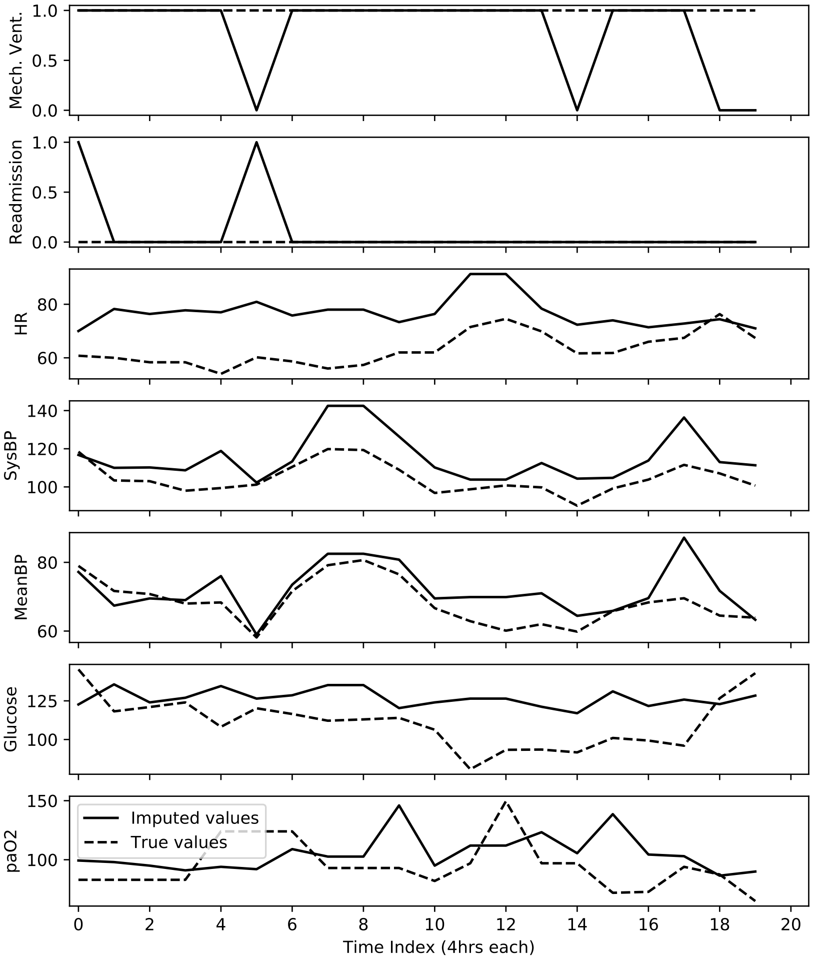


Figure 2-28: *Comparison of the imputed values (by taking the median of each feature for each cluster) and the actual values for the same patient.*

2.8 Conclusion

Given the desire to deploy RL policies in high-risk settings (e.g., healthcare), it is important to develop more tools and techniques to introspect these models and the policies they learn. In this chapter, we have presented a general method for doing so, which we call *counterfactual policy introspection.* Our approaches requires two inputs: A policy to be inspected, and a model of the relevant decision-making problem. This model could be a MDP or POMDP, or it could be any other learned graphical model of the environment which can be represented as a directed acyclic graph. By making general assumptions regarding the structure of causal mechanisms, we convert such a model into a structural causal model which can be used to generate counterfactuals. These counterfactuals serve several purposes:

1. First, they can be used to get a sense for which patients are driving the overall model-based reward. Theoretically, if the SCM is well specified, the expected counterfactual reward will be equivalent to the model-based reward. Anecdotally, in both our real-data and synthetic experiments (where the model was presumably not well-specified), we also found this to hold approximately.

2. Second, they can be used to highlight particularly interesting trajectories for further manual inspection. In our experiments, we give the example of highlighting patients who the model believes would have lived under the counterfactual, despite dying in the real world.

3. Finally, they serve to provide a detailed 'rationale' for the estimated performance of the policy, in terms of an expected counterfactual trajectory. These trajectories seek to isolate the differences in intermediate and final outcomes that are due to difference in actions, and can be reviewed along with side information (e.g., chart review in the medical setting) to identify flaws in the conclusions, which may suggest flaws in the original model.

However, this approach does not come without its limitations. It requires knowing, or making an untestable assumption about, the structural causal model: Here we propose the Gumbel-Max SCM, which is an example of an SCM that may be realistic in some settings. As revealed through our real-data experiment, our approach may also work best when the environment is modelled directly in the original state space, and the model of dynamics is not too brittle to handle unseen trajectories that may arise in test data. Nonetheless, our real-data experiments give us hope that this might be useful to researchers in the future, as a relatively straightforward method to debug models and generate hypotheses for improving them.

Chapter 3

Characterization of Overlap in Observational Studies

This chapter (and accompanying appendix) was previously published as (Oberst et al., 2020) at AISTATS 2020.

3.1 Introduction

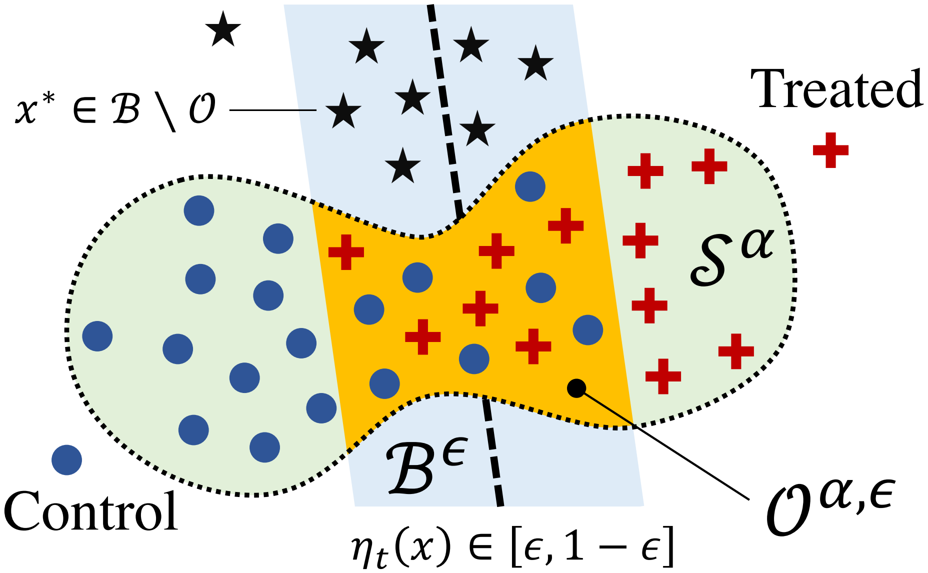
To accurately estimate the causal effect of an intervention, it is essential that intervention alternatives have been observed in comparable contexts, i.e., that there is *overlap* between the distributions of individuals receiving each intervention (Rosenbaum and Rubin, 1983b; D'Amour et al., 2017). In randomized experiments, overlap is guaranteed for the study population by randomizing the intervention. However, this is not the case in observational studies where interventions are chosen according to an existing, in some cases deterministic, policy. In such settings, overlap may hold only for an unidentified subset of cases, with the causal effect being unidentifiable outside of this subset. We motivate this chapter with the following use cases:

Scenario 1: From study to policy. When researchers publish the findings of a clinical trial, they also share the eligibility criteria (e.g., *Age* $\geq$ *18, Serum M protein* $\geq$ *1g/dl* *or Urine M protein ≥ 200 mg/24 hrs, Recent diagnosis* (National Cancer Institute, 2012)) and cohort statistics in order to characterize the cohort of study subjects. This gives policy makers means to assess the external validity of the results, i.e., to whom the results apply. We seek to provide the same for observational studies, with our algorithms producing an interpretable description of subjects with treatment group overlap.

Scenario 2: Evaluating guidelines. There are over 471 different guidelines for how to manage hypertension (Benavidez and Frakt, 2019). We could evaluate these—and new guidelines—using off-policy evaluation methods (Precup et al., 2000) on observational data derived from electronic medical records. Off-policy evaluation of a guideline is only possible on subsets of the population where there is some probability that the guideline was followed (which we will also call overlap). The estimated policy value should be accompanied by a description of the validity (overlap) region.

Beyond causal estimation, overlap is of interest in many other branches of machine learning: In domain adaptation, the overlap between source and target domains is the set of inputs for which we can expect a trained model to transfer well (Ben-David et al., 2010; Johansson et al., 2019); In classification, overlap between inputs with different labels signifies regions that are hard to classify; In algorithmic fairness (Dwork et al., 2012), overlap between protected groups may shed light on disparate treatment of individuals from different groups who are otherwise comparable in task-relevant characteristics; In reinforcement learning, lack of overlap has been identified as a failure mode for deep Q-learning using experience replay (Fujimoto et al., 2019).

Our main contributions are as follows: (i) We propose desiderata in overlap estimation, and note how existing methods fail to satisfy them. (ii) We give a method for *interpretable characterization of distributional overlap*, which satisfies these desiderata, by reducing the problem to two binary classification problems, and using a linear programming relaxation of learning optimal Boolean rules. (iii) We give generalization bounds for rules minimizing empirical loss. (iv) We demonstrate that small rules often perform comparably to black-box estimators on a suite of real-world tasks.



$x^* \in \mathcal{B} \setminus \mathcal{O}$
Treated
$\mathcal{S}^{\alpha}$
$\mathcal{B}^{\epsilon}$
$\mathcal{O}^{\alpha,\epsilon}$
Control
$\eta_t(x) \in [\epsilon, 1-\epsilon]$

Figure 3-1: *Overlap $\mathcal{O}^{\alpha,\epsilon}$ between treatment groups with joint support $\mathcal{S}^{\alpha}$. A point x^* has group propensity η_t bounded away from 0 and 1, but is outside of $\mathcal{O}^{\alpha,\epsilon}$.*

(v) We evaluate the interpretability of rules for describing treatment group overlap in post-surgical opioid prescription in a user study with medical professionals. (vi) We show how a generalized definition and method applies to policy evaluation and apply it to describing overlap in policies for antibiotic prescription.

3.2 Related work

Treatment group overlap is a central assumption in the estimation of causal effects from observational data. Comparing group-specific covariate bounds and lower-order moments is a common first step in assessing overlap (Rosenbaum et al., 2010; Zubizarreta, 2012; Fogarty et al., 2016) but fails to identify local regions of overlap when they exist (see the example of $\mathcal{O}^{\alpha,\epsilon}$ in Figure 3-1). An alternative is to estimate the *treatment propensity*—the probability that a subject was prescribed treatment. Treatment propensities bounded away from 0 and 1 at a point X indicates that treatment groups overlap at X (Rosenbaum and Rubin, 1983b; Li et al., 2018b).

In studies with partial overlap, it is common to restrict the study cohort by thresholding treatment propensity or discarding unmatched subjects after applying matching methods (Rosenbaum, 1989; Iacus et al., 2012; Kallus, 2016; Visconti and Zubizarreta, 2018). For example, Crump et al. (2009) proposed an optimal propensity threshold that minimizes the variance of the estimated average treatment effect on a sub-population. However, neither propensity thresholding nor matching are sufficient for guiding policy in new cases: they do not provide a self-contained, interpretable description of where treatment groups overlap *within* the study, nor do they provide insight into *external* validity by describing the limits of the study cohort.

Fogarty et al. (2016) address the first concern above by learning "interpretable study populations" through identifying the largest axis-aligned box that contains only subjects with bounded propensity. However, this approach is very limited in capacity and does not address external validity. For this reason, we strive to provide interpretable descriptions of overlap, both in terms of treatment propensity and the study support.

Rule-based models have been considered in classification tasks (Rivest, 1987; Angelino et al., 2017; Yang et al., 2017; Lakkaraju et al., 2016; Wang et al., 2017; Dash et al., 2018; Freitas, 2014; Wang and Rudin, 2015), subgroup discovery (Herrera et al., 2011) and density estimation (Ram and Gray, 2011; Goh and Rudin, 2015) but have to the best of our knowledge not been applied or tailored to support or overlap estimation.

3.3 Defining Overlap

We address *interpretable description of population overlap*. Our primary motivation is to aid *policy making* based on observational studies, the success of which relies on understanding and communicating the studies' validity region—the set of cases for which there is evidence that a particular policy decision is preferable. We identify the following desiderata for descriptions of overlap: (D.1) They cover regions where all populations (treatment groups) are well-represented; (D.2) They exclude all other regions, including those outside the support of the study (see Figure 3-1); (D.3) They can be expressed using a small set of simple rules. Next, we define overlap according to (D.1) and (D.2). We address (D.3) in Section 3.4.

Let subjects $i = 1, ..., m$ be observed through samples (x_i, t_i) of covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and a group indicator $T \in \mathcal{T}$. In our running example, X represents patient attributes and T their treatment. We assume that subjects are independently and identically distributed according to a density $p(X, T)$, and that $\mathcal{X}$ is bounded. Let $p_t(X) := p(X \mid T = t)$ denote the covariate density of group $t \in \mathcal{T}$ and $\eta_t(x) := p(T = t \mid X = x)$ the propensity of membership in group $t \in \mathcal{T}$ for subjects with covariates $x \in \mathcal{X}$. We denote the probability mass of a set $S \subseteq \mathcal{X}$ under p by $P(S) := \int_{x \in S} dp$ and the support of p by $\text{supp}(p) := \{x \in \mathcal{X} : p(x) > 0\}$.

In the common case of two groups, $\mathcal{T} = \{0, 1\}$, overlap is typically defined as either a) the intersection of supports, $\text{supp}(p_0) \cap \text{supp}(p_1)$, or b) the set of covariate values for which all group propensities η_t are bounded away from zero (D'Amour et al., 2017; Li et al., 2018b). We let $\mathcal{B}^\epsilon$ denote this latter set of values with *ϵ-bounded propensity* for a fixed parameter $\epsilon \in (0, 1)$ and an arbitrary set of groups $\mathcal{T}$,

$$\mathcal{B}^\epsilon := \{x \in \mathcal{X}; \forall t \in \mathcal{T} : \eta_t(x) > \epsilon\} \ . \tag{3.1}$$

Neither $\mathcal{B}^\epsilon$ nor the support intersection fully capture our desired notion of overlap: The former does not satisfy (D.2) since a point may have bounded propensity (true or estimated) but lie outside the population support $\text{supp}(p)$ (see Figure 3-1). Note that interpretable description alone does not address this. The latter is non-informative for variables with infinite support (e.g., a normal random variable), and even with finite support, we may wish to exclude distant outliers.

Our preferred definition of overlap combines the requirement of bounded propensity with a generalization of support called *α-minimum-volume sets* (Schölkopf et al., 2001). Let $\mathfrak{C}$ be a set of measurable subsets of $\mathcal{X}$, let $V(C)$ denote the volume of a set $C \in \mathfrak{C}$. An *α-minimum-volume* set $\mathcal{S}^\alpha$ of p is then

$$\mathcal{S}^\alpha := \arg\min_C \{V(C) \, ; P(C) \geq \alpha, C \in \mathfrak{C}\} \ , \tag{3.2}$$

with $\mathcal{S}^1 = \text{supp}(p)$. For $\alpha < 1$, $\mathcal{S}^\alpha$ is not always unique, but the intersection S of two α-MV sets has mass $P(S) \geq 2\alpha - 1$. In this work, we let $\alpha < 1$ in order to handle distributions with infinite support and unwanted outliers, and *refer to $\mathcal{S}^\alpha$ as the support of p*. We define the α, ϵ-*overlap set*, for $\alpha, \epsilon \in (0, 1)$, to be

$$\mathcal{O}^{\alpha,\epsilon} := \mathcal{S}^\alpha \cap \mathcal{B}^\epsilon \ . \tag{3.3}$$

We define the problem of overlap estimation under definition (3.3) as *characterizing the set $\mathcal{O}^{\alpha,\epsilon}$ given thresholds α and ϵ*. In line with (D.3), these characterizations should be useful in policy making, and interpretable by domain experts, at small cost in accuracy. For notational convenience, we sometimes leave out superscripts from $\mathcal{S}^\alpha, \mathcal{B}^\epsilon$ and $\mathcal{O}^{\alpha,\epsilon}$, assuming that α, ϵ are fixed.

Remark. Defining overlap instead as the intersection of group-specific α-MV sets is feasible, but scales poorly with $|\mathcal{T}|$; it does not facilitate the generalization to policy evaluation described below; and the intersection of many descriptions may be hard to interpret.

3.3.1 Generalization to Policy Evaluation

The definition of $\mathcal{B}^\epsilon$ in (3.1) is motivated by causal effect estimation—comparison of outcomes under two or more alternative interventions. We may instead be interested in policy evaluation, which involves estimating the expected outcome under a conditional intervention π, which assigns a treatment t to each x following a conditional distribution $\pi(T|X)$ (Precup et al., 2000). To perform this evaluation, we only require that the propensity $p(T|X)$ of observed treatments be bounded away from zero for treatments which have non-zero probability under π. To describe the inputs for which this is satisfied, we generalize $\mathcal{B}^\epsilon$ to be a function of the target policy π,

$$\mathcal{B}^\epsilon(\pi) := \{x \in \mathcal{X}; \forall t : \pi(t \mid x) > 0 : \eta_t(x) > \epsilon\} \ . \tag{3.4}$$

More details are given in the supplement regarding the use of OverRule in this setting.

3.4 Overrule: Boolean Rules for Overlap

We propose OverRule[1], an algorithm for identifying the overlap region $\mathcal{O}$ in (3.3) by first estimating the α-MV support set $\mathcal{S}$ (3.2) and then the bounded-propensity set $\mathcal{B}$ (3.1) restricted to $\mathcal{S}$, thereby satisfying desiderata (D.1)–(D.2). We aim to fulfill desideratum (D.3) by using Boolean rules—logical formulae in either disjunctive (DNF) or conjunctive (CNF) normal form—which have received renewed attention because of their interpretability (Dash et al., 2018; Su et al., 2016). See Figures 3-3–3-4 for examples of learned rules. OverRule proceeds in the following steps:

(i) Fit α-MV set $\hat{\mathcal{S}}^{\alpha}$ of $p(X)$ using Boolean rules

(ii) Fit model of group propensity $\hat{\eta}_{(\cdot)}$ over $\hat{\mathcal{S}}^{\alpha}$ and let $\tilde{b}(x) = \prod_{t\in\mathcal{T}} \mathbb{1}[\hat{\eta}_t(x) > \epsilon]$ define membership in $\tilde{\mathcal{B}}^{\epsilon}$

(iii) Approximate $\tilde{\mathcal{B}}^{\epsilon}$ using Boolean rules to yield $\hat{\mathcal{B}}^{\epsilon}$ and estimate overlap region by $\hat{\mathcal{O}}^{\alpha,\epsilon} = \hat{\mathcal{B}}^{\epsilon} \cap \hat{\mathcal{S}}^{\alpha}$.

In this section, we demonstrate how steps (i) & (iii) can be reduced to binary classification. This enables us to exploit the many existing methods for rule-based classification (Freitas, 2014) to improve the interpretability of $\hat{\mathcal{O}}$. Finally, we give results bounding the generalization error of estimates of both $\mathcal{S}$ and $\mathcal{S} \cap \mathcal{B}$.

Remark. It was observed in evaluations with a medical practitioner that fitting rules for $\mathcal{S}$ and $\mathcal{B}$ separately improved interpretability as it makes clear which rules apply to which task and prevents the bulk of the rules from being consumed by one of the two tasks.

3.4.1 Estimation of S^{α} as Binary Classification

In the first step of OverRule, we learn a Boolean rule to approximate the α-MV set $\mathcal{S}^{\alpha}$ of the marginal distribution $p(X)$ by reducing the problem to binary classification

[1] Code available at https://github.com/clinicalml/overlap-code

between observed samples $\mathcal{D} := \{x_i\}_{i=1}^{m}$ and uniform background samples. For clarity, we focus only on DNF rules—disjunctions of conjunctive clauses such as $(\text{Age} < 30 \wedge \text{Female}) \vee (\text{Married})$. As pointed out by Su et al. (2016), a CNF rule can be learned by swapping class labels and fitting a DNF rule.

We adapt previous notation and let $\mathfrak{C}$ be a class of candidate α-MV sets $\mathcal{C}$ corresponding to Boolean rules, i.e., each $\mathcal{C}$ consists of the points in $\mathcal{X}$ that satisfy a rule. We will often not distinguish between a rule and its corresponding set $\mathcal{C}$ and thus will speak of the "volume" of a rule or clause. We aim to solve a normalized and regularized version of the α-MV problem in (3.2),

$$\arg\min_{\mathcal{C}\in\mathfrak{C}} \; Q(\mathcal{C}) := \underbrace{\bar{V}(\mathcal{C})}_{\text{Volume}} + \underbrace{R(\mathcal{C})}_{\text{Regularization}} \quad \text{s.t.} \;\; \underbrace{P(\mathcal{C}) \geq \alpha}_{\text{Coverage}} \tag{3.5}$$

where the volume $\bar{V}(\mathcal{C}) = V(\mathcal{C})/V(\mathcal{X}) \in [0,1]$ is normalized to that of $\mathcal{X}$. We assume that the regularization term $R(\mathcal{C})$ controls complexity by placing penalties λ_0 on each clause in the rule and λ_1 on each condition in a clause. Thus, for a Boolean rule with clauses $k = 1, \ldots, K$, each with p_k conditions, we have[2]

$$R(\mathcal{C}) = K\lambda_0 + \lambda_1 \sum_{k=1}^{K} p_k. \tag{3.6}$$

It is also assumed that the trivial "all-true" and "all-false" rules have complexity $R(\mathcal{C}) = 0$.

The volume $\bar{V}(\mathcal{C})$ may be difficult to compute repeatedly during optimization and $\mathfrak{C}$ is often too large to allow pre-computation of $\bar{V}(\mathcal{C})$ for all $\mathcal{C}$. In particular, for DNF rules, each $\mathcal{C}$ is a union of potentially several overlapping clauses (see Figures 3-3–3-4 or the illustration in the supplement); even if the volume spanned by each clause is quick to compute on the fly, the overall volume may not be. As an alternative, the normalized volume $\bar{V}(\mathcal{C})$ can be estimated by means of uniform samples $\{x_{m+1}, \ldots, x_{m+n}\}$ over $\mathcal{X}$. Let $\mathcal{U}$ be the index set of these uniform samples. Then, $\frac{1}{n}\sum_{i\in\mathcal{U}} \mathbb{1}[x_i \in \mathcal{C}]$ is distributed

[2] It is possible to generalize (3.6) to place different penalties on different conditions but we adopt (3.6) for simplicity.

as a scaled binomial random variable with mean $\bar{V}(\mathcal{C})$ and variance $\bar{V}(\mathcal{C})(1-\bar{V}(\mathcal{C}))/n$. Theorem 3.1 below provides guidance in selecting the number of uniform samples n to ensure a good estimate.

Given the above empirical estimator of volume, we reduce problem (3.5) to a classification problem between the marginal density $p(X)$ and a uniform distribution over $\mathcal{X}$. This reduction was also mentioned in the conclusion of Scott and Nowak (2006). We also replace the probability mass constraint with its empirical version over $\mathcal{D}$ with $\mathcal{I} = \{1, \ldots, m\}$. The result is a Neyman-Pearson-like classification problem with a false negative rate constraint of $1-\alpha$ (instead of the usual false positive constraint), as given below.

$$\begin{aligned} \hat{\mathcal{S}} := \arg\min_{\mathcal{C}} \quad & \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} \mathbb{1}[x_i \in \mathcal{C}] + R(\mathcal{C}) \\ \text{subject to} \quad & \sum_{i\in\mathcal{I}} \mathbb{1}[x_i \in \mathcal{C}] \geq \alpha m \ . \end{aligned} \tag{3.7}$$

The following theorem bounds the regret of the minimizer of (3.7) with respect to (3.5) and is proven in the supplement. The assumption of binary variables simplifies the analysis and is not a fundamental limitation.

Theorem 3.1. *Let $q^*(\alpha)$ denote the minimum regularized volume attained in (3.5) over the class of DNF rules with probability mass α. Assume that a) the regularization R follows (3.6) with fixed parameters λ_0, λ_1, b) all variables X_j are binary-valued, and c) the class $\mathcal{C}$ is restricted to rules satisfying necessary conditions of optimality for (3.5) (see Lemmas in the supplement). Then with probability greater than $1-2\delta$, the empirical estimate $\hat{\mathcal{S}}$ in (3.7) satisfies*

$$Q(\hat{\mathcal{S}}) \leq q^*(\alpha+\epsilon_m) + 2\epsilon_n \quad \textit{and} \quad P(\hat{\mathcal{S}}) \geq \alpha - \epsilon_m,$$

where $\epsilon_m = \sqrt{\frac{\lambda_1^{-1}\log(2d) + \lfloor 1+\log_2 \lambda_1^{-1} \rfloor \log \lambda_1^{-1} + \log(4/\delta)}{2m}}$ *and ϵ_n is defined analogously.*

Remark. The error term ϵ_m bounds the amount by which the probability constraint may be violated and contributes $q^*(\alpha+\epsilon_m) - q^*(\alpha)$ to the possible regret. Given the number of data samples m, penalty λ_1 (λ_0 does not appear in this simplified bound) could be chosen to keep ϵ_m small, although user preferences for rule complexity are likely to be more important in setting λ_0, λ_1. Given λ_1, the number of uniform samples n could in turn be chosen to reduce ϵ_n. Note that ϵ_m, ϵ_n are largely controlled by λ_1 and depend only logarithmically on the dimension d.

3.4.2 Estimation of $\mathcal{B}^\epsilon$ as Binary Classification

To estimate the set $\mathcal{B}^\epsilon$ of inputs with bounded group propensity $\eta_t(X) := p(T = t \mid X)$, we follow in the tradition of using black-box (potentially non-parametric) estimators of propensity to identify overlapping or balanced cohorts in the study of causal effects (Crump et al., 2009; Fogarty et al., 2016). This is typically done by fitting a classifier (e.g., logistic regression) for predicting T given X, and letting $\hat{\eta}_t(x)$ be the estimated probability of class t for input x. Given such an estimate, we assign a label $\tilde{b}_i$ to each data point $x_i \in \mathcal{D}$ indicating significant propensity for every group,

$$\forall i \in [m] : \tilde{b}_i = \prod_{t \in \mathcal{T}} \mathbb{1}[\hat{\eta}_t(x_i) \geq \epsilon] \ . \tag{3.8}$$

Let $\tilde{\mathcal{B}} = \{x_i : \tilde{b}_i = 1\}$. Similar to the case of $\mathcal{S}^\alpha$, we may now reduce estimation of $\mathcal{B}^\epsilon$ to binary classification. Given $\hat{\mathcal{S}}$, the minimizer of (3.7), we again set up a Neyman-Pearson-like classification problem, now regarding the intersection $\hat{\mathcal{S}} \cap \tilde{\mathcal{B}}$ as the positive class:

$$\begin{aligned} \hat{\mathcal{B}} := \arg\min_{C} \quad & \frac{1}{|\hat{\mathcal{S}} \setminus \tilde{\mathcal{B}}|} \sum_{i : x_i \in \hat{\mathcal{S}} \setminus \tilde{\mathcal{B}}} \mathbb{1}[x_i \in \mathcal{C}] + R(\mathcal{C}) \\ \text{subject to} \quad & \sum_{i : x_i \in \hat{\mathcal{S}} \cap \tilde{\mathcal{B}}} \mathbb{1}[x_i \in \mathcal{C}] \geq \beta |\hat{\mathcal{S}} \cap \tilde{\mathcal{B}}| \ . \end{aligned} \tag{3.9}$$

The sets $\hat{\mathcal{S}} \setminus \tilde{\mathcal{B}}$ and $\hat{\mathcal{S}} \cap \tilde{\mathcal{B}}$ are defined by the solution to (3.7) and the base estimator (3.8). To accommodate the policy evaluation setting described in Section 3.3, we can modify the pseudo-labels defined in (3.8) to be $\tilde{b}_i(\pi) = \prod_{t \in \pi(x_i)} \mathbb{1}[\hat{p}(T = t \mid X = x_i) \geq \epsilon]$, where $\pi(x_i) := \{t : \pi(t|x_i) > 0\}$, and solve (3.9) using $\tilde{\mathcal{B}}(\pi) = \{x_i : \tilde{b}_i(\pi) = 1\}$ in place of $\tilde{\mathcal{B}}$. The resulting full procedure is given in the supplement.

Generalization of the final estimator. In the supplement, we state and prove a theorem bounding the generalization error of our final estimator, $\hat{\mathcal{O}} = \hat{\mathcal{S}} \cap \hat{\mathcal{B}}$. It shows that for good base estimators $\hat{\mathcal{S}}, \tilde{\mathcal{B}}$, the error of $\hat{\mathcal{O}}$ with respect to the true overlap $\mathcal{O}$ is dominated by its error with respect to the base estimators. Hence, practitioners may make an informed tradeoff between accuracy and interpretability based on this metric.

3.4.3 Optimizing Boolean Rules

Next, we describe a procedure for optimizing (3.7) over a class $\mathcal{C}$ of Boolean DNF rules. The same procedure also solves (3.9).

We assume that base features X have been binarized to form literals such as (Age > 30) or (Sex $=$ Female), as is standard in e.g. decision tree learning. A conjunction may thus be represented as the product of binary indicators of these literals. We let $\mathcal{K}$ index the set of all possible (exponentially many) conjunctions of literals, e.g. (Age $> 30) \wedge$ Female. Then, for $k \in \mathcal{K}$, let $a_{ik} \in \{0, 1\}$ denote the value taken by the k-th conjunction at sample x_i. Let the DNF rule be parametrized by $r \in \{0, 1\}^{|\mathcal{K}|}$ such that $r_k = 1$ indicates that the k-th conjunction is used in the rule.

Define an error variable ξ_i for i in $\mathcal{U} \cup \mathcal{I}$ representing the penalty for covering or failing to cover point i, depending on its set membership. Then, problem (3.7) may be reformulated as follows,

$$\begin{aligned}
&\underset{r}{\text{minimize}} \quad \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} \xi_i + R(r) \qquad (3.10)\\
&\text{subject to} \begin{cases} r_k \in \{0,1\},\ k \in \mathcal{K}, \\ \xi_i \geq 1 - \displaystyle\sum_{k\in\mathcal{K}} a_{ik} r_k, \quad \xi_i \geq 0,\ i \in \mathcal{I}, \\ \displaystyle\sum_{i\in\mathcal{I}} \xi_i \leq (1-\alpha)m \\ \xi_i = \displaystyle\max_{k\in\mathcal{K}}(a_{ik} r_k),\ i \in \mathcal{U}. \end{cases}
\end{aligned}$$

Problem (3.10) is an IP with an exponential number of variables and is intractable as written. We follow the column generation approach of Dash et al. (2018) to effectively manage the large number of variables and solve (3.10) approximately. As in that previous work, we bound from above the max in the last constraint of (3.10) with the sum (Hamming loss instead of zero-one loss) as it gives better numerical results. The choice of regularization in (3.6) implies $R(r) = \sum_{k\in\mathcal{K}} \lambda_k r_k$ with $\lambda_k = \lambda_0 + \lambda_1 p_k$. Thus the objective becomes linear in r, $\sum_{k\in\mathcal{K}} \left(1/|\mathcal{U}| \sum_{i\in\mathcal{U}} a_{ik} + \lambda_k\right) r_k$, and the ξ_i, $i \in \mathcal{U}$ constraints are absorbed into the objective. We then follow the overall procedure in (Dash et al., 2018) of solving the linear programming (LP) relaxation, using column generation to add variables only as needed.

We make the following departures from Dash et al. (2018). As noted, (3.10) has a constraint on false negative rate instead of a corresponding objective term and a complexity penalty $R(r)$ while Dash et al. (2018) use a constraint. As a result, the LP reduced costs, needed for column generation, are different. With dual variables $\mu_i \geq 0$, $i \in \mathcal{I}$ corresponding to the ξ_i, $i \in \mathcal{I}$ constraints in (3.10), the reduced cost of conjunction k is now $1/|\mathcal{U}| \sum_{i\in\mathcal{U}} a_{ik} + \lambda_k - \sum_{i\in\mathcal{I}} \mu_i a_{ik}$, which remains a linear function of a_{ik}, allowing the same column generation method to be used. We also avoid the need for an IP solver as used in Dash et al. (2018) by a) solving the column generation problem using a beam search algorithm from (Wei et al., 2019), and b) restricting (3.10) to the final columns once column generation terminates, converting to a weighted set cover problem, and applying a greedy algorithm to obtain an integer solution.

3.5 Experiments

In our experiments, we seek to address the following questions, while relating the performance of OverRule to that of MaxBox (MB) (Fogarty et al., 2016), which is also designed to produce interpretable study populations. (i) **Why is support estimation important?** In Section 3.5.1 we give a conceptual illustration using the Iris dataset, where MaxBox returns a description that empirically includes a large space outside of the true overlap region. (ii) **How well does OverRule approximate the base estimators / true overlap region?** In Section 3.5.2 we use the Jobs (LaLonde, 1986) dataset to show that performance of OverRule is comparable to that of the base estimators, and generally surpasses the performance of MaxBox. (iii) **Do the resulting rules yield any insights?** We apply OverRule to overlap estimation in two real-world clinical datasets on (1) post-surgical opioid prescriptions, and (2) policy evaluation in antibiotic prescriptions. For the former, we conducted a user study with three clinicians to interpret and critique the output, with additional comparison to the output of MaxBox.

OverRule and MaxBox algorithms are both *meta-algorithms* in the sense that they take (as input) labels indicating whether each data point is in the overlap set. To generate these labels, we use a variety of base overlap estimators: (i) *Covariate Bounding Boxes*: The intersection of covariate (marginal) bounding boxes (CBB), analogous to classical balance checks in causal inference. The bounding boxes are selected to cover the $[(1-\alpha)/2, (1+\alpha)/2]$ quantiles of the data. (ii) *Propensity Score Estimators*: Standard propensity score estimators as described in (3.8) and Crump et al. (2009) with logistic regression (PS-LR) or k-nearest neighbors (PS-kNN) estimates of the propensity. These can be viewed as a binary version of overlap weights (Li et al., 2018b). (iii) *One-Class SVMs*: One-Class Support Vector Machines (OSVM) to first estimate conditional supports and then use their intersection as overlap labels. Details

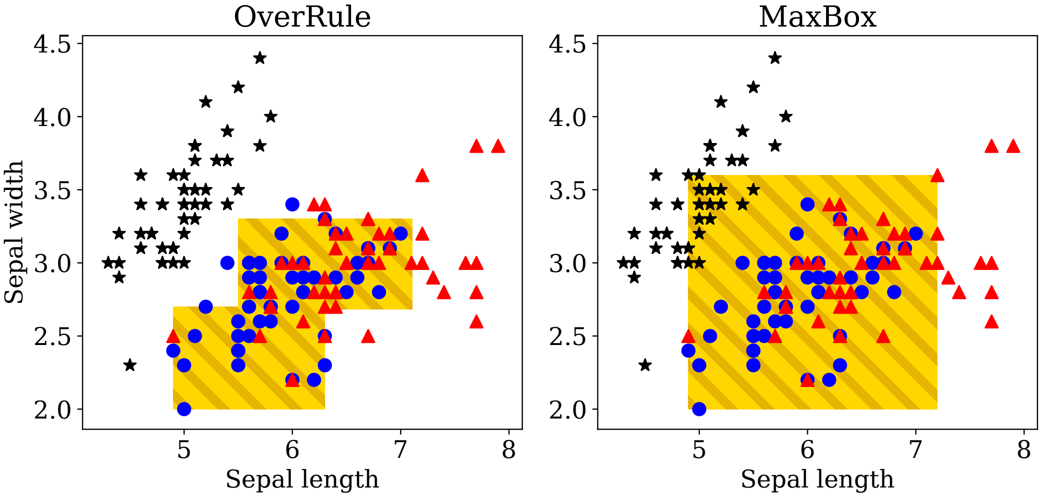


Figure 3-2: *Overlap (orange stripes) between Versicolor (blue circles) and Virginica (red triangles) species in the Iris dataset as identified by OverRule (left) and MaxBox (right) using the same base estimator of propensity. Black stars indicate samples of the (unobserved) Setosa species. We see that MaxBox identifies several of the Setosa samples as being in the overlap set, despite it being outside of the support of the observed data.*

on hyperparameter selection and feature binarization are given in the supplement, along with general guidance on hyperparameter selection depending on user goals, from optimizing an observable metric (e.g., accuracy w.r.t the base estimator), to generating shorter rule sets, to exploring structure in the data.

3.5.1 Illustrative Example: Iris

We use the Iris dataset to illustrate the importance of combining explicit support estimation (lacking in MaxBox) with an interpretable characterization of the overlap region (lacking in propensity score models). We use OverRule to identify the overlap between members of two species of Iris, as represented by their sepal and petal dimensions. In Figure 3-2, we visualize the estimates $\hat{\mathcal{O}}$ learned using OverRule and MaxBox in the space of sepal length and width. In contrast, the coefficients of a logistic regression propensity score model, $[-1.7, -1.5, 2.5, 2.6]^\top$ reveal very little about which points lie in the overlap set.

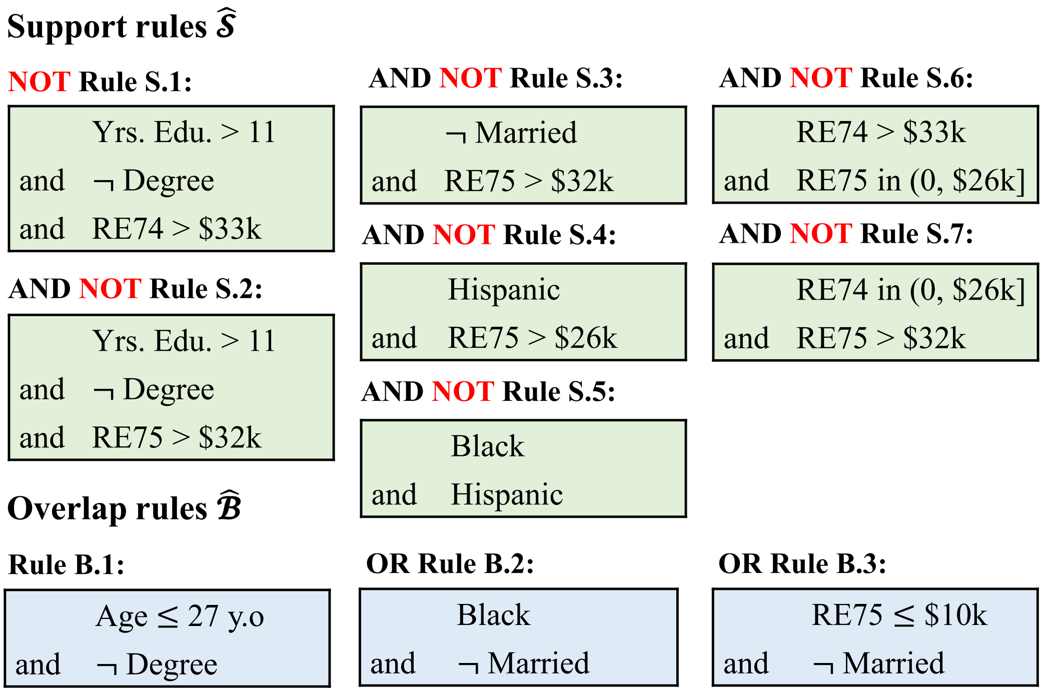
Table 3.1: *Overlap estimation in Jobs. Balanced accuracy (Acc), false positive rate (FPR), false negative rate (FNR), and number of literals (L) with standard deviations over 5-fold CV. MB and OR indicate MaxBox and OverRule. MB did not run with CBB.*

	Acc	**FPR**	**FNR**	**L**
Baselines (base estimators):				
CBB	0.75 ± 0.02	0.12 ± 0.01	0.38 ± 0.03	—
OSVM	0.82 ± 0.01	0.22 ± 0.03	0.14 ± 0.02	—
PS-k-NN	0.90 ± 0.02	0.14 ± 0.02	0.05 ± 0.02	—
PS-LR	0.96 ± 0.01	0.10 ± 0.01	0.09 ± 0.03	—
MaxBox with base estimator:				
OSVM	0.68 ± 0.01	0.09 ± 0.02	0.54 ± 0.01	16
PS-kNN	0.84 ± 0.01	0.03 ± 0.01	0.29 ± 0.02	16
PS-LR	0.80 ± 0.02	0.04 ± 0.01	0.35 ± 0.04	16
OverRule with base estimator:				
CBB	0.83 ± 0.01	0.16 ± 0.01	0.19 ± 0.02	20
OSVM	0.84 ± 0.02	0.25 ± 0.03	0.07 ± 0.02	23
PS-kNN	0.89 ± 0.02	0.16 ± 0.02	0.06 ± 0.02	40
PS-LR	0.88 ± 0.02	0.15 ± 0.04	0.09 ± 0.01	21

3.5.2 Job Training Programs

In this section, we demonstrate that OverRule compares favorably to MaxBox in terms of approximating both the derived overlap labels (using a base estimator), as well as the "ground truth" overlap labels in a real dataset. To do so, we use data from a famous trial performed to study the effects of job training (LaLonde, 1986; Smith and Todd, 2005), in which eligible US citizens were randomly selected into ($T = 1$), or left out of ($T = 0$) job training programs. The RCT ($E = 1$), which satisfies overlap by definition, has since been combined with non-experimental control samples ($E = 0, T = 0$), forming a larger observational set (Jobs), to serve as a benchmark for causal effect estimation (LaLonde, 1986). Here, we aim to characterize the overlap between treated and control subjects.

Due to the trial's eligibility criteria, the experimental and non-experimental cohorts barely overlap; standard logistic regression separates the experimental and non-experimental groups with held-out balanced accuracy of 0.96. Since all treated



Support rules $\widehat{\mathcal{S}}$

NOT Rule S.1:
Yrs. Edu. > 11
and ¬ Degree
and RE74 > \$33k

AND NOT Rule S.2:
Yrs. Edu. > 11
and ¬ Degree
and RE75 > \$32k

AND NOT Rule S.3:
¬ Married
and RE75 > \$32k

AND NOT Rule S.4:
Hispanic
and RE75 > \$26k

AND NOT Rule S.5:
Black
and Hispanic

AND NOT Rule S.6:
RE74 > \$33k
and RE75 in (0, \$26k]

AND NOT Rule S.7:
RE74 in (0, \$26k]
and RE75 > \$32k

Overlap rules $\widehat{\mathcal{B}}$

Rule B.1:
Age ≤ 27 y.o
and ¬ Degree

OR Rule B.2:
Black
and ¬ Married

OR Rule B.3:
RE75 ≤ \$10k
and ¬ Married

Figure 3-3: *OverRule description of the overlap region $\mathcal{O}$ in the Jobs dataset learned using the LR propensity base estimator, achieving held-out balanced accuracy of 0.88. ¬ indicates a negation, and CNF support rules are given with rule-level negations applied for readability. If* none *of the support rules (top) and* any *of the overlap rules (bottom) apply, a subject is in $\mathcal{O}$.*

subjects were part of the experiment, the experimental cohort perfectly represents the overlap region. For this reason, we use the experiment indicator E as ground truth for $\mathcal{O}$, at the risk of introducing a small number of false negatives. In studies of causal effects in this data, the following features were included to adjust for confounding: Age, #Years of education (Educ), Race (black/hispanic/other), Married, No degree (NoDegr), Real earnings in 1974 (RE74) and 1975 (RE75). These are the features X for which we estimate overlap.

We present results in Table 3.1 and Figure 3-3, where all balanced accuracies are w.r.t. the ground truth indicator E. For the propensity base estimators, the OverRule approximations achieve slightly lower balanced accuracies than the base estimator, but with a simpler description, while for the other base estimators the accuracy is actually better. OverRule compares favorably to MaxBox on balanced accuracy, although MaxBox generally achieves a lower FPR, likely because it does not try to retain a fixed fraction β of the overlap set. In the supplement, we show that the held-out balanced accuracy quickly converges as the number of literals in the rules increases and correlates strongly with the quality by which the rule set approximates the base estimator.

The learned support rules in Figure 3-3 demonstrate that support estimation can find gaps in the dataset that are intuitive, such as a lack of individuals with high income but no degree (Rules S.1-2) or whose income changes dramatically from 1974 to 1975 (Rules S.6-7). The learned overlap rules conform to expectations, as the eligibility criteria for the RCT allow only subjects who were currently unemployed and had been so for most of the time leading up to the trial—factors that correlate with age and education (Rule B.1), previous income (Rule B.3), and marital status (Rules B.2-3) (Smith and Todd, 2005).

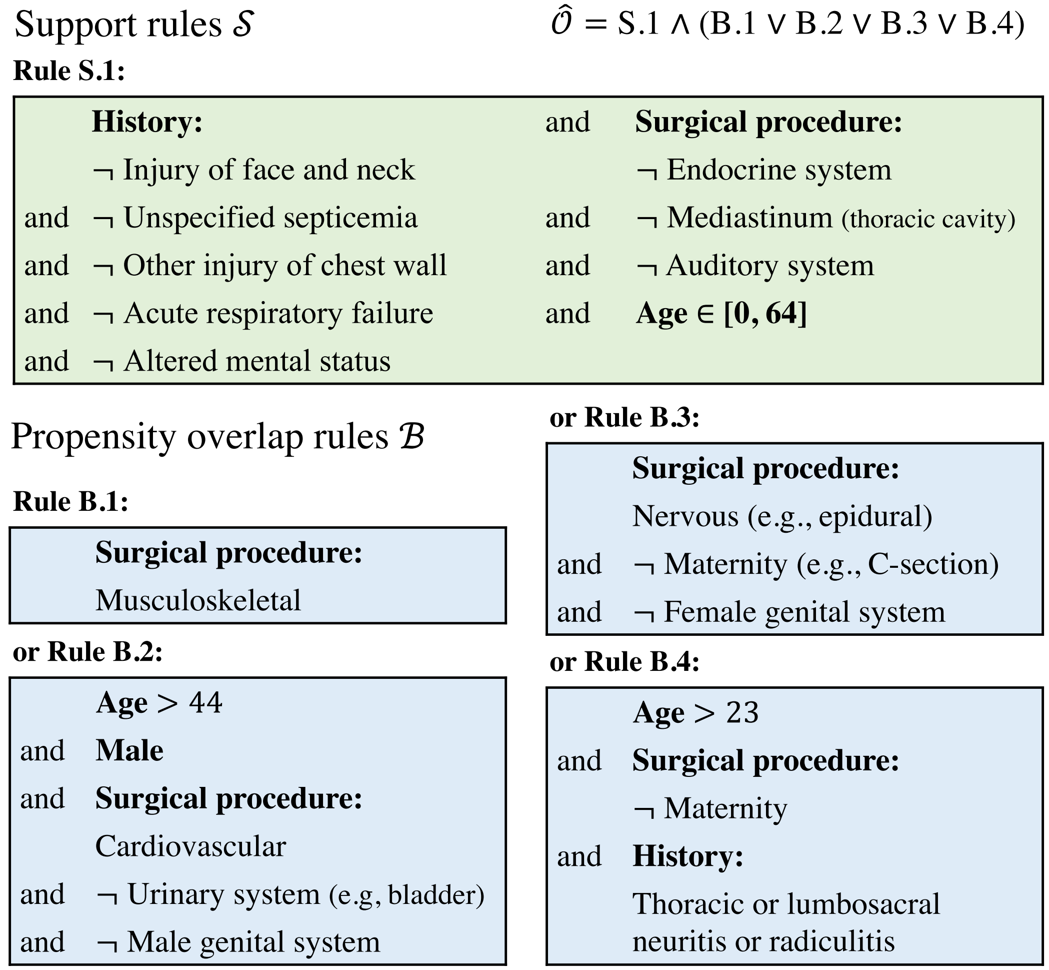
3.5.3 Post-surgical Opioid Prescriptions

Opioid addiction affects millions of Americans. Understanding the factors that influence the risk of addiction is thus of great importance. To this end, Brat et al. (2018) and Zhang et al. (2017) study the effect of choices in opioid prescriptions on the risk of future misuse. Here, we study a group of *post-surgical* patients who were given opioid prescriptions within 7 days of surgery, replicating the cohort eligibility criteria of Brat et al. (2018) using a subset of the MarketScan insurance claims database. We compare groups of patients with morphine milligram equivalent (MME) doses above and below the 85th percentile in the cohort, MME=450. Subjects were represented by basic demographics (age, sex), diagnosis history, and procedures billed as surgical on the index date (not mutually exclusive). Cohort statistics are given in the supplement. We fit three models: An OverRule model (OR) using DNF support rules and a random forest base estimator, a MaxBox model (MB) (Fogarty et al., 2016) with the same base estimator, and another OverRule model describing the complement of $\mathcal{O}$ (OR-C). The balanced accuracies of these models w.r.t. the base were 0.90 (OR), 0.77 (MB) and 0.92 (OR-C). Learning took 10 minutes for OverRule (Python) and 7 minutes for MaxBox (R). Other hyperparameter details are in the supplement.

In Figure 3-4, we summarize the rules learned by OR which cover 27% of the overall population. MB learned: (Musculoskeletal surg. ∧ ¬Mediastinum surg. ∧ ¬Male genital surg. ∧ ¬Maternity surg. ∧ ¬Lumbosacral spondylosis without myelopathy) which covers 17% of patients. The rules learned by OR-C are presented in the supplement.

To evaluate the interpretability of the output, we conducted a qualitative user study through a moderated discussion with three participants: two attending surgeons (P1 & P2) and a 4th year medical student (P3) at a large US teaching hospital. Before seeing the outputs of any method, the participants were asked to give their expectations for what to find in the overlap set.

The participants expected that the overlap set would mostly correspond to patients in the higher dose range, as these patients are often considered also for smaller doses,



Support rules $\mathcal{S}$ $\hat{O} = \text{S.1} \land (\text{B.1} \lor \text{B.2} \lor \text{B.3} \lor \text{B.4})$

Rule S.1:

History:
¬ Injury of face and neck
and ¬ Unspecified septicemia
and ¬ Other injury of chest wall
and ¬ Acute respiratory failure
and ¬ Altered mental status
and **Surgical procedure:**
¬ Endocrine system
and ¬ Mediastinum (thoracic cavity)
and ¬ Auditory system
and **Age ∈ [0, 64]**

Propensity overlap rules $\mathcal{B}$

Rule B.1:

Surgical procedure:
Musculoskeletal

or Rule B.2:

Age > 44
and **Male**
and **Surgical procedure:**
Cardiovascular
and ¬ Urinary system (e.g, bladder)
and ¬ Male genital system

or Rule B.3:

Surgical procedure:
Nervous (e.g., epidural)
and ¬ Maternity (e.g., C-section)
and ¬ Female genital system

or Rule B.4:

Age > 23
and **Surgical procedure:**
¬ Maternity
and **History:**
Thoracic or lumbosacral neuritis or radiculitis

Figure 3-4: *OverRule description of post-surgical patients likely to receive both high and low opioid doses. A patient is in the overlap set if the support rule (top) applies and* any *propensity overlap rule (bottom) applies. ¬ indicates negation. The rules cover 27% of patients with balanced accuracy of 0.90 w.r.t. the base estimator. Surgical procedures are not mutually exclusive.*

and that overlap would be driven largely by surgery type. All participants expected Musculoskeletal and Cardiovascular surgery patients to be predominantly in the higher dose group, and sometimes in the lower, and one suggested that Maternity surgeries (e.g., C-sections) would be only in the lower range. These comments are all consistent with the findings of OverRule, which identified all of these surgery types as important. MaxBox identified only Musculoskeletal surgery patients as overlapping. One participant expected history of psychiatric disease and Tobacco use disorder to be predictive of higher prescription doses for some patients, and thus overlap. Neither method identified psychiatric disease, but Tobacco use disorder was identified by OR-C as predictive (see supplement).

The participants found the support rules ($\hat{\mathcal{S}}$) output by OR (Figure 3-4 top) intuitive. P1 stated that Endocrine surgeries are not typically followed by opioid prescriptions. They found the MaxBox and OR rule descriptions easy to interpret, and discussion focused on their clinical meaning. The first three propensity overlap rules B.1-B.3 were all consistent with expectation as described above, with the caveat that Cardiovascular patients are not typically stratified by Urinary and Genital surgeries. This was later partially explained by catheters being billed as Urinary and P3 interpreted this as a proxy for more severe Cardiovascular surgeries. P1 pointed out the value in discovering such surprising patterns that may be hidden in black-box analyses. The OR-C rules were found hard to interpret due to many double negatives ("excluded from exclusion"), but were ultimately deemed clinically sound.

Remark: We noted that these support rules primarily exclude individually rare features, in lieu of e.g., finding that certain non-rare surgery types do not co-occur. This motivated both (1) an empirical study (w/semi-synthetic data) of how support rule hyperparameters influence the recovery of these interactions, and (2) the generation of new rules. Both are in the supplement.

3.5.4 Policy Evaluation of Antibiotic Prescription Guidelines

Using the policy evaluation formulation of $\mathcal{B}^\epsilon(\pi)$ (Section 3.3.1), we apply OverRule to assess overlap for a policy that follows clinical guidelines published by the Infectious Disease Society of America (IDSA) for treatment of uncomplicated urinary tract infections (UTIs) in female patients (Gupta et al., 2011). Using medical records from two academic medical centers, we apply OverRule to a cohort of 65,000 UTI patients to test whether it can recover a clinically meaningful overlap set. From a qualitative perspective, we discussed the resulting rules with an infectious disease specialist, who verified that they have a clear clinical interpretation as identifying primarily outpatient cases and uncomplicated inpatient cases, which are where the guidelines are applied in practice. Detailed results (including quantitative results) are given in the supplement.

3.6 Conclusion

We have presented OverRule—an algorithm for learning rule-based characterizations of overlap between populations, or the inputs for which policy evaluation from observational data is feasible. The algorithm learns to exclude points that are marginally out-of-distribution, as well as points where some population/policy has low density. We gave theoretical guarantees for the generalization of our procedure and evaluated the algorithm on the task of characterizing overlap in observational studies. These results demonstrated that our rule descriptions often have similar accuracy to black-box estimators and outperform a competitive baseline. In an application to study treatment-group overlap in post-surgical opioid prescription, a qualitative user study found the results interpretable and clinically meaningful. Similar observations were made in an application to evaluation of antibiotic prescription policies. Future research challenges include investigating the scalability of the method with the dimensionality of the input.

Chapter 4

Falsification before Extrapolation in Causal Effect Estimation

This chapter (and accompanying appendix) was previously published as (Hussain et al., 2022) at NeurIPS 2022, and is presented here with minor typographical changes.

4.1 Introduction

Policy guidelines often rely on conclusions from Randomized Controlled Trials (RCTs), whether considering treatment decisions in healthcare, classroom interventions in education, or social programs in economics (Keum et al., 2019; Cloyd et al., 2020; Prete et al., 2018). In healthcare, when a target population has reasonable overlap with the inclusion criteria of RCTs, current clinical treatment guidelines rely primarily on RCTs (Guyatt et al., 2008b,a). For target populations not well-represented in RCTs, observational studies are often used to infer treatment effects. However, different observational estimates can give conflicting conclusions. We give an example of this tension when looking at a new chemotherapy for multiple myeloma.

Example 4.1 (*Carfilzomib-based Combination Therapy for Newly Diagnosed Multiple Myeloma (NDMM)*)**.** Until 2020, the effect of Carfilzomib-based combination therapy in the NDMM subpopulation had not been studied via an RCT. However, a trial (ASPIRE) in 2015 measured the effect of Carfilzomib-based therapy on survival in Relapsed & Refractory Multiple Myeloma (RRMM) patients (Stewart et al., 2015). The CoMMpass trial, an observational dataset, was also available in which the Carfilzomib regimen was given to both NDMM and RRMM patients (NIH, 2016). Several analyses on the CoMMpass dataset to estimate the effect of Carfilzomib-based therapy on NDMM patients led to different, sometimes opposing, conclusions on the benefit of the therapy in this subpopulation (Li et al., 2018a; Landgren et al., 2018).

A traditional meta-analysis approach would combine observational estimates under the assumption that differences arise only due to random variation, and not e.g., differences in confounding bias (Higgins et al., 2019, Section 10.10.4.1). This is unlikely to be true in practice. For instance, in Example 4.1, the two studies in question made different choices in e.g., how to adjust for confounders. *In the work presented in this chapter, we relax the assumption that all observational estimates are valid.* Instead, we assume that at least one observational estimate is valid across all subpopulations. In the context of Example 4.1, we might assume that at least one of the candidate observational studies yields consistent and asymptotically normal estimates of the effects in both the NDMM and RRMM populations. While we cannot *verify* that any given estimator is valid for all subpopulations, we can *falsify* this claim of validity if an estimator is inconsistent for the causal effects identified by the RCT (e.g., RRMM). Hence, we use the term *validation effects* to refer to causal effects in subpopulations that overlap between the observational and randomized datasets (e.g., RRMM), and use the term *extrapolated effects* to refer to those only covered by observational datasets (e.g., NDMM).

We propose a meta-algorithm that combines two key ideas: falsification of estimators, and pessimistic combination of confidence intervals. We first aim to falsify candidate estimators using hypothesis testing, rejecting those that fail to replicate the RCT estimates of validation effects. In Section 4.2.2, we motivate this approach with examples of observational estimates based on different causal assumptions, showing that hypothesis tests based on asymptotic normality can be applied even when causal assumptions fail to hold. Then, we combine accepted estimators to get confidence intervals on the extrapolated effects. Since failure to reject does not imply validity,[1] we return an interval that contains every confidence interval of the accepted estimators. We demonstrate theoretically that if *at least one* candidate estimator is consistent for both the validation and extrapolated effects, then the intervals returned by our algorithm provide valid asymptotic coverage of the true effects.

In scenarios where the covariate distribution differs across datasets, estimators that "transport" the causal effect should be used Pearl and Bareinboim (2014); Dahabreh et al. (2019, 2020). Furthermore, in the case of high-dimensional covariates, flexible machine learning methods are required to estimate nuisance functions, which can affect the hypothesis tests due to their slower convergence rates. In light of this, we adapt estimators of the average treatment effect in this setting to provide estimates of group-wise treatment effects, and show (via the framework of double machine learning (Chernozhukov et al., 2018; Semenova and Chernozhukov, 2021)) that this estimator enjoys asymptotic normality under mild conditions on convergence rates of the nuisance function estimators. Our conclusions are supported by semi-synthetic experiments, based on the IHDP dataset, as well as real-world experiments, based on clinical trial and observational data from the Women's Health Initiative (WHI), that demonstrate various characteristics of our meta-algorithm.

4.2 Setup and Motivating Examples

4.2.1 Notation and Assumptions

Let $Y \in \mathcal{Y}$ denote an outcome of interest, and $A \in \{0, 1\}$ denote a binary treatment. We use Y_a to denote the potential outcome of an individual under treatment $A = a$. We use $X \in \mathcal{X}$ to denote all other covariates. To distinguish between different sampling distributions (i.e., datasets), we use the random variable $D \in \{0, \dots J\}$, where $J \geq 1$

[1] For instance, we could fail to reject due to low power, or because falsification is impossible, due to differences in causal structure across subpopulations, as discussed in Appendix B.1.

is the number of observational datasets, and $D = 0$ is reserved for the sampling distribution of the randomized trial. We let $\mathbb{P}(Y_1, Y_0, Y, A, X, D)$ denote the joint distribution over all variables, including unobserved potential outcomes. For instance, $\mathbb{P}(Y_1, Y_0, X \mid D = 0)$ denotes the distribution of potential outcomes and covariates in the RCT.

We seek to estimate conditional average treatment effects for a finite set of I subgroups $\{\mathcal{G}_i\}_{i=1}^{I}$. We assume subgroups are defined a-priori by a function $G : \mathcal{X} \mapsto \{1, \ldots, I\}$, such that $G = i$ indicates that $X \in \mathcal{G}_i$. We use observational data precisely because not all groups are supported on the RCT dataset. To this end, we use $\mathcal{I}_R = \{i : \mathbb{P}(G = i \mid D = 0) > 0\}$ to denote the set of subgroups supported on the RCT dataset, and we let $\mathcal{I}_O$ denote the complement $\{1, \ldots, I\} \setminus \mathcal{I}_R$. We use $|\mathcal{I}_R|$ to denote the cardinality of a set, and assume that every observational dataset has support for all groups.

Assumption 4.1 (Support)**.** We assume that for each $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J\}$, $\mathbb{P}(G = i, D = j) > 0$, i.e., all observational datasets ($D \geq 1$) have support for all groups.

Definition 4.1 (Validation and Extrapolated Effects)**.** We define the group average treatment effect (GATE)[2] as

$$\tau_i := \begin{cases} \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0], & \text{if } i \in \mathcal{I}_R \\ \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 1], & \text{if } i \in \mathcal{I}_O \end{cases} \tag{4.1}$$

and refer to τ_i for $i \in \mathcal{I}_R$ as a validation effect, and τ_i for $i \in \mathcal{I}_O$ as an extrapolated effect.

Here, we focus on discrete subgroups, in part to reflect the practical reality of comparing RCTs to observational studies, where we may have large observational datasets with rich covariates but only have access to the published results of the RCT, which often provides estimates (with confidence intervals) for subgroup effects but not the raw

[2] We use this term in line with the literature (Chernozhukov et al., 2017; Jacob, 2019; Park and Kang, 2019; Semenova and Chernozhukov, 2021) and to distinguish it from the CATE function.

data itself (SPRINT Research Group et al., 2015, Figure 4, for example). In Def. 4.1, we allow for the fact that different datasets may have different distributions of effect modifiers. To have a well-defined effect of interest, we have chosen the reference dataset $D = 1$ arbitrarily, but in principle we could choose any of the observational datasets. We discuss further nuances of this definition under Assumption 4.3. By Def. 4.1, we often write these effects as a vector $\tau \in \mathbb{R}^I$. We use $\hat{\tau}(k) \in \mathbb{R}^I$ to denote an estimator, where $k \in \{0, \ldots, K\}$, with $\hat{\tau}(0)$ reserved to denote the estimator derived from the RCT data. The remainder are observational estimators.[3] In general, we use "hat" notation to refer to estimators, and refer to their population quantities without a hat. We use N_k to denote the number of samples used by each estimator. Throughout, we will assume that the RCT estimator is consistent.

Assumption 4.2. The RCT estimator $\hat{\tau}(0)$ is a consistent estimator of the (supported) dimensions of τ, such that for each $i \in \mathcal{I}_R$, $\hat{\tau}_i(0)$ is consistent for τ_i.

Below, our central assumption states that at least one observational estimator also enjoys consistency. We discuss examples of specific observational estimators in Section 4.2.2.

Assumption 4.3. There exists at least one observational estimator $\hat{\tau}(k) \in \mathbb{R}^I$, $k \geq 1$ that is a consistent estimator of $\tau \in \mathbb{R}^I$, such that for each $i \in \{1, \ldots, I\}$, $\hat{\tau}_i(k)$ is consistent for τ_i.

Remark 1. Assumption 4.3 is our primary non-trivial assumption, and in Appendix B.2, we give one example of causal assumptions (for a given observational study) under which the entire GATE vector τ is **identifiable** from observational data, and give an **estimator** of the resulting observational quantity which is asymptotically normal (Pearl and Bareinboim, 2011, 2014; Pearl, 2015; Dahabreh et al., 2020; Degtiar and Rose, 2021). In order to compare observational estimates with experimental ones, Assumption 4.3 requires not only that the observational data is free of confounding, but also that the causal effect can be transported to the RCT population. This

[3] We define $\hat{\tau}(0)$ as a vector in $\mathbb{R}^I$ for simplicity of notation, allowing the entries $\hat{\tau}_i(0), i \in \mathcal{I}_O$ to be arbitrary.

can be done so long as relevant effect modifiers are observed in both the RCT and observational study, but the latter requirement is satisfied automatically (without requiring RCT data) if e.g., treatment effects are constant within each subgroup G, or if the distribution of effect modifiers is the same between the RCT and observational study, in which case $\mathbb{E}[Y_1 - Y_0 \mid D, G] = \mathbb{E}[Y_1 - Y_0 \mid G]$. This represents one (conservative) failure mode of our approach, in which we may reject an observational estimator due to failures in transportability, even if it yields unbiased estimates of the extrapolated effects.

Assumptions 4.2 and 4.3 imply that there exists an observational estimator $\hat{\tau}(k)$ such that both $\hat{\tau}_i(k)$ and the RCT estimate $\hat{\tau}_i(0)$ are both consistent for the validation effects τ_i, $\forall i \in \mathcal{I}_R$. To validate this implication in finite samples, we will construct a statistical test to compare $\hat{\tau}_i(k)$ and $\hat{\tau}_i(0)$. Our general approach could be modified to use any valid test, but to facilitate further analysis, as well as explicit construction of confidence intervals, we additionally assume the following:

Assumption 4.4. All GATE estimators are pointwise[4] asymptotically normally distributed. That is, for all $(k, i) \in \{1, , ..., K\} \times (\mathcal{I}_R \cup \mathcal{I}_O)$ and for all $(k, i) \in \{0\} \times \mathcal{I}_R$,

$$\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))/\hat{\sigma}_i(k) \overset{\mathrm{d}}{\to} \mathcal{N}(0, 1) \tag{4.2}$$

Here, $\overset{\mathrm{d}}{\to}$ denotes convergence in distribution, and $\hat{\sigma}_i^2(k)$ is an estimate of the variance that converges in probability to $\sigma_i^2(k)$, the asymptotic variance of $\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))$.

Assumption 4.4 requires each estimator $\hat{\tau}(k)$ to be consistent and asymptotically normal for some $\tau(k)$, which may **not** be equal to τ. This is not a particularly strong assumption, as we discuss below.

[4] Here, “pointwise” refers to the fact that each subgroup effect estimate is asymptotically normal.

4.2.2 Asymptotic Normality of Biased Estimators

In this section, we give two simple examples to illustrate the principle that multiple estimators $\hat{\tau}(k)$ may be asymptotically normal, even if they are asymptotically biased (i.e., $\tau(k) \neq \tau$). In both cases, there is a distinction between the *statistical* assumptions required to obtain asymptotic normality, and the *causal* assumptions required for $\tau(k)$ to identify the causal effect τ. For simplicity in both examples, we restrict to the setting of comparing one-dimensional estimates $\tau(k) \in \mathbb{R}$, which estimate the GATE, τ, in a single group $G = 1$ covered by all datasets. The statistical claims here also extend to GATE estimation with multiple groups (Semenova and Chernozhukov, 2021).

Example 4.2 (Variation in confounding across datasets)**.** Suppose that there is one estimator of the GATE per observational dataset, and each estimator seeks to estimate the population quantity, $\tau(k) = \mathbb{E}[g_k(1, X_k) - g_k(0, X_k) \mid G = 1, D = k]$, where X_k denotes the controls used in each study, and $g_k(A, X_k) \coloneqq \mathbb{E}[Y \mid A, X_k, D = k]$ and $m_k(X_k) \coloneqq \mathbb{P}(A = 1 \mid X_k, D = k)$. We assume that there exists some $\eta > 0$ such that for all x, k, $\eta < m_k(x) < 1 - \eta$. Note that $\tau(k)$ is only a *statistical* quantity: identifying this with the *causal* quantity (the GATE) requires additional assumptions like unconfoundedness, that $Y_a \perp\!\!\!\perp A \mid X_k$ for the given dataset D. This assumption may hold for some datasets, but not others, particularly if the set of observed confounders X_k differs across datasets.

Regardless of the interpretation of $\tau(k)$, one can construct estimators of it that are consistent and asymptotically normal using flexible machine learning estimators.[5] One approach, given in Chernozhukov et al. (2018), is to use double machine learning (DML), which employs cross-fitting to produce estimates $\hat{\tau}(k)$ based on the doubly-robust score (Robins and Rotnitzky, 1995), while using plug-in estimates $\hat{g}_k, \hat{m}_k$ based on machine learning models. This approach achieves asymptotic normality, $\sqrt{N_k}(\hat{\tau}(k) - \tau(k))/\hat{\sigma}^2(k) \xrightarrow{\text{d}} \mathcal{N}(0, 1)$, under regularity conditions that allow for flexible

[5] A rich literature focuses on establishing such results, beyond the approach in this example (Athey et al., 2016; Farrell, 2018; Wager and Athey, 2018; Oprescu et al., 2019; Athey et al., 2019).

machine learning estimators that converge at slower than parametric rates, and where $\hat{\sigma}^2(k)$ converges in probability to the variance of the doubly robust score (See Theorem 5.1 of Chernozhukov et al., 2018, for additional details). These results hold whether or not $\tau(k) = \tau$, as discussed in Footnote 9 of Chernozhukov et al. (2018). For simplicity, we have focused on the case where $\mathbb{E}[Y_1 - Y_0 \mid G = 1]$ is constant across datasets. When this does not hold, certain conditions enable valid transportation of treatment effects across datasets (Degtiar and Rose, 2021) with the use of transported estimators (Dahabreh et al., 2020) (see Appendix B.2 for details).

Example 4.3 (Selection of Adjustment Strategy)**.** Consider the two causal graphs given in Figure 4-1, and assume that all variables are binary. Each graph suggests a different identification strategy for the causal effect, $\mathbb{E}[Y \mid do(A = a), G = 1]$. In Figure 4-1a, this is identified by the (observational) quantity $\mathbb{E}[Y \mid A = a, G = 1]$, and in Figure 4-1b, by front-door adjustment (Pearl, 1995) as $\sum_M P(M \mid a, G = 1) \sum_{A'} \mathbb{P}(Y \mid M, A', G = 1)\mathbb{P}(A' \mid G = 1)$.

These observational quantities will typically differ: the one that represents the true interventional effect depends on which graph reflects the true causal structure. However, in the case where all variables are discrete and low-dimensional, we can still construct asymptotically normal estimators for both observational quantities.[6] For more complex settings (e.g., requiring regularized ML models for estimating conditional distributions) asymptotic normality has been established under certain conditions for general graphs (Bhattacharya et al., 2020; Jung et al., 2021)

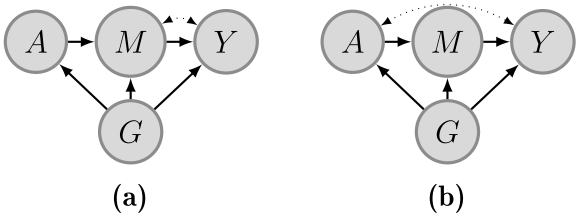


Figure 4-1: *(Ex. 4.3) In (a), M and Y are confounded by unobservables (bi-directional dotted arrow). In (b), A and Y are confounded, but the causal effect is identified via front-door adjustment.*

[6] This follows from the use of maximum likelihood (i.e., empirical counts) for estimating each conditional distribution, and applying the delta method to the front-door estimator.

Remark 2. In each example, there are multiple estimators available, each asymptotically normal under basic statistical assumptions, but potentially biased in the sense that $\tau(k) \neq \tau$. In the first example, this bias occurs if X is not sufficient to control for confounding in all observational datasets. In the second, this bias arises in a given estimator if the causal graph is incorrectly specified. Assumption 4.3 corresponds to assuming that both the statistical assumptions and causal assumptions hold for one of the candidate estimators, e.g., X is sufficient to control for confounding in at least one study (Example 4.2), or that one of the causal graphs is correct (Example 4.3).

4.2.3 Asymptotic Normality of GATE Estimators with Transportation

Example 4.2 assumes that $\mathbb{E}[Y_1 - Y_0 \mid G = i]$ is constant across datasets. In practice, it may be necessary to correct for differences (not captured by group indicators) between the observational and RCT populations. There exist estimators for the ATE in this setting under mild additional assumptions Dahabreh et al. (2020, 2019). These extend in a straightforward way to estimators of the GATE, but proving asymptotic normality is nuanced in high-dimensional settings when using flexible machine learning methods to estimate nuisance functions. For completeness, inspired by Semenova and Chernozhukov (2021), we demonstrate that a doubly-robust GATE estimator for this setting is asymptotically normal under reasonable conditions (Assumption B.3.1 to B.3.5). Details on the estimator, and the corresponding proof of normality, are given in Appendix B.3, and may be of independent interest.

4.2.4 Testing for Bias under Asymptotic Normality

Under Assumption 4.4, each observational estimate $\hat{\tau}_i(k)$ can be compared to the estimate from the randomized trial $\hat{\tau}_i(0)$ for $i \in \mathcal{I}_R$, the groups with common support. Since the observational and randomized datasets are distinct, we can conclude that each $\hat{\tau}_i(k)$ is independent of $\hat{\tau}_i(0)$, and use this to test for the hypothesis that $\tau_i(k) = \tau_i$.

Proposition 4.1. *For an observational estimator $\hat{\tau}(k)$, assume Assumptions 4.2 and 4.4 hold. Furthermore, let $N = N_k + N_0$ with fixed proportions, where $N_k = \rho N, N_0 = (1-\rho)N$ for $\rho \in (0,1)$. Define the test statistic*

$$\hat{T}_N(k,i) \coloneqq \frac{\hat{\tau}_i(k) - \hat{\tau}_i(0) - \mu_i(k)}{\hat{s}} \tag{4.3}$$

where $\hat{s}^2 \coloneqq \frac{\hat{\sigma}_i^2(k)}{N_k} + \frac{\hat{\sigma}_i^2(0)}{N_0}$ is the estimated variance, and $\mu_i(k) \coloneqq \tau_i(k) - \tau_i$. This test statistic converges in distribution to a normal distribution as $N \to \infty$, $\hat{T}_N(k,i) \xrightarrow{d} \mathcal{N}(0,1)$.

We present the proof for Proposition 4.1 in Appendix B.4. This asymptotic normality allows for the construction of simple hypothesis tests. For instance, one can construct a Wald test for $H_0 : \tau_i(k) = \tau_i$, with asymptotic level α by setting $\mu_i(k) = 0$ in Equation (4.3) and rejecting H_0 whenever, $|\hat{T}_N(k,i)| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the normal CDF. Moreover, the asymptotic power of this test (the probability of correctly rejecting H_0) is given by

$$1 - \Phi\left(\frac{\mu_i(k)}{\sigma_{k,0}} + z_{\alpha/2}\right) + \Phi\left(\frac{\mu_i(k)}{\sigma_{k,0}} - z_{\alpha/2}\right) \tag{4.4}$$

where $\sigma_{k,0}^2 \coloneqq \frac{\sigma^2(k)_i}{N_k} + \frac{\sigma_i^2(0)}{N_0}$ (see Theorems 10.4, 10.6 of Wasserman, 2004). Likewise, Assumption 4.4 implies an asymptotic $1-\alpha$ confidence interval for $\tau_i(k)$ as

$$[\hat{L}_i(k)(\alpha), \hat{U}_i(k)(\alpha)] \coloneqq \left[\hat{\tau}_i(k) - \frac{z_{\alpha/2} \cdot \hat{\sigma}_i(k)}{\sqrt{N_k}}, \hat{\tau}_i(k) + \frac{z_{\alpha/2} \cdot \hat{\sigma}_i(k)}{\sqrt{N_k}}\right] \tag{4.5}$$

4.3 Meta-Algorithm for Conservative Extrapolation

In this section, we more formally introduce our algorithm (Algorithm 1). There are two primary steps: falsification of estimators, and combination of confidence intervals. First, we attempt to falsify candidate estimators via hypothesis testing, rejecting

Algorithm 1 Extrapolated Pessimistic Confidence Sets

```
Input: Desired coverage $1-\alpha$. For each $i \in \mathcal{I}_R$, RCT estimate $\hat{\tau}_i(0)$ and variance $\hat{\sigma}_i^2(0)$. For each $i \in \mathcal{I}_R \cup \mathcal{I}_O$, $K$ candidate estimators $\hat{\tau}_i(k)$ and variances $\hat{\sigma}_i^2(k)$. Sample sizes $N_0, \dots, N_K$.
Initialize: Empty candidate set $\hat{\mathcal{C}} \leftarrow \varnothing$
for $k = 1$ to $K$ do
    Compute $\hat{T}_N(k,i), \forall i \in \mathcal{I}_R$, with $\mu_i(k) = 0$ (Eq. 4.3)
    if $\forall i \in \mathcal{I}_R, \left|\hat{T}_N(k,i)\right| \leq z_{\alpha/4|\mathcal{I}_R|}$, then $\hat{\mathcal{C}} \leftarrow \hat{\mathcal{C}} \cup \{k\}$
end for
for $i \in \mathcal{I}_O$ do
    $\hat{L}_i \leftarrow \min_{k \in \hat{\mathcal{C}}} \hat{L}_i(k)(\alpha/2)$ and $\hat{U}_i \leftarrow \max_{k \in \hat{\mathcal{C}}} \hat{U}_i(k)(\alpha/2)$ (Eq. 4.5)
end for
Return: $\hat{L}_i, \hat{U}_i$ for each $i \in \mathcal{I}_O$.
```

estimator k whenever we are able to reject the null hypothesis $H_0 : \tau_i(k) = \tau_i, \forall i \in \mathcal{I}_R$. We use Bonferroni correction to control the false positive rate of the test. For the combination of confidence intervals, while we are unlikely to reject the "correct" estimator if one exists (Assumption 4.3), we may be unable to reject all "incorrect" (i.e., biased) estimators. This motivates the combination of confidence intervals (for the extrapolated effects) of the accepted estimators by taking the maximum and minimum bounds over all such intervals. Our main result characterizes the properties of our procedure, with proof in Appendix B.4.

Theorem 4.1 (Properties of Algorithm 1)**.** *Under Assumptions 4.1 and 4.2, the output of Algorithm 1 has the following asymptotic properties as $N \to \infty$, where N denotes the total sample size, and the samples used for all estimators are of the same order $N_k = \rho_k N_0, \forall k \geq 1$, for some $\rho_k > 0$.*

1. *Under Assumptions 4.3 and 4.4, for each $i \in \mathcal{I}_O$,*

$$\lim_{N \to \infty} \mathbb{P}(\tau_i \in [\hat{L}_i, \hat{U}_i]) \geq 1 - \alpha \tag{4.6}$$

2. *Under Assumption 4.4, for each estimator where $\tau_i(k) \neq \tau_i$ for some $i \in \mathcal{I}_R$,*

$$\lim_{N \to \infty} \mathbb{P}(k \in \hat{\mathcal{C}}) = 0 \tag{4.7}$$

The first point says that for each extrapolated effect τ_i, the coverage of the final confidence interval $[\hat{L}_i, \hat{U}_i]$ is at least $1-\alpha$ in the limit. It follows from Assumption 4.3 and 4.4 that at least one estimator provides intervals $[\hat{L}_i(k)(\alpha/2), \hat{U}_i(k)(\alpha/2)]$ that achieve asymptotic coverage of $1-\alpha/2$. The result follows from our choice of threshold for the significance test as well as application of union bounds. The second point says that we will reject estimators that are not consistent for the validation effects, in the limit. Assumption 4.4 ensures that Proposition 4.1 holds for all estimators, so that this rejection is a consequence of the asymptotic power in Equation (4.4), going to 1 for a fixed bias as $N \to \infty$.

Remark 3. Equations (4.4) and (4.5) are useful for building further intuition. All of the candidate confidence intervals shrink at a rate of $O(1/\sqrt{N})$ as the overall sample size increases. For sufficiently large N, the width of our generated intervals will depend largely on our power to reject biased estimators, which will be higher for observational estimates with larger biases for validation effects.

4.4 Semi-Synthetic Experiments

4.4.1 Setup of Simulation

We generate semi-synthetic RCTs and observational datasets with covariates from the Infant Health and Development Program (IHDP), a randomized experiment on premature infants assessing the effect of home visits from a trained provider on the future cognitive performance (Brooks-Gunn et al., 1992). The outcomes are simulated. Our data generation is based on the partial IHDP dataset used in (Hill, 2011), which includes $n_0 = 985$ observation, 28 covariates, and a binary treatment variable. We construct a scenario where there are four subgroups, defined by the infant's birth weight and maternal marital status: (high [$\geq$ 2000g], married), (low [$<$ 2000g], married), (high, single) and (low, single), which we shorthand as HM, LM, HS and LS. We include all subgroups in the observational studies, but exclude the latter two subgroups for the simulated RCT (i.e. only infants with married mothers are in the RCT).

For each simulated dataset, we generate 1 RCT and K observational studies. For the observational studies, we resample the rows of the IHDP dataset to the desired sample size $n = r \cdot n_0$. We performed weighted sampling to induce a different covariate distribution for observational studies, such that male infants, infants whose mothers smoked, and infants whose mothers worked during pregnancy are less prevalent. Then, we introduce confounding in the observational data, generating m_c continuous confounders and m_b binary confounders. Finally, we simulate outcomes in each dataset, modifying the response surface given in Hill (2011). In our experiments, we may choose to conceal some confounders in each observational study to mimic unobserved confounding, denoting the number of concealed variables across the K studies as $\mathbf{c_z} = (c_{z1}, c_{z2}, ..., c_{zK})$. For further details on confounder generation, outcome simulation, and confounder concealment, see Appendix B.6. Data generation parameters include K, r, m_c, m_b, $\mathbf{c_z}$, and the significance level α. By default, we set $K = 5$, $r = 10$, $m_c = 4$, $m_b = 3$, $\mathbf{c_z} = (0, 0, 2, 4, 6)$, and $\alpha = 0.05$. The full hyperparameter search is provided in Appendix B.6, and details of hyperparameter tuning can be found in Appendix B.3.

4.4.2 Implementation and Evaluation of Meta-Algorithm

To implement Algorithm 1, we first obtain GATE estimates for the four subgroups and their estimated variances in each observational study, combining techniques from the DML and trasportability literature (Semenova and Chernozhukov, 2021; Dahabreh et al., 2020). Estimation details are shown in Appendices B.2 and B.3. For the RCT, we stratify the data into the subgroups HM, LM and estimate the GATEs as the difference of mean outcomes between the treated and untreated. The z tests in Algorithm 1 are applied to both GATE estimates in the HM and LM subgroups ($|\mathcal{I}_R| = 2$), and the significance level of the tests is set at $\alpha/4$.

We evaluate performance using two main metrics: (1) the coverage probability of the output confidence intervals (ideally at least $1-\alpha$), and (2) the width of the confidence intervals (narrower is better). In addition to assessing the intervals produced by Algorithm 1, which we call *Extrapolated Pessimistic Confidence Sets (ExPCS)*, we will evaluate intervals produced by a variant of our algorithm, called *Extrapolated Optimistic Confidence Sets (ExOCS)*. In *ExOCS*, after falsifying estimators, we combine confidence intervals using a random-effects meta-analysis on the non-falsified observational studies. We compare *ExPCS* and *ExOCS* against two baselines. *Meta-Analysis* is a random-effects meta-analysis on all observational studies, as described in Section 4.6, with heterogeneity variance estimated via the DerSimonian-Laird moment method (DerSimonian and Laird, 1986). This baseline is the current standard for aggregating observational study results. The second baseline, *Simple Union*, uses the maximum upper bound and minimum lower bound of the $1-\alpha$ confidence intervals across all observational studies, with no falsification procedure.[7]

4.4.3 Results

We perform three semi-synthetic experiments to assess the performance of our proposed meta-algorithm under different scenarios. The first experiment applies our algorithm under the default settings given in Section 4.4.1. In the second experiment, we vary the sample size ratio between the observational studies and the original RCT, r, from 1 to 10. In the third experiment, we vary the proportion of biased observational studies by setting $\mathbf{c_z}$ to be $(0,0,0,0,0)$, $(0,0,0,0,3)$, $(0,0,0,3,3)$ or $(0,3,3,3,3)$, corresponding to $0,1,2,4$ studies being biased out of a total of 5 observational studies. Results for the latter two experiments are shown over 100 simulations of the datasets. Results for all experiments are shown in Figures 4-2, 4-3, and Figure B-2 in Appendix B.7, respectively. We observe the following:

Meta-algorithm produces confidence intervals that cover the true GATE with nominal probability: We demonstrate in Figure 4-2 the application of our meta-algorithm

[7] Note that *Simple Union* combines $1-\alpha$ confidence intervals, while our approach combines $1-\alpha/2$ confidence intervals to account for the probability of rejecting the "correct" estimator, if one exists. As a result, *Simple Union* intervals do not always strictly cover the intervals produced by *ExPCS*.

(*ExPCS*), a variant of it (*ExOCS*), and two other baselines on one dataset. Our goal is to produce narrow confidence intervals that still cover the true GATEs in the extrapolated subgroups. The confidence intervals of *ExPCS* cover the true GATEs in the extrapolated subgroups with reasonable widths. In contrast, intervals produced by *Meta-Analysis* fail to cover the true GATE in both extrapolated subgroups due to the false assumption of unbiasedness across all studies. The *ExOCS* approach produces narrow intervals for the extrapolated effects, though it barely covers the true effect in the HS subgroup. This hints at the need for a conservative combination of non-falsified studies. However, an overly conservative approach (e.g. *Simple Union*) produces wide intervals that may be of little use for meaningful inference.

Although *Meta-Analysis* produces confidence intervals with inadequate coverage, its intervals for the married subgroups still have considerable overlap with the intervals produced by the RCT. This suggests that testing the meta-analyzed GATE estimates against the RCT GATE estimates may not be enough to demonstrate their validity. Compared to our *ExPCS* intervals, the lower bounds of the *Simple Union* intervals are higher in several subgroups, since we use a higher confidence level for the candidate intervals corresponding to each study to account for probable error in study falsification.

An analysis of increasing observational study size: In Figure 4-3, we find that the coverage of the *Meta-Analysis* intervals is quite low across all sample sizes and particularly decreases at higher sample sizes. This result is intuitive, as three out of five studies are biased, meaning that meta-analysis will converge to a biased estimate as the amount of data increases. One could attempt to fix this issue through *ExOCS*, which does meta-analysis after falsification. However, *ExOCS* has poor coverage when the sample size of the observational studies is small, since the falsification tests are underpowered (evidenced by the high probability of selecting biased studies in Appendix B.7, Table B.1). Both *ExPCS* and the *Simple Union* intervals have adequate coverage across all sample sizes. However, the widths of the intervals reported at the bottom of Figure 4-3 show that *ExPCS* intervals are narrower when there is adequate power, i.e. at higher sample sizes. Ultimately, *ExPCS* will tend to provide intervals

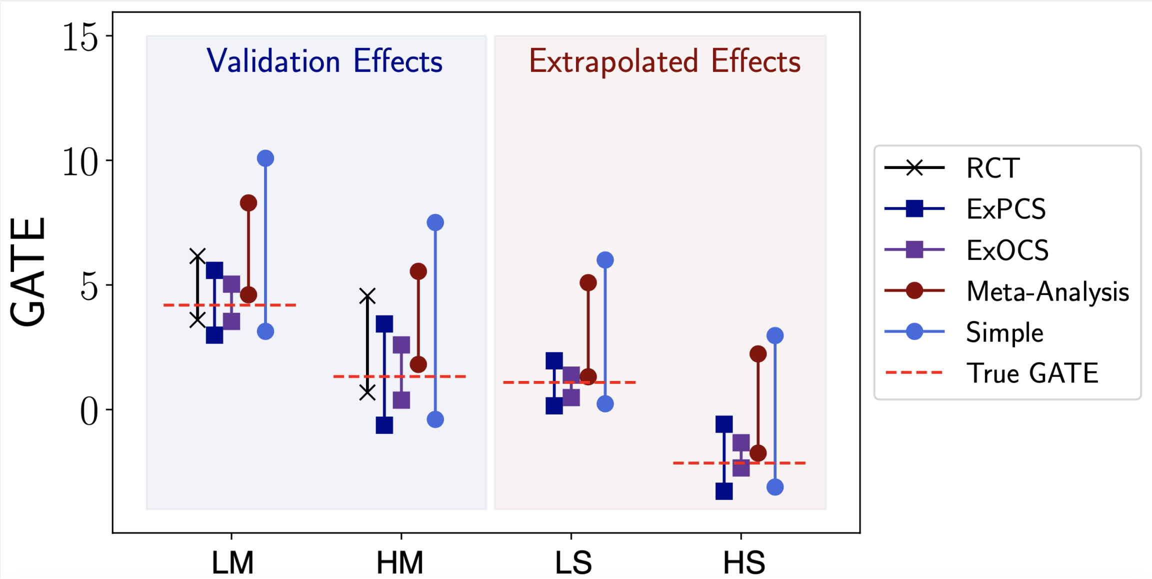
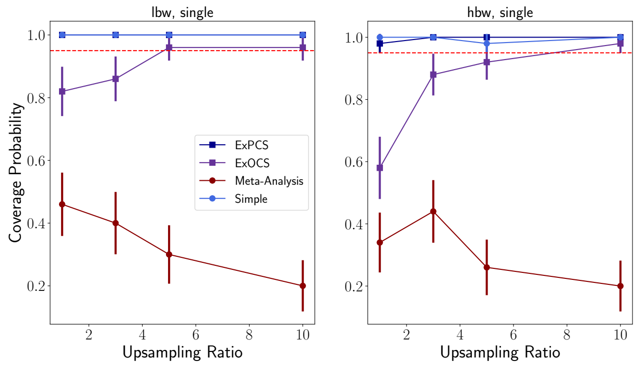


Figure 4-2: *The confidence intervals for group average treatment effects (GATE) within the four subgroups output by our algorithm (ExPCS), our algorithm variant (ExOCS), random-effects meta-analysis on all observational studies (Meta-Analysis), simple union bound on all observational studies (Simple), and RCT, for one dataset generated using the default parameter settings laid out in Section 4.4.1. LM, HM, LS, HS represent four subgroups defined in Section 4.4.1*

that cover the true effect regardless of sample size, and in the case we have sufficient power, these intervals will both have good coverage and narrower width, allowing for more meaningful inference.



Mean width of 95% confidence intervals					
		Upsampling ratio			
Group	Estimator	1	3	5	10
LS	*ExPCS*	21.00	9.51	5.53	3.58
	Simple	21.00	10.20	6.95	5.69
HS	*ExPCS*	25.00	8.29	5.36	4.34
	Simple	26.00	9.15	6.90	6.33

Figure 4-3: *Coverage probabilities of confidence intervals shown as a function of the size of the observational studies relative to the RCT. Dotted red lines stand for 95% coverage. Vertical bars are the 95% confidence intervals of the coverage probabilities. LS / HS stand for groups with low / high birth weight and single mother. Between ExPCS and Simple which have adequate coverage, ExPCS generally has narrower intervals.*

4.5 Women's Health Initiative (WHI) Experiments

In order to assess our approach in a real-world setting, we use clinical trial and observational data available from the WHI. Each subgroup is supported in both RCT and observational data, which proves useful for evaluation. At a high level, we "hide" some number of subgroups from the RCT, estimate a confidence interval of the effect estimate using our algorithm on the remaining data, and compare the result to the hidden RCT estimate. We do this over a large set of possible "held-out" subgroups, yielding >2000 different scenarios on which to test our approach. Because the original observational datasets replicate the RCT results fairly well using standard methods, we create additional "biased" datasets by sub-selecting the original observational dataset in a way that induces selection bias. We evaluate each method, for each held-out subgroup, according to the length of the intervals as well as coverage of the RCT point estimates. Below, we describe the specifics of the data, the experimental setup, and the main results of the analysis. For additional details on data preprocessing, setup, and evaluation, see Appendix B.5.

4.5.1 Setup

The Postmenopausal Hormone Therapy (PHT) trial, i.e. the RCT used in this analysis, was run on postmenopausal women aged 50-79 years who had an intact uterus. It studied the effect of hormone combination therapy on several types of cancers, cardiovascular events, and fractures. The observational study (OS) was run in parallel, had a similar follow-up time to the RCT, and tracked similar outcomes. In our analysis, we use a composite outcome, where $Y = 1$ if any of the following events are **observed** to occur in the first 7 years of follow-up, and $Y = 0$ otherwise: coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. This represents a binarization of the "global index" time-to-event outcome from the original study, where $Y = 0$ could also occur due to censoring. We establish treatment and control groups in the OS based on explicit confirmation or denial of usage of both estrogen and progesterone in the first three years. We use only covariates measured in both the RCT and OS to simplify analysis.

4.5.2 Evaluation

Our empirical evaluation consists of several steps. In the first step, we replicate the principal results from the PHT trial, given in Table 2 of (Rossouw et al., 2002), by fitting a doubly robust estimator (of the style given in Appendix B.3) on the WHI OS data. Then, while treating the WHI OS dataset as the "unbiased" observational dataset, we simulate additional "biased" observational datasets by inducing selection bias into the WHI OS. The exact mechanism of selection bias and its clinical intuition is given in Appendix B.5. Importantly, this is the only part of the evaluation that involves any simulation.

The second step is to construct a large suite of tasks on which to evaluate our method, by considering different sets of validation-extrapolation subgroups. To construct the subgroups, we consider all pairs of a selected set of binary covariates (see

	Coverage	**Length**	**OS %**
Simple	0.39	0.416	–
Meta-Analysis	0.03	0.260	–
ExOCS	0.28	0.058	–
ExPCS (ours)	0.45	0.081	0.99
Oracle	0.44	0.068	–

Table 4.1: *Coverage, length, and unbiased OS % of ExPCS and baselines. ExPCS achieves comparable coverage to the oracle method with highly efficient intervals. Additionally, we do not reject the unbiased OS in* 99% *of the tasks.*

Appendix B.5.6), where each pair defines four subgroups. For example, one covariate pair is ("current smoker", "currently drinks alcohol"). We treat two of the subgroups as validation subgroups and two as extrapolated subgroups. For the latter groups, we apply our algorithm without access to the RCT data, and only use the RCT data for final evaluation. The total number of covariate pairs is 592, leading to 1184 distinct "tasks" (i.e., extrapolated groups). For each task, we evaluate ExPCS (our method), ExOCS, Simple, and Meta-Analysis (described in Section 4.4.2). Additionally, we evaluate an "oracle" method, which is identical to ExPCS, except that it always selects only the original observational study (i.e. the base WHI OS to which we have not added any selection bias). For each method, we compute the following metrics, averaged across all tasks – **Length**: length of the confidence interval, **Coverage**: percentage of tasks where the interval covers the RCT point estimate. In addition, we report the **Unbiased OS Percentage**: the percentage of tasks where the ExPCS approach retains the unbiased study after the falsification step.

4.5.3 Results

Table 4.1 reports the metrics above, averaged across all extrapolated subgroups.

Compared to the "simple" baseline, our approach has better coverage with much shorter confidence intervals. Our falsification procedure retains the unbiased observational study 99% of the time, yielding near-oracle coverage rates, but produces substantially shorter intervals than the “simple” baseline. Recall that the simple baseline takes a union over all $1-\alpha$ intervals estimated from each observational dataset, while ExPCS takes a union of a smaller number of slightly wider $(1-\alpha/2)$ confidence intervals.

Compared to the Meta-Analysis and ExOCS baselines, we achieve comparable (or much better) length with substantially better coverage. In particular, compared to meta-analysis, we achieve tighter intervals and also cover the RCT estimate with higher frequency. This result is intuitive, since one will get a biased estimate if biased observational studies are included in the meta-analysis. Additionally, conservatively combining the non-falsified estimates (as opposed to *ExOCS*, which does a meta-analysis on the non-falsified estimates) is important to achieve good coverage (0.45 vs 0.28).

We get comparable coverage and interval lengths to the oracle method. Our coverage rate is nearly identical (0.45) to that of the oracle method (0.44), with intervals that are marginally wider (0.081 vs. 0.068). Our slightly improved coverage is possible due to the wider intervals. Note that our measure of “coverage” may be pessimistic, because we track coverage of the RCT point estimate, as opposed to the true causal effect (which is unknown), and the confidence intervals are designed to cover the latter. Indeed, we report the oracle method precisely as a means of providing a more suitable comparison. Overall, our real-world results suggest that our method of falsification followed by a conservative combination of intervals may be useful for biostatisticians and clinicians when doing meta-analyses.

4.6 Related Work

Meta-analysis for combining observational estimates Among the quantitative approaches for meta-analysis to account for potential bias, our *Meta-Analysis* baseline is standard for meta-analysis of observational data (Higgins et al., 2019) to account for heterogeneity. Allowing for heterogeneity of treatment effects among studies produces wider confidence intervals and thus more conservative inference. If additional study-level covariates are available (e.g. study designs, drop-out rate), several approaches aim to adjust for potential bias, either by modeling the bias magnitude (Eddy et al., 1990; Wolpert and Mengersen, 2004; Anglemyer et al., 2014; Greenland, 2005), down-weighting studies with higher risk of bias (Ibrahim and Chen, 2000; Neuenschwander et al., 2009), or using Bayesian hierarchical regression to account for difference between subgroups of studies (Prevost et al., 2000; Welton et al., 2009). Our work differs from these approaches, in that (1) we use information from outside the population of interest to assess bias, and (2) we do not place any assumptions on the patterns of bias across studies.

Partial identification and sensitivity analysis These methods seek to place bounds on causal effects when they cannot be point-identified. Our method can be seen as an alternative way of doing so, with a fundamentally different type of assumption. Methods for partial identification rely on having discrete variables and a known causal graph (typically including unobserved confounders) (Duarte et al., 2021, Section 9). Methods for sensitivity analysis, on the other hand, translate assumptions about the strength and nature of unobserved confounding into bounds on causal effects (Rosenbaum and Rubin, 1983a; Rosenbaum et al., 2010; Yadlowsky et al., 2018). In contrast, we do not make any such assumptions, e.g., we allow for continuous variables, and when some candidate estimators are biased due to unmeasured confounding, we do not place any limit a-priori on the bias. An extended related work is given in Appendix B.8.

4.7 Discussion and Limitations

We have presented a meta-algorithm that constructs conservative confidence intervals for group average treatment effects of subgroups that are not represented in RCTs, but are represented in observational studies. Under the assumption that there exists at least one candidate estimator that is asymptotically normal and consistent for both the validation and extrapolated effects, these intervals will achieve the correct asymptotic coverage of the true effect. However, our method is not without limitations. Most notably, we may fail to reject the null hypothesis due to low power, e.g., when an observational estimate $\hat{\tau}(k)$ has high variance. In practice, we expect that our approach will be most useful when the observational studies in question have large sample sizes, leading to higher-precision estimates of potential bias, and smaller confidence intervals on the extrapolated effects. Our hope is that methods such as ours will lead to higher confidence in observational estimates when RCT data is available to falsify observational studies that do not replicate known causal relationships. Finally, great care should be taken to appropriately validate and soundly interpret the results of our method in practice, especially with more sensitive subgroups (e.g. with respect to race or gender).

Part II

Robust prediction via causal knowledge

Chapter 5

Regularizing towards Causal Invariance: Linear Models with Proxies

This chapter (and accompanying appendix) was previously published as (Oberst et al., 2021a) at ICML 2021.

5.1 Introduction

Ideally, predictive models would generalize beyond the distribution on which they are trained, e.g., across geographic regions, across time, or across individual users. However, models often learn to rely on signals in the training distribution that are not stable across domains, causing a drop-off in predictive performance. This problem is broadly known as dataset shift (Quiñonero-Candela et al., 2009).

Tackling this problem requires a formalization of how dataset shift arises, and how that shift impacts the conditional distribution of our target Y given features X. One way to formalize this shift is in terms of an underlying causal graph (Pearl, 2009), where changes between distributions are seen as arising from causal interventions on variables.

Conceptual example: In the causal graph given in Figure 5-1, the variable A serves as

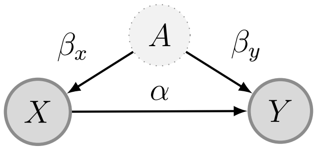


Figure 5-1: *Conceptual Example: A represents an (unobserved) socioeconomic variable, X represents current health status, and Y represents a long-term health outcome. All relationships are assumed to be linear, and coefficients are given. We consider a broader class of graphs in this work, see Figure 5-2.*

a confounder. In a medical setting, A could represent smoking habits or socioeconomic status, which have a causal effect on current health status (X) as well as longer-term outcomes (Y). Importantly, A may not be recorded in our training data, and the distribution of A could vary across geography and time.

In the context of this causal graph, interventions which change the distribution of A will also alter the conditional mean $\mathbb{E}(Y \mid X)$. Under the linear relationships in Figure 5-1, the optimal least-squares predictor $\hat{Y} = \gamma^* X$ under the test distribution depends on the test-time variance in A, in that

$$\gamma^* = \begin{cases} \alpha, & \text{if after intervention } A = 0 \\ \alpha + \frac{\beta_Y}{\beta_X}, & \text{if after intervention } \mathrm{Var}(A) \to \infty. \end{cases}$$

The first predictor encodes the direct causal effect of X on Y, but is only optimal in the setting where the correlations induced by A are removed by fixing it to a constant value of zero (the same holds when including intercepts and allowing for non-zero means). The second predictor, on the other hand, renders the distribution of the residual $Y - \hat{Y}$ independent of A, and is therefore robust to arbitrary interventions upon A. However, this is only optimal under arbitrarily strong interventions on A.

Balancing performance and invariance: Instead of seeking an invariant predictor that is robust to arbitrary interventions on A (like the second predictor above), we instead seek to minimize a worst-case loss under bounded interventions of a given strength. We contrast this with work that seeks to discover causal relationships as a route to invariance (Rojas-Carulla et al., 2018; Magliacane et al., 2018), optimize for invariance directly across environments (Arjovsky et al., 2019), or use known causal structure to select predictors with invariant performance (Subbaswamy et al., 2019).

Our proposed objective takes the form of a standard loss, plus a regularization term that encourages invariance. This builds upon Rothenhäusler et al. (2021), who introduce a similar objective, and prove that their objective optimizes a worst-case loss over bounded interventions on A, under a large class of linear structural causal models.

In contrast to Rothenhäusler et al. (2021), we do not assume that A is observed. Instead we assume that, during training, we have access to noisy proxies of A. For most of this chapter, we assume that neither A nor proxies are available during testing. With this in mind, our contributions are as follows

- *Distributional robustness to bounded shifts*: In Section 5.3, we show that a single proxy can be used to construct estimators with distributional robustness guarantees under bounded interventions on A. However, these estimators are robust to a strictly smaller set of interventions, compared to when A is used directly, and the size of this set depends on the (unidentifiable) noise in the proxy. When two proxies are available, we propose a modified estimator that can be used to recover the same guarantees as when A is observed.
- *Targeted shifts*: In Section 5.4, we show how to target our loss to interventions on A contained in a specified robustness set. We show that this formulation includes Anchor Regression as a special case, but also allows for sets that are not centered around the mean of A. In this setting we give an estimator, using two proxies, that identifies the target loss.

In Section 5.5, we evaluate our theoretical findings on synthetic experiments, and in Section 5.6 we demonstrate our method on a real-world dataset consisting of hourly pollution readings across five major cities in China.

5.2 Preliminaries

5.2.1 Notation

We use upper case letters X to denote (possibly vector-valued) random variables, and lower-case letters x to denote values in the range of those random variables. Vectors are assumed to be column vectors, so that $X \in \mathbb{R}^{d_X}$ indicates that $X = (X_1, \ldots, X_{d_X})^\top$, a column vector of d_X random variables. We use $\Sigma_X \in \mathbb{R}^{d_X \times d_X}$ to denote the covariance matrix of a variable X. We use bold upper-case letters $\mathbf{X}$ to denote a data matrix in $\mathbb{R}^{n \times d_X}$, consisting of n i.i.d. observations of X, and $\mathbf{1}\{\cdot\}$ as an indicator random variable. When dealing with matrices C, D, we use $C \prec D$ and $C \preceq D$ to indicate the positive definite and positive semi-definite partial order, respectively. That is, $C \prec D$ if $D - C$ is positive definite (PD), and $C \preceq D$ if $D - C$ is positive semi-definite (PSD). We use Id to denote the identity matrix, whose dimension is given by context. All proofs are provided in the supplementary material.

5.2.2 Linear structural causal model

We assume the general class of causal graphs represented in Figure 5-2, where $X \in \mathbb{R}^{d_X}$ denotes observed covariates that can be used in prediction, $Y \in \mathbb{R}^{d_Y}$ is the target we seek to predict, $H \in \mathbb{R}^{d_H}$ are unobserved variables, and $A \in \mathbb{R}^{d_A}$ represents anchor variables, which are assumed to have no causal parents in the graph. We assume the linear structural causal model (SCM) given in Assumption 5.1.

Assumption 5.1 (Linear SCM)**.** We assume the SCM

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + M_A A + \epsilon, \tag{5.1}$$

where A, ϵ have zero mean, bounded covariance, and are independently distributed. We assume that $\mathbb{E}[AA^\top]$ and $\mathrm{Id} - B$ are invertible, where Id is the identity matrix.

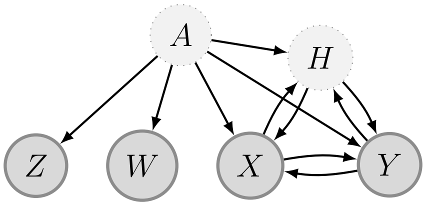


Figure 5-2: *In contrast to Rothenhäusler et al. (2021), we assume that anchor variables (denoted A) are unobserved, but that we have access to either one or two proxies W, Z. Observed variables are shown in dark grey and unobserved variables in light grey. We do not assume knowledge of the causal structure between A, X, H, Y (except that A has no causal parents). The relationship between X, H, Y could be cyclic, but all relationships are linear.*

See Figure 5-2 for a graphical representation.

Note that we do not assume here (or anywhere in this chapter) that either A or ϵ is Gaussian. The invertibility of $\mathrm{Id} - B$ is satisfied if the causal graph is a directed acyclic graph. The matrices B, M_A encode the linear causal relationships. For instance, Figure 5-1 can be represented in this form by $B = \begin{bmatrix} 0 & 0 \\ \alpha & 0 \end{bmatrix}$, $M = \begin{bmatrix} \beta_X \\ \beta_Y \end{bmatrix}$. In general, $\epsilon \in \mathbb{R}^D$, $B \in \mathbb{R}^{D \times D}$, and $M \in \mathbb{R}^{D \times d_A}$, where $D := d_X + d_Y + d_H$. We assume that $d_Y = 1$ for simplicity.

5.2.3 Distributional robustness of anchor regression

Our goal is to learn a predictor $f^*(X)$ of Y that minimizes a worst-case risk of the following form

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\ell(Y, f(X))], \tag{5.2}$$

where $\mathcal{F}$ denotes a hypothesis class of possible predictors, $\mathcal{P}$ denotes a set of possible distributions, and ℓ represents our loss function. We take the class $\mathcal{P}$ to consist of distributions that arise as the result of causal interventions on A, and seek to learn a linear predictor to minimize mean-squared error.

We use $\mathbb{P}$ to refer to the observational distribution, and $\mathbb{P}_{do(A:=\nu)}$ to refer to the distribution under interventions on A, where the variable A is replaced by the random variable ν, and ν is assumed to be independent of the noise vector ϵ. We often write

$$R(\gamma) \coloneqq Y - \gamma^\top X$$

as a random variable that represents the residual of a predictor $\gamma \in \mathbb{R}^{d_X}$. Importantly, Assumption 5.1 implies that for any γ, $\mathbb{E}[R(\gamma) \mid A]$ can be written as a linear function in A.

In this setting, Rothenhäusler et al. (2021) propose the following objective, defined here with respect to the observational distribution $\mathbb{P}$ (rather than a finite sample)

Definition 5.1 (Anchor Regression)**.**

$$\ell_{AR}(A; \gamma, \lambda) \coloneqq \ell_{LS}(X, Y; \gamma) + \lambda \ell_{PLS}(X, Y, A; \gamma), \tag{5.3}$$

where $\lambda \geq -1$ is a hyperparameter and

$$\ell_{LS}(X, Y; \gamma) \coloneqq \mathbb{E}\left[R(\gamma)^2\right] \tag{5.4}$$

$$\ell_{PLS}(X, Y, A; \gamma) \coloneqq \mathbb{E}\left[(\mathbb{E}\left[R(\gamma) \mid A\right])^2\right]. \tag{5.5}$$

The first term ℓ_{LS} encodes the least-squares objective, while the second term ℓ_{PLS} encodes the residual error which can be predicted from A, which we refer to as the projected least-squares error. For $\lambda > 0$, the second term adds an additional penalty (beyond that of ordinary least squares) when the bias varies across values of A. The second term (5.5) can also be written in the linear setting of Assumption 5.1 as

$$\ell_{PLS}(A; \gamma) = \mathbb{E}[R(\gamma) A^\top] \mathbb{E}[A A^\top]^{-1} \mathbb{E}[A R(\gamma)^\top], \tag{5.6}$$

where we drop the dependence on X, Y for notational simplicity. Under Assumption 5.1, Equation (5.3) corresponds to a worst-case loss under distributional shift caused by bounded intervention on A (Rothenhäusler et al., 2021, Theorem 1)

$$\ell_{AR}(A;\gamma,\lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y-\gamma^\top X)^2], \tag{5.7}$$

where the robustness set is given by

$$C_A(\lambda) \coloneqq \{\nu : \mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]\}. \tag{5.8}$$

Since minimizing ℓ_{AR} is equivalent to ordinary least squares (OLS) regression when $\lambda = 0$, this also provides a natural robustness guarantee for the OLS estimator, where $C_{OLS} \coloneqq \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top]\}$. In an identifiable instrumental variable setting, the minimizer converges against the causal parameter for $\lambda \to \infty$ (e.g. Jakobsen and Peters, 2020, eq. (71)); the ℓ_{PLS} term has therefore been referred to as 'causal regularization' (e.g. Bühlmann and Ćevid, 2020), and has also been denoted by ℓ_{IV} (Rothenhäusler et al., 2021), as $\mathrm{Cov}(A, R(\gamma)) = \mathbf{0}$ if and only if $\ell_{PLS}(\gamma) = 0$.

5.3 Distributional robustness to bounded shifts

We first assume the existence of a noisy proxy W, conditionally independent of (X, Y, H) given A (see Figure 5-2).

Assumption 5.2 (Single proxy with additive noise)**.** In the context of Assumption 5.1, W is generated as follows

$$W \coloneqq \beta_W^\top A + \epsilon_W,$$

where ϵ_W has mean zero, bounded covariance, and is independent of (A, ϵ). In addition, we assume that the second moment matrix $\mathbb{E}[WW^\top]$ is invertible.

Under mild identifiability conditions (e.g., that β_W is full rank) one can show (see Section C.3.2) that

$$\ell_{PLS}(A;\gamma) = 0 \iff \ell_{PLS}(W;\gamma) = 0, \tag{5.9}$$

Hence, a single proxy is enough (in the population case) to identify whether the sharp constraint $\ell_{PLS}(\gamma) = 0$ holds, representing invariance to interventions of arbitrary strength. This corresponds to the fact that if A is a valid instrumental variable, then so is W (Hernán and Robins, 2006).

However, we consider interventions on A that are not of arbitrarily large strength. With that in mind, in Section 5.3.1, we demonstrate that (i) when a single proxy W is used in place of A, a robustness guarantee holds, but the robustness set is reduced relative to (5.8), (ii) the extent of this reduction depends on the signal-to-variance relationship in W, and (iii) this relationship is not generally identifiable from the observational distribution over (X, Y, W) alone. In Section 5.3.2, we show that in the setting where two proxies are available, the same guarantees as for an observed A can be obtained. We do so constructively, giving a regularization term whose population version is equal to $\ell_{PLS}(A; \gamma)$.

5.3.1 Robustness with a single proxy

First, we establish the robustness set of Anchor Regression when a single proxy is used in place of A. We refer to this as Proxy Anchor Regression, to distinguish it from the case when A is observed, but the only difference from Definition 5.1 is that W is used in place of A.

Definition 5.2 (Proxy Anchor Regression)**.** Let ℓ_{LS}, ℓ_{PLS} be defined as in Equations (5.4) and (5.6). We define

$$\ell_{PAR}(W; \gamma, \lambda) \coloneqq \ell_{LS}(\gamma) + \lambda \ell_{PLS}(W; \gamma), \tag{5.10}$$

where $\lambda \geq -1$ is a hyperparameter and we suppress the dependence on X, Y in the notation.

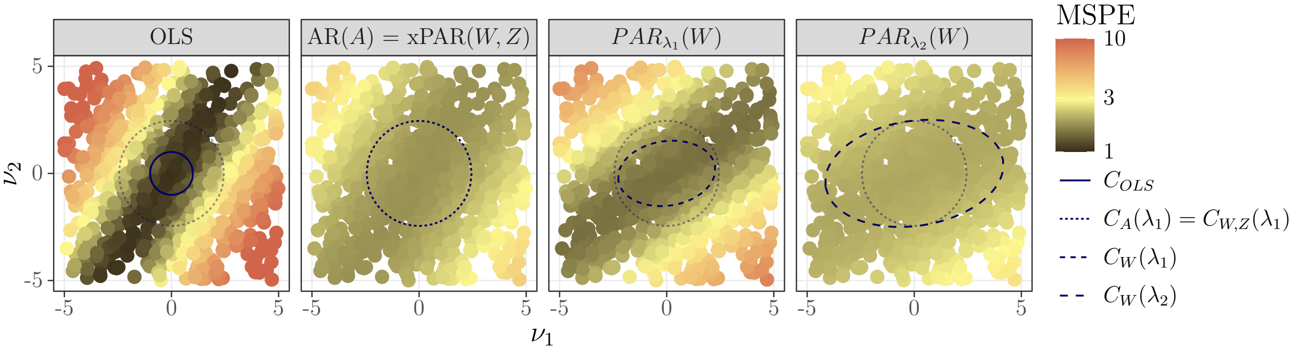


Figure 5-3: *Test performance under interventions $do(A \coloneqq (\nu_1, \nu_2))$ which give rise to different test distributions over X and Y. Each dot corresponds to a different intervention (i.e., test distribution on X, Y), and the color gives the resulting mean squared prediction error (MSPE).* ***(Far Left)*** *OLS performs well for interventions in the set C_{OLS} (solid circle), corresponding to the training covariance of A. However, it performs poorly under interventions far from this region (e.g., top left).* ***(Middle Left)*** *Anchor Regression (AR) minimizes the worst-case loss over interventions on A within the region $C_A(\lambda_1)$ (cf., (5.8)), a re-scaling of C_{OLS}. There is a trade-off, with better performance than OLS under large interventions, but worse performance under small interventions. Given two proxies W, Z, we introduce Cross-Proxy Anchor Regression (xPAR, cf., (5.14)) and prove that it minimizes the same worst-case loss.* ***(Middle Right)*** *When only a single proxy W is used in place of A, the result is a weaker guarantee, in the form of a smaller robustness set $C_W(\lambda_1)$ (cf., (5.11)) for the same value of λ_1. The shape of this set depends on the noise in the proxy along different dimensions.* ***(Far Right)*** *As a result, there does not generally exist a λ_2 such that $C_W(\lambda_2) = C_A(\lambda_1)$. If we choose some $\lambda_2 > \lambda_1$ such that $C_A(\lambda_1) \subset C_W(\lambda_2)$, we enforce a stronger constraint than intended, resulting in an unwanted trade-off between performance and robustness.*

Theorem 5.1. *Under Assumptions 5.1 and 5.2, for all $\gamma \in \mathbb{R}^{d_X}$ and for all $\lambda \geq -1$*

$$\ell_{PAR}(W; \gamma, \lambda) = \sup_{\nu \in C_W(\lambda)} \mathbb{E}_{do(A \coloneqq \nu)}[(Y - \gamma^\top X)^2],$$

where the robustness set is given by

$$C_W(\lambda) \coloneqq \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega_W\} \tag{5.11}$$

and where Ω_W is defined as

$$\Omega_W := \mathbb{E}[AW^\top]\big(\mathbb{E}[WW^\top]\big)^{-1}\mathbb{E}[WA^\top]. \tag{5.12}$$

Intuitively, Ω_W defines a signal-to-variance relationship in W, and this determines the robustness guarantee. In the case where both $A, W \in \mathbb{R}$ are one-dimensional, and A has unit variance, the robustness sets simplify to

$$\begin{aligned}
C_{OLS} &= \{\nu : \mathbb{E}[\nu^2] \leq 1\} \\
C_W(\lambda) &= \{\nu : \mathbb{E}[\nu^2] \leq 1 + \lambda \cdot \rho_W\} \\
C_A(\lambda) &= \{\nu : \mathbb{E}[\nu^2] \leq 1 + \lambda\},
\end{aligned}$$

where $\rho_W := \beta_W^2/(\beta_W^2 + \mathbb{E}\epsilon_W^2) < 1$ is the signal-to-variance ratio of W, also referred to as the reliability ratio in the measurement error literature (Fuller, 1987). Thus, in the one-dimensional case, the robustness set using W is strictly smaller than the one obtained by using A when $\lambda > 0$, except in the case where $\epsilon_W = 0$ a.s. This result generalizes to higher dimensions.

Proposition 5.1. *Assume Assumptions 5.1 and 5.2 and that $\mathbb{E}[\epsilon_W\epsilon_W^\top] \in \mathbb{R}^{d_W \times d_W}$ is positive definite. Then for $\lambda > 0$*

$$C_{OLS} \subseteq C_W(\lambda) \subset C_A(\lambda),$$

and the set $C_W(\lambda)$ increases monotonically when $\mathbb{E}[\epsilon_W\epsilon_W^\top]$ decreases w.r.t. the partial matrix ordering. If $d_W = d_A$, β_W is full rank, and $\epsilon_W = 0$ a.s., then $C_W(\lambda) = C_A(\lambda)$.

If Ω_W were known, we could choose a larger λ^* such that $C_A(\lambda) \subseteq C_W(\lambda^*)$. In contrast to the one-dimensional case, where we could choose $\lambda^* = \lambda/\rho_W$ to obtain an equality $C_A(\lambda) = C_W(\lambda^*)$, we cannot generally achieve equality in higher dimensions (see Figure 5-3).

However, Ω_W is not generally identifiable from the observed distribution over (X, Y, W)

alone. Moreover, SCMs compatible with the observed distribution react differently under interventions on A and yield different coefficients that are optimal w.r.t. interventions in $C_A(\lambda)$. Consequently, in this setting, it is not possible to recover the guarantees of Anchor Regression without further assumptions (e.g., on Ω_W). See Supplement C.2 for an example.

Note that these results apply regardless of whether or not β_W is full rank. However, if β_W is not full rank, then there will be directions of variation in A that are not reflected in W, and we will not be able to achieve additional robustness (beyond that of OLS) against interventions along these directions.

5.3.2 Robustness with two proxies

We now show that if we have two (sufficiently different) proxies for A, then it is possible to recover the original robustness set using a different regularization term. We denote these proxies by W, Z, as shown in Figure 5-2. In this setting, the structural causal model over (X, Y, H, A) can still be written in the form of Equation (5.1), where we make the following additional assumptions.

Assumption 5.3 (Proxies with additive noise)**.** In the context of Assumption 5.1, Z, W are generated as follows

$$W := \beta_W^\top A + \epsilon_W \qquad \text{and} \qquad Z := \beta_Z^\top A + \epsilon_Z,$$

where ϵ_W, ϵ_Z are mean-zero with bounded covariance, and $\epsilon_W, \epsilon_Z, \epsilon, A$ are jointly independent.

Assumption 5.4. The dimensions of A, W, Z are equal, $d_A = d_W = d_Z$, and β_W, β_Z are full-rank.

Note that Assumption 5.4 also implies that the second moment matrix $\mathbb{E}[ZW^\top]$ is invertible.

To build intuition, note that this assumption is trivially satisfied in the setting where $W = A + \epsilon_W$ and $Z = A + \epsilon_Z$, i.e., where W and Z are two noisy observations of A. More generally, Assumption 5.4 rules out directions of variation in A that are undetectable in W or Z.

In this setting we introduce the following loss, and prove that it is equal to the worst-case loss obtained when A is observed (c.f., (5.7))

Definition 5.3 (Cross-Proxy Anchor Regression)**.**

$$\ell_{\times PAR}(W, Z; \gamma, \lambda) \coloneqq \ell_{LS}(X, Y; \gamma) + \lambda \ell_{\times}(W, Z; \gamma),$$

where we refer to

$$\ell_{\times}(W, Z; \gamma) \coloneqq \mathbb{E}[R(\gamma) W^\top] \mathbb{E}[Z W^\top]^{-1} \mathbb{E}[Z R(\gamma)^\top], \tag{5.13}$$

as the cross-proxy regularization term.

Theorem 5.2. *Under Assumptions 5.1, 5.3 and 5.4, for any $\gamma \in \mathbb{R}^{d_X}$ and any $\lambda \geq -1$*

$$\ell_{\times PAR}(W, Z; \gamma, \lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2], \tag{5.14}$$

where $C_A(\lambda) = \{\nu : \mathbb{E}[\nu\nu^\top] \preceq (1 + \lambda)\mathbb{E}[AA^\top]\}$.

$\ell_{\times PAR}$ is convex in γ and has a closed form solution for its minimizer based only on the population moments of X, Y, W and Z (see Proposition A4 in the supplement).

To build intuition for why Assumption 5.4 is required for this result, consider an example where W, Z are both scalars ($d_W = d_Z = 1$) and A has two independent dimensions (A_1, A_2). In this example, if both proxies measure the same dimension A_1, then variation in A_2 is not detectable in either proxy, and we cannot optimize for robustness to interventions on A_2. On the other hand, if W only measures A_1 (e.g., $W = A_1 + \epsilon_W$), and Z only measures A_2 (e.g., $Z = A_2 + \epsilon_Z$), then we cannot use Z to identify the signal-to-variance ratio of W, and vice-versa. In this case, (W, Z) is effectively a single two-dimensional proxy in the framework of Section 5.3.1, where we showed that recovering the guarantees of Anchor Regression is not generally possible. Intuitively, we need all directions of variation in A to have some influence on both proxies (i.e., β_W, β_Z full rank), and hence require that W, Z have sufficiently large dimension.

5.4 Targeted anchor regression: Incorporating additional shift information

We now generalize Anchor Regression to an estimator that is targeted to be robust against particular shifts, and demonstrate that we can similarly handle this setting when only proxies of A are observed. In Section 5.2.3 we showed that Anchor Regression minimizes the worst-case loss over the set $C_A(\lambda)$ of all interventions $do(A \coloneqq \nu)$ where $\mathbb{E}[\nu\nu^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]$. For deterministic ν, $C_A(\lambda)$ is an ellipsoid centered at 0, and its width in each direction is proportional to the variation of A in that direction. However, we may desire a different robustness set: For instance, if we anticipate a particular shift μ_ν in the mean of A, or if we want to add extra protection against particular directions of variation in A. This can be formalized as a robustness set defined by an ellipsoid that may not be centered at 0, nor be proportional to $\mathbb{E}[AA^\top]$. The estimator developed in this section can incorporate such prior beliefs.

More formally, instead of considering robustness against interventions $do(A \coloneqq \nu)$ over the set $\nu \in C_A(\lambda)$, we now assume that we have additional information on the nature of ν, which is specified in the form of a vector μ_ν and a symmetric PSD matrix Σ_ν. We introduce a new method, Targeted Anchor Regression, minimizing what we refer to as the *targeted loss*. We prove in Propositions 5.2 and 5.3 that minimizing this objective can be interpreted in two ways: First, as minimizing an expected loss over interventions ν with a known mean and covariance, or minimizing a worst-case loss over deterministic interventions ν contained in an ellipsoid robustness set (as discussed above). This is visualized in Figure 5-4.

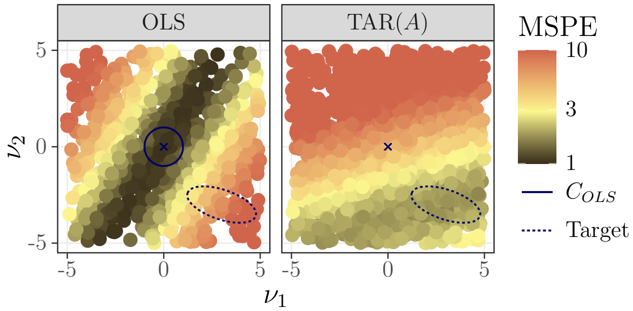


Figure 5-4: *Targeted Anchor Regression allows for minimizing the worst-case loss in regions (dashed ellipse) that may differ in location, size, and shape from the regions in Figure 5-3 (OLS copied for reference). Every point ν represents a test distribution $do(A := \nu)$, the color indicating the mean squared prediction error in this distribution. Cross marks the origin. The TAR estimator achieves its minimal test loss at the center of the targeted region.*

5.4.1 Targeting when A is observed

We first consider the case when A is observed during training, and the mean and covariance of ν are known, given by μ_ν, Σ_ν. Importantly, for a given γ we have $\mathbb{E}[R(\gamma) \mid A = a] = b_\gamma^\top a$, where, writing $\Sigma_A := \mathbb{E}[AA^\top]$,

$$b_\gamma^\top := \mathbb{E}[R(\gamma)A^\top]\Sigma_A^{-1}. \tag{5.15}$$

Definition 5.4 (Targeted Anchor Regression)**.** Let $\mu_\nu \in \mathbb{R}^{d_A}$, and $\Sigma_\nu \in \mathbb{R}^{d_A \times d_A}$, where Σ_ν is a symmetric PSD matrix.

$$\begin{aligned} &\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha) \\ &:= \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma + \left(b_\gamma^\top \mu_\nu - \alpha\right)^2, \end{aligned} \tag{5.16}$$

where b_γ is defined in (5.15), and Σ_A is the covariance of A.

Proposition 5.2. *Under Assumption 5.1, and the assumption that $\nu \perp\!\!\!\perp \epsilon$, we have, for all $\gamma \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}$,*

$$\ell_{TAR}(A; \mu_\nu, \Sigma_\nu; \gamma, \alpha) = \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2],$$

where $\mu_\nu = \mathbb{E}[\nu]$ and Σ_ν is the covariance matrix of ν.

Importantly, the objective in Equation (5.16) is convex in (γ, α), and has a closed-form solution (see Proposition A5 in the supplement). If ν is a known constant, then this corresponds to performing OLS using both X and A as predictors during training, and using the known value of ν for A for prediction (see Supplement C.3.3). However, if for example ν exhibits more variance than A along certain directions, and less variance along others, then the targeted regression parameter differs from standard solutions. Optimizing the objective in Equation (5.16) can also be interpreted as optimizing a worst-case loss over interventions $do(A := \nu)$ in a certain set.

Proposition 5.3. *Under Assumption 5.1, we have, for all $\mu_\nu \in \mathbb{R}^{d_A}$ and $\Sigma_\nu \in \mathbb{R}^{d_A \times d_A}$ being a symmetric positive definite matrix, that*

$$\begin{aligned}&\arg\min_{\gamma,\alpha} \ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha)\\ &\qquad = \arg\min_{\gamma,\alpha} \sup_{\nu \in T(\mu_\mu, \Sigma_\nu)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2],\end{aligned}$$

where the supremum is taken over (deterministic or random) shifts ν of the form $\nu = \mu_v + \delta$, where δ satisfies the constraint that $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$. If δ is random, we require that it is independent of all other random variables. In other words, we can write that ν lies in the set

$$T(\mu_\nu, \Sigma_\nu) := \{\nu : \mathbb{E}[(\nu - \mu_\nu)(\nu - \mu_\nu)^\top] \preceq \Sigma_\nu\}.$$

Note that the expectation in the constraint T is with respect to the random variable ν. This covers the case in which ν (and hence δ) is deterministic, in which case it is equal to a fixed value with probability one.

Proposition 5.3 shows that Targeted Anchor Regression generalizes Anchor Regression to a broader class of robustness sets, that need not depend explicitly on $\mathbb{E}[AA^\top]$. In particular, Anchor Regression can be viewed as a special case, where $\Sigma_\nu = (1 + \lambda)\Sigma_A$ and $\mathbb{E}[\nu] = 0$, in which case the objectives are equal for $\alpha = 0$. In the following, we adopt the interpretation of μ_ν, Σ_ν as specifying a mean and covariance of ν (Proposition 5.2).

5.4.2 Targeting with proxies

In the single-proxy setting, we define Proxy Targeted Anchor Regression as using W in place of A in Equation (5.16). We assume a known mean and covariance of W under $\mathbb{P}_{do(A:=\nu)}$, used in place of μ_ν, Σ_ν. By similar arguments to those in Section 5.3.1, this approach does not generally yield the optimal predictor, in a way that depends on the (unidentified) signal-to-variance relationship in W. Given the similarity, we defer details to Supplement C.4.

When two proxies W, Z are available, we can recover the statement from Proposition 5.2 using a modified estimator, by similar arguments to those in Section 5.3.2. The core observation is that we can construct a linear term

$$a_\gamma^\top := \mathbb{E}[R(\gamma) Z^\top](\mathbb{E}[W Z^\top])^{-1}, \tag{5.17}$$

which, if $\beta_Z = \beta_W = \text{Id}$ can be seen as a linear IV estimate of $b_\gamma^\top$ in Equation (5.15), an estimator used in the measurement error literature given repeated noisy measurements of a single variable (Fuller, 1987). In our case, Equation (5.17) identifies $b_\gamma^\top$ only up to the linear transformation β_W, but this is sufficient to identify the targeted loss.

Definition 5.5 (Cross-Proxy Targeted Anchor Regression)**.** Let $\tilde{\mu} \in \mathbb{R}^{d_W}$, and $\tilde{\Sigma}_W \in \mathbb{R}^{d_W \times d_W}$, where $\tilde{\Sigma}_W$ is a symmetric positive semi-definite matrix. We define

$$\begin{aligned}&\ell_{\times TAR}(W, Z; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha)\\ &:= \ell_{LS}(\gamma) + a_\gamma^\top \left(\tilde{\Sigma}_W - \Sigma_W\right) a_\gamma + \left(a_\gamma^\top \tilde{\mu} - \alpha\right)^2,\end{aligned}$$

where a_γ is defined in (5.17).

In Theorem C.1 (Supplement C.4) we prove, analogous to Theorem 5.2, that this population objective is equal to that of Targeted Anchor Regression (5.16).

5.5 Synthetic experiments

In Section 5.5.1, we show that Cross-Proxy Anchor Regression (xPAR) outperforms Proxy Anchor Regression (PAR) in settings with noisy proxies. As the noise increases, xPAR continues to match Anchor Regression (AR) test performance under intervention, while PAR approaches OLS. In Section 5.5.2, we demonstrate the risks of attempting to correct for this noise by assuming a certain signal-to-variance ratio. In Section 5.5.3 we demonstrate another benefit of xPAR over PAR, giving an example where it places more weight on causal predictors relative to PAR. Finally, in Section 5.5.4, we highlight the trade-off between using Targeted Anchor Regression (TAR) vs. OLS and AR, showing that TAR improves performance under the targeted shift, at the cost of incurring additional error on the training distribution. Code for experiments is available at https://github.com/clinicalml/proxy-anchor-regression.

5.5.1 Mean squared prediction error under intervention

We demonstrate on synthetic data that xPAR recovers similar test performance to AR, while the performance of PAR degrades as the signal-to-variance ratio (SVR) of the proxies decreases. We simulate training data (at different levels of signal-to-variance) from an SCM with the structure given in Figure 5-2, fix $\lambda := 5$ and fit PAR and xPAR. We then choose a fixed intervention ν, and simulate test data under the intervened distribution, evaluating our learned predictors.

In Figure 5-5, we see that the test errors for xPAR and AR coincide (see Theorem 5.2) while PAR interpolates between OLS and AR, depending on the signal-to-variance ratio (see Proposition 5.1). Section C.5 gives additional implementation details on this and remaining experiments.

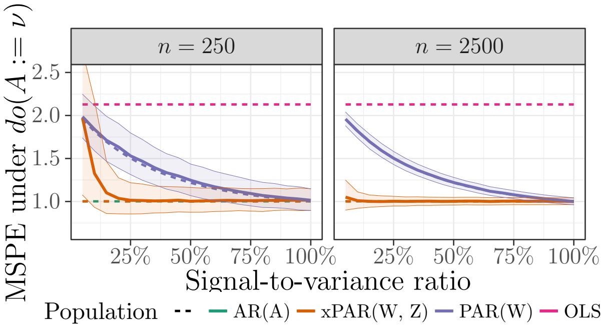


Figure 5-5: *Mean squared prediction error (MSPE) under interventions $do(A := \nu)$ for estimators PAR and xPAR. We display population losses for the population parameters as dashed lines, and median empirical MSPE when fit from data as solid lines, with shaded regions covering the 25% to 75% quantiles.*

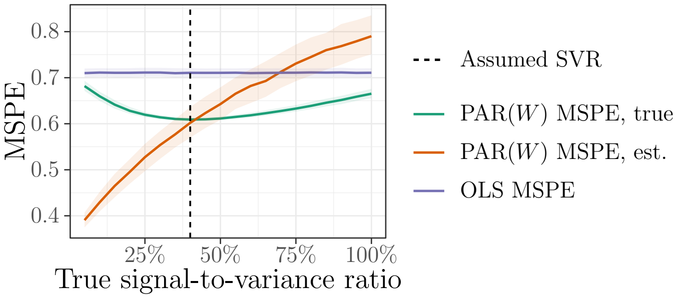


Figure 5-6: *Estimates of worst-case mean squared prediction error (MSPE) over a robustness set C. PAR is applied assuming that the signal-to-variance ratio is 0.4, which gives an estimate of the worst-case MSPE over C (orange). Green line shows actual worst-case MSPE over C at different underlying signal-to-variance ratios.*

5.5.2 Misspecified signal-to-variance ratio

In Section 5.3.1, we noted that if the (unidentified) signal-to-variance ratio (SVR) were known, we could correct for it when using PAR with a single proxy. Here we demonstrate the implications of incorrectly specifying this correction. We simulate data from the same SCM as in Section 5.5.1, with varying (true) signal-to-variance ratio.

In Figure 5-6, for the predictor chosen by PAR, we plot the estimated worst-case MSPE (in orange), using a correction factor assuming that the signal-to-variance ratio is 0.4, against the true worst-case MPSE (in green). We observe that if the true signal-to-variance ratio is smaller than our assumption of 0.4, then our estimate is too conservative, and vice versa if the true signal-to-variance ratio is larger.

5.5.3 Causal and anti-causal predictors

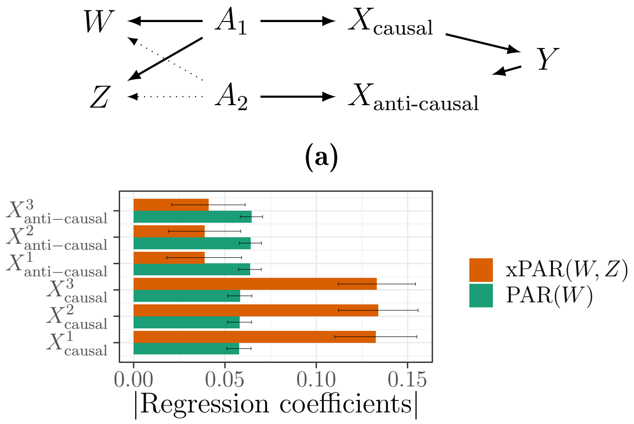
We demonstrate the ability of xPAR to select causal predictors, in a synthetic setting where predictors X may contain both causal and anti-causal predictors. We simulate data from an SCM (Figure 5-7 [top]), where one anchor, A_1, is a parent of the causal predictors, while the other, A_2, is a parent of the anti-causal predictors. We consider two identically distributed noisy proxies W, Z of $A \coloneqq (A_1, A_2)$. The challenge is that A_2 is measured with significantly more noise than A_1, across both proxies.

As seen in Figure 5-7 [bottom] PAR places more weight on anti-causal features. In effect, the noise in the measurement of A_2 causes $X_{\text{anti-causal}}$ to appear less sensitive to shifts in A_2. This is an ideal scenario for xPAR, as it is designed to deal with additional noise by leveraging both proxies. Consequently, when two proxies W, Z are available, xPAR places more weight on the causal predictors, relative to PAR.

5.5.4 Targeted shift

We demonstrate the trade-off made by Targeted Anchor Regression (TAR) versus Anchor Regression (AR), considering the case when A is observed for simplicity. We simulate training data and fit estimators γ_{OLS}, γ_{AR} and γ_{TAR}, where γ_{TAR} is targeted to a particular mean and covariance of a random intervention ν, and we select λ for γ_{AR} such that this intervention is contained within $C_A(\lambda)$.

We then simulate test data from two distributions: $\mathbb{P}_{do(A:=\nu)}$ (i.e., the shift occurs), and $\mathbb{P}$ (where it does not), and evaluate the mean squared prediction error (MSPE). The results are shown in Figure 5-8, and demonstrated that TAR performs better than AR and OLS in the first scenario, but this comes at the cost of worse performance on the training distribution.



(a)

(b)

Figure 5-7: *(a) SCM with A_1, A_2 (unobserved), target Y and predictor variables $X_{causal}, X_{anti\text{-}causal} \in \mathbb{R}^3$. Dotted lines indicate higher noise. (b): Absolute value of regression coefficients. PAR places more weight on anti-causal predictors, while xPAR places more weight on causal predictors.*

5.6 Real-data experiment: Pollution

We test our approach on a real-world heterogeneous dataset of hourly pollution readings in five cities in China, taken over several years (Liang et al., 2016), with most data available from 2013-15. Our prediction target is PM2.5 concentration, a measure of pollution, and covariates are primarily weather-related, including dew point, temperature, humidity, pressure, wind direction / speed, and precipitation.

Real-World Proxy (Temperature): Pollution tends to be seasonal in this dataset, and so we construct our training and test environments using seasons: For each of the four seasons, we train only on the other three seasons, and evaluate on the held-out season. We do this for each city, treating each city and held-out season as a separate evaluation. This leads to 20 separate scenarios.

With this variation in mind, we use temperature as a real-world proxy, and treat it as unavailable at test time. We also construct two noisier copies of temperature, which we refer to as W, Z, adding independent Gaussian noise while controlling the signal-to-variance ratio (in the training distribution) at $\text{Var(Temp)}/\text{Var}(W) = 0.9$.

Estimators / Benchmarks: For Proxy and Cross-Proxy AR (PAR, xPAR, see Sec-

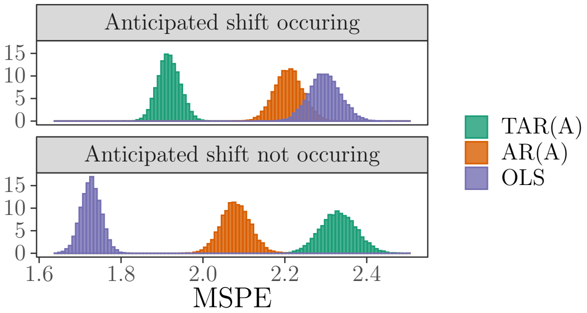


Figure 5-8: *Empirical mean squared prediction error of TAR, OLS and AR under the shifted distribution and the training distribution.*

tion 5.3), we choose $\lambda \in [0, 40]$ by leave-one-group-out cross-validation on the three training seasons, using the first year (2013) of data. For instance, if "winter" is the test season, then we choose the value of λ that performs best on average across combinations of the other seasons e.g., training on the fall & summer data and evaluating on the spring data.

When using temperature as a single proxy in PAR, we observe that in 9 out of 20 scenarios, $\lambda = 40$ is chosen, but in the remaining 11, $\lambda = 0$ is chosen, which is equivalent to OLS. For comparability, we use the same values of λ for PAR(W) and xPAR(W, Z). For Proxy Targeted AR and Cross-Proxy Targeted AR (PTAR, xPTAR, see Section 5.4), we use the mean and variance of the relevant variables (e.g., temperature, W, Z) in the held-out season to target our predictors.

Our primary benchmark is OLS (without temperature). We also compare to (a) OLS that uses temperature during train and test [OLS (TempC)], and (b) OLS that includes the temperature during training, and uses the mean test value for temperature during prediction [OLS + Est. Bias]. We present the results for the 9 scenarios where $\lambda > 0$ in Table 5.1, since PAR with $\lambda = 0$ is equivalent to OLS (aggregate results in Table C.1 in the supplement).

Results: For both PAR and PTAR, we see improvement over OLS on average across scenarios, with limited downside (e.g., in the worst scenario for PTAR relative to OLS, the additional MSE incurred is 0.001). In Figure C-4 (Supplement), we observe that PAR and PTAR achieve gains in two different ways: PAR increases the coefficients of

Table 5.1: *Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.*

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
PAR (TempC)	0.531	6	-0.041	0.006
PAR (W)	0.531	6	-0.037	0.006
xPAR (W, Z)	0.531	6	-0.039	0.007
PTAR (TempC)	0.525	8	-0.061	0.001
PTAR (W)	0.529	8	-0.038	0.001
xPTAR (W, Z)	0.526	7	-0.059	0.001

humidity and dew point relative to OLS, while PTAR reduces them and incorporates a correction into the intercept.

5.7 Discussion and related work

Learning a predictive model that performs well under arbitrarily strong causal interventions is an ambitious goal. In this work, we have argued that even if causal invariance is achievable, it may not be desirable: A model whose performance is invariant to arbitrarily strong interventions may have poor performance when the test distribution does not differ too much from the training distribution.

There is a large body of work that seeks to learn causal models as a route to achieving invariance (Rojas-Carulla et al., 2018; Magliacane et al., 2018), or that uses knowledge of the causal graph to select predictors with invariant performance under a set of known interventions (Subbaswamy et al., 2019). Similarly, invariant risk minimization (IRM) seeks a predictor Φ such that $\mathbb{E}(Y \mid \Phi(X))$ is invariant across a set of discrete environments (Arjovsky et al., 2019; Xie et al., 2020; Krueger et al., 2020; Bellot and van der Schaar, 2020). Recent work has pointed to the theoretical and practical difficulty of learning such a predictor for IRM (Rosenfeld and Risteski, 2020; Kamath et al., 2021; Guo et al., 2021), in part due to the fact that recovering a truly invariant model, even in linear settings, requires a large number of environments. Generalization in non-linear settings requires sufficient overlap between environments and strong restrictions on the model class (e.g., Christiansen et al., 2020). In contrast to all of the above, we trade off between in-distribution performance and invariance explicitly, instead of seeking invariance as a primary goal. Moreover, since we allow for A to influence Y directly and through hidden variables, invariance may not even be achievable, but we can still formulate a worst-case loss for bounded interventions.

We argue for incorporating prior knowledge about potential shifts by (1) identifying proxies for relevant factors of variation (i.e., anchor variables), and (2) specifying plausible sets of interventions on these factors of variation. We build upon the causal framework of Anchor Regression (Rothenhäusler et al., 2021), extending it in two important ways.

To start, we relax the assumption that the anchor variables are directly observed. Instead, we only assume access to proxies, and prove that identification of the worst-case loss is feasible with two proxies. The challenge of identifying the worst-case loss is related to the problem of identifying causal effects with noisy proxies of unmeasured confounders (Tchetgen Tchetgen et al., 2020; Miao and Tchetgen, 2018; Shi et al., 2018; Kuroki and Pearl, 2014), and the challenge of learning under classical measurement error (Fuller, 1987; Hyslop and Imbens, 2001; Bound et al., 2001). Our observation that a single proxy will underestimate the worst-case loss is related to the well-known problem of regression dilution bias (Frost and Thompson, 2000), where performing linear regression under measurement error leads to bias in parameter estimation. In contrast, we are not concerned with causal / structural parameter estimation, which is generally not possible in the models we consider, but rather estimating a worst-case loss under a class of interventions. Srivastava et al. (2020) also consider distributional shift in unmeasured variables for which proxies are available, and apply techniques for handling worst-case sub-populations from DRO (Duchi et al., 2020b). In contrast, we consider causal interventions on A that could lie outside the support of the training data, which cannot be represented as a sub-population. Moreover, they consider the single-proxy case, and give a generalization bound that incorporates the impact of noise, while under our assumptions we are able to recover guarantees as if A were observed, using two proxies.

We then introduce Targeted Anchor Regression, a method for incorporating additional prior knowledge on the strength and direction of shifts in anchor variables. This method can be interpreted as allowing for specification of a broader class of robustness sets, beyond those considered in Rothenhäusler et al. (2021), or as specifying the mean and covariance of the anchors at test time. We prove analogous results with proxies in this setting, and evaluate this strategy empirically in Section 5.6, targeting our loss to a particular mean and variance over temperature in the held-out season.

Our work contributes to a growing body of literature that seeks to generalize Anchor Regression to new settings, whether allowing for unobserved anchors and a broader class of robustness sets (as in our work), or generalizing to discrete and censored outcomes, as in Kook et al. (2022).

Chapter 6

Evaluating Robustness to Dataset Shift via Parametric Robustness Sets

This chapter (and accompanying appendix) was previously published as (Thams et al., 2022) at NeurIPS 2022.

6.1 Introduction

Predictive models may perform poorly outside of the training distribution, a problem broadly known as dataset shift (Quiñonero-Candela et al., 2008). In high-stakes applications, such as healthcare, it is important to understand the limitations of a model in advance (Finlayson et al., 2021): given a model trained on data from one hospital, how will it perform under changes in the population of patients, in the incidence of disease, or in the treatment policy?

In this chapter, our goal is to **proactively** understand the sensitivity of a predictive model to dataset shift, using only data from the training distribution. This requires domain knowledge, to specify what type of distributional changes are plausible. Formally, for a model $f(X)$ trained on data from $\mathbb{P}(X, Y)$, with loss function $\ell(f(X), Y)$, we seek to understand the loss of the model under a set of *plausible* future distributions

$\mathcal{P}$. We seek to evaluate the worst-case loss over $\mathcal{P}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(f(X), Y)], \tag{6.1}$$

and provide an interpretable description of a distribution P which maximizes this objective. If the value of the worst-case loss is low, this can build confidence prior to deployment, and otherwise, examining the worst-case distribution P can help identify weaknesses of the model. To illustrate, we use the following running example, inspired by Subbaswamy et al. (2021).

Example 6.1 (Changes in laboratory testing)**.** We seek to classify disease (Y) based on the age (A) of a patient, whether a laboratory test has been ordered (O), and test results (L) if a test was ordered. The performance of a predictive model may be sensitive to changes in testing policies, as the *fact that a test has been ordered* itself is predictive of disease. Figure 6-1a gives a plausible causal relationship between variables. Let $\mathbb{P}(O = 1|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid function and $\eta(A, Y)$ is the log-odds. In Figure 6-1b, we show the loss under a set of new distributions parameterized by $\delta = (\delta_0, \delta_1)$, where we modify $\mathbb{P}_\delta(O = 1|A, Y) = \sigma(\eta(A, Y) + s(Y; \delta))$ for a *shift function* $s(Y; \delta) = \delta_1 \cdot Y + \delta_0 \cdot (1 - Y)$, which modifies the log-odds of testing for both sick and healthy patients. If δ_0, δ_1 are unconstrained, the worst-case occurs when all healthy patients are tested, and no sick patients are tested.

The first challenge is to define a set of possible distributions $\mathcal{P}$ such that each distribution $P \in \mathcal{P}$ satisfies two desiderata: First, they should be *causally interpretable and simple to specify*, without placing unnecessary restrictions on the data-generating process. Second, they should be *realistic*, which often entails bounding the magnitude of the shift. We construct causally interpretable shifts by defining perturbed distributions $\mathbb{P}_\delta$ using changes in causal mechanisms, parameterized by a finite-dimensional parameter δ. Our main requirement is that the shifting mechanisms follow a conditional exponential family distribution. For discrete variables, this places no restriction on $\mathbb{P}$: In Example 6.1, O is binary and the log-odds $\eta(A, Y)$ can be any function of A, Y. We also demonstrate that constraining δ can ensure that shifts are realistic:

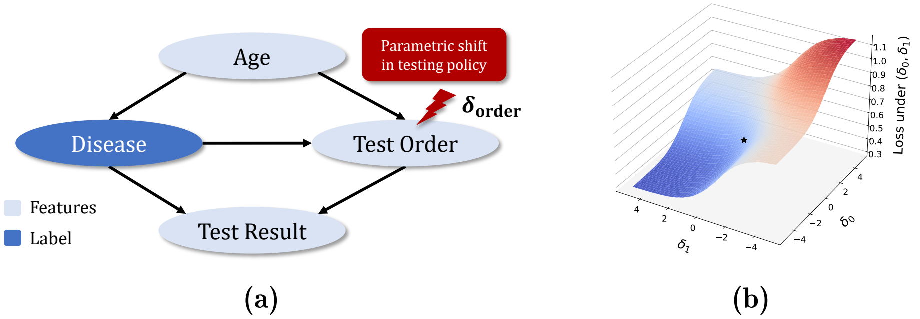


Figure 6-1: *(a) Causal graph for Example 6.1, with a shift in conditional testing rates, parameterized by δ_{order}. (b) We illustrate a shift using $s(Y; \delta_{order}) = \delta_1 \cdot Y + \delta_0(1 - Y)$, where $\delta_{order} = (\delta_0, \delta_1)$. Here we plot the (non-concave) landscape of the expected cross-entropy loss of a fixed model over distributions parameterized by (δ_0, δ_1), with the training distribution given as the black star. Simulation details are given in Appendix D.1.*

The unconstrained worst-case shift in Example 6.1 is implausible, where all healthy patients (and no sick patients) are tested. Equation (6.1) becomes

$$\sup_{\delta \in \Delta} \mathbb{E}_\delta[\ell(f(X), Y)], \tag{6.2}$$

where $\mathbb{E}_\delta$ is the expectation in the shifted distribution $\mathbb{P}_\delta$ and Δ is a bounded set of shifts.

The second challenge is evaluation of the expected loss under shift, as well as finding the worst-case shift. Under our definition of shifts, we show that the test distribution can always be seen as a reweighting of the training distribution, allowing for reweighting approaches, such as importance sampling, to estimate the expected loss under shifts. While this is practical for some distribution shifts, for others, importance sampling can lead to extreme variance in estimation. Further, finding the worst-case shift using a reweighted objective involves maximization over a non-concave objective (see Figure 6-1), a problem that is generally NP-hard. We derive a second-order approximation to the expected loss under shift, and show how it can be estimated without the use of reweighting. When Δ is a single quadratic constraint, we can approximate the general non-convex optimization problem in Equation (6.2) with a particular non-convex, quadratically constrained quadratic program (QCQP) for which efficient solvers exist (Conn et al., 2000, Section 7). We bound the approximation error of this surrogate objective, and show in experiments that it tends to find impactful adversarial shifts.

Our contributions are as follows:

1. We provide a novel formulation of robustness sets which are defined using parametric shifts. This formulation only require that the shifting mechanisms (i.e., conditional distributions) can be modelled as a conditional exponential family (see Section 6.2).

2. We derive a second-order approximation to the expected loss and provide a bound on the approximation error. We show that this translates the general non-convex problem into a particular non-convex quadratic program, for which efficient solvers exist (see Section 6.3).

3. In a computer vision task, we find that this approach finds more impactful shifts than a reweighting approach, while taking far less time to compute, and that the resulting estimates of accuracy are substantially more reliable (see Section 6.4).

6.1.1 Related Work

Distributionally robust optimization/evaluation: Distributionally robust optimization (DRO) seeks to learn models that minimize objectives like Equation (6.1) with respect to the model (Duchi and Namkoong, 2021; Duchi et al., 2020b; Sagawa et al., 2020). We focus on proactive worst-case evaluation of a fixed model, not optimization, similar to Subbaswamy et al. (2021); Li et al. (2021), but we also differ in our **definition of the set of plausible future distributions** $\mathcal{P}$, often called an "uncertainty set" in the optimization literature. Prior work often defines these sets using distributional distances (such as f-divergences): For instance, Joint DRO (Duchi and Namkoong, 2021) allows for shifts in the entire joint distribution (i.e., all distributions in an f-divergence ball around $\mathbb{P}(X, Y)$), which may be overly conservative. Marginal DRO (Duchi et al., 2020b) considers shifts in a marginal distribution (e.g., $\mathbb{P}(X)$), while assuming that the remaining conditionals (e.g., $\mathbb{P}(Y \mid X)$) are fixed. However, this assumption is not applicable in all scenarios: In Example 6.1, for instance, this assumption does not hold for a shift in testing policy. Conditional shifts are considered in recent work that focuses on evaluation (Subbaswamy et al., 2021), using worst-case conditional subpopulations. However, choosing a plausible size of conditional subpopulation is often non-obvious. In Appendix D.4 we give a simple lab-testing example where taking worst-case 20% conditional subpopulations corresponds to an implausible shift: Healthy patients are always tested, and sick patients never tested.

In contrast, our approach uses explicit parametric perturbations to define shifts, as opposed to distributional distances or subpopulations. In addition, our approach allows for shifts in multiple marginal or conditional distributions simultaneously: In Example 6.1, for instance, we could model a simultaneous change in both the marginal distribution of age $\mathbb{P}(A)$, as well as the conditional distribution of lab testing $\mathbb{P}(O \mid A, Y)$, leaving other factors unchanged.

Causality-motivated methods for learning robust models: Several approaches proactively specify shifting causal mechanisms/conditional distributions, and then seek to learn predictors that have good performance under arbitrarily large changes in these mechanisms (Subbaswamy et al., 2019; Veitch et al., 2021; Makar et al., 2022; Puli et al., 2022). Other approaches use environments (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2019) or identity indicators (Heinze-Deml and Meinshausen, 2021) to learn models that rely on invariant conditional distributions.

However, when shifts are not arbitrarily strong, causality-motivated predictors can be overly conservative. In Example 6.1, a model that ignores all test-related features (and only uses age as a predictor) is a particularly simple example of a causality-motivated predictor, with invariant risk over changes in testing policy. Closer to our setting is a line of work that considers bounded mechanism changes in linear causal models (Rothenhäusler et al., 2021; Oberst et al., 2021b), where estimation of the worst-case loss enables learning of worst-case optimal models. Our work can be seen as extending this idea to more general non-linear causal models, where we focus on evaluation rather than optimization.

Evaluating out-of-distribution performance with unlabelled samples: A recent line of work has focused on predicting model performance in out-of-distribution settings, where unlabelled data is available from the target distribution (Garg et al., 2022; Jiang et al., 2022; Chen et al., 2021). In contrast, our method operates using only samples from the original source distribution, and seeks to estimate the worst-case loss over a set of possible target distributions.

In Appendix D.6 we give a more detailed discussion of these approaches and others.

6.2 Defining parametric robustness sets

Notation: Let $\mathbf{V}$ denote all observed variables, where $(X, Y) \subseteq \mathbf{V}$ for features X and labels Y, and use $\mathbb{P}(\mathbf{V})$ to denote the probability density/mass function in the training distribution. We also refer to $\mathbb{P}$ as simply "the training distribution". $\mathbb{E}[\cdot]$ and $\mathrm{Cov}(\cdot, \cdot)$ refer to the mean and covariance in $\mathbb{P}$, and for a shifted distribution $\mathbb{P}_\delta$ (Definition 6.1) we use $\mathbb{E}_\delta[\cdot]$, $\mathrm{Cov}_\delta(\cdot, \cdot)$. For a random variable Z, we use $\mathcal{Z}$ to denote the space of realizations, and d_Z for dimension e.g., $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$. For a set of random variables $\mathbf{V} = \{V_1, \dots, V_d\}$, we use V_i to denote an individual element, and use $\mathrm{PA}_\mathcal{G}(V_i)$ to denote the set of parents in a directed acyclic graph (DAG) $\mathcal{G}$, omitting the subscript when otherwise clear.

We begin with a general definition of a parameterized robustness set of distributions $\mathcal{P}$.

Definition 6.1. A *parameterized robustness set around* $\mathbb{P}(\mathbf{V})$ is a family of distributions $\mathcal{P}$ with elements $\mathbb{P}_\delta(\mathbf{V})$ indexed by $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$, with $0 \in \Delta$, where $\mathbb{P}_0(\mathbf{V}) = \mathbb{P}(\mathbf{V})$.

We give examples shortly that satisfy this general definition. To construct such a robustness set, we consider distributions $\mathbb{P}_\delta$ that differ from $\mathbb{P}$ in one or more conditional distributions (Assumption 6.1). We require that the relevant conditional distributions can be described by an exponential family.

Definition 6.2 (Conditional exponential family (CEF) distribution)**.** $\mathbb{P}(W|Z)$ is a conditional exponential family distribution if there exists a function $\eta(Z) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$

such that the conditional probability density (for continuous W) or probability mass function (for discrete W) is given by

$$\mathbb{P}(W|Z) = g(W)\exp\left(\eta(Z)^\top T(W) - h(\eta(Z))\right), \tag{6.3}$$

where $T(W)$ is a vector of sufficient statistics, $T(W) \in \mathbb{R}^{d_T}$, $g(\cdot)$ specifies the density of a base measure and $h(\eta(Z))$ is the log-partition function.

Definition 6.2 does not restrict $\mathbb{P}(W|Z)$ for binary/categorical W, and captures a wide range of distributions, including the conditional Gaussian (see Appendix D.2.1 for other examples). Definition 6.2 extends to marginal distributions where $Z = \varnothing$ and $\eta(Z)$ is a constant function.

Example 6.1 (Continued)**.** Suppose the probability of ordering a test (O) depends on age (A) and disease (Y), such that $\mathbb{P}(O = 1|A, Y) = \sigma(\eta(A, Y))$, where σ is the sigmoid, and η is an arbitrary function. Here, Definition 6.2 is satisfied with $W = O$, $Z = (A, Y)$, and sufficient statistic $T(O) = O$.

We now state our main assumption, where we distinguish between the terms in the joint distribution of $\mathbb{P}$ that shift, which we will need to model, and those that remain fixed, which we do not.

Assumption 6.1 (Factorization into CEF distributions)**.** Let $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ be a "intervention set" of variables and let

$$\mathbb{P}(\mathbf{V}) = \underbrace{\prod_{W_i \in \mathbf{W}} \mathbb{P}(W_i|Z_i)}_{\text{Conditionals that shift}} \underbrace{\prod_{V_j \in \mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)}_{\text{Conditionals we do not model}} \tag{6.4}$$

be a factorization, where $Z_i, U_j, V_j \subseteq \mathbf{V}$ are possibly overlapping (or empty) sets of variables, where $\mathbb{P}(V_j \mid \varnothing) := \mathbb{P}(V_j)$. For each W_i we assume Z_i is known and $\mathbb{P}(W_i|Z_i)$ satisfies Definition 6.2.

If $\mathbb{P}(\mathbf{V})$ factorizes according to a DAG $\mathcal{G}$, the factorization in Assumption 6.1 is always satisfied by $Z_i = \mathrm{PA}_{\mathcal{G}}(W_i)$. While we assume data is generated according to

Equation (6.4), we do not require knowledge of the full distribution, but only the conditionals that shift. In Appendix D.2.2 we show that we can also consider shifts that extend Z_i to include additional variables, subject to an acyclicity constraint. We now define parametric perturbations and give the general form of the robustness sets that we consider in this work, involving simultaneous perturbations to multiple W_i.

Definition 6.3 (Parameterized shift functions and δ-perturbations)**.** Let $s(Z;\delta) : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_T}$ be a *parameterized shift function* with parameters $\delta \in \Delta \subseteq \mathbb{R}^{d_\delta}$ which is twice-differentiable with respect to δ and which satisfies $s(Z;0) = 0$ for all Z. For $\mathbb{P}(W|Z)$ satisfying Equation (6.3), we refer to

$$\mathbb{P}_\delta(W|Z) = g(W)\exp\left(\eta_\delta(Z)^\top T(W) - h(\eta_\delta(Z))\right)$$

as a δ-perturbation of $\mathbb{P}(W|Z)$ with shift function $s(Z;\delta)$, where $\eta_\delta(Z) := \eta(Z)+s(Z;\delta)$. Note that this differs from Equation (6.3) in that $\eta(Z)$ is replaced by $\eta_\delta(Z)$.

Example 6.1 (Continued)**.** A model developer may be concerned about a uniform change in testing rates across all types of patients. This can be modelled by choosing $s(Z;\delta) = \delta$, for $\delta \in \mathbb{R}$, an additive intervention on the log-odds scale. A separate change in testing rates for sick and healthy patients could instead be modeled using $s(Z;\delta) = \delta_0(1-Y) + \delta_1 Y$, using $\delta \in \mathbb{R}^2$. This reasoning extends readily to more complex shifts (e.g., allowing for age-specific changes in testing rates, with a non-linear dependence on age), as long as $s(Z;\delta)$ remains a parametric function.

While the shift function $s(Z;\delta)$ is parametric, $\eta(Z)$ is unconstrained in Definitions 6.2 and 6.3. Note that this formulation includes multiplicative shifts $\eta_\delta(Z) = (1+\delta)\eta(Z)$ by letting $s(Z;\delta) = \delta \cdot \eta(Z)$.

Definition 6.4 (CEF parameterized robustness set)**.** For a distribution $\mathbb{P}$ and intervention set $\mathbf{W} = \{W_1, \ldots, W_m\} \subseteq \mathbf{V}$ satisfying Assumption 6.1, let each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ be a

δ_i-perturbation (Definition 6.3) of $\mathbb{P}(W_i|Z_i)$. Then

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i|Z_i)\right)\left(\prod_{V_j \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_j|U_j)\right)$$

is called a δ-perturbation of $\mathbb{P}(\mathbf{V})$, and the robustness set $\mathcal{P}$ consists of all $\mathbb{P}_\delta$ for $\delta \in \Delta_1 \times \cdots \Delta_m$.

To estimate the expected loss under $\mathbb{P}_\delta$, we will typically[1] need to estimate $\eta(Z_i)$ for each $W_i \in \mathbf{W}$. However, we make no distributional assumptions on the remaining variables $\mathbf{V} \setminus \mathbf{W}$. This is useful in applications such as computer vision, where we do not need to restrict the generative model of images given attributes (e.g., background, camera type, etc), but can still model the expected loss under changes in the joint distribution of those attributes.

Remark 4 (Causal Interpretation of Shifts)*.* If available, causal knowledge helps identify which factors in the joint distribution are subject to shifts (e.g., $\mathbb{P}(O \mid Y, A)$ in Example 6.1), and which remain stable. It is worth noting, however, that our methodology can be used to model any change in distribution that satisfies Assumption 6.1, including choices of "non-causal" factorizations and shifting factors. For example, in the context of Example 6.1, we could choose the factorization $\mathbb{P}(Y)\mathbb{P}(O \mid Y)\mathbb{P}(L, A \mid O, Y)$, and model a change only in the conditional $\mathbb{P}(O \mid Y)$ while keeping other factors unchanged. This shift is not interpretable as a change in causal mechanisms: The shifted distribution would imply a change in the marginal distribution of age, which should be unaffected by a real-world change in laboratory testing. Nonetheless, we can still estimate a worst-case loss over such non-causal shifts in distribution. In short, our machinery can model shifts in non-causal conditionals (for example because the causal structure is unknown), though the resulting shifted distribution is not interpretable as a plausible shift in the ground-truth data generating mechanism.

[1] As a special case, in Appendix D.3.2, we show the second-order approximation (Theorem 6.1) can be estimated in the case of variance-scaled mean-shifts in a conditional Gaussian without estimation of all of $\eta(Z)$.

6.3 Evaluation of the worst-case loss

For a fixed predictor and loss function, we can use data from $\mathbb{P}(\mathbf{V})$ to estimate the expected loss $\mathbb{E}_\delta[\ell] := \mathbb{E}_\delta[\ell(f(X), Y)]$ for a fixed δ, and estimate the worst-case loss over all δ of bounded magnitude. In Section 6.3.1, we show that $\mathbb{P}_\delta$ shares support with $\mathbb{P}$, suggesting the use of reweighting estimators. However, these estimators can exhibit high variance for shifts that produce large density ratios (see Appendix D.3.5 for an example), and maximizing a reweighted objective over δ is generally a non-convex problem. In Section 6.3.2 we derive an approximation to the expected loss under $\mathbb{P}_\delta$, yielding a tractable surrogate optimization problem under quadratic constraints such as $\|\delta\|_2 \leq \lambda$.

Remark 5. The methods here can be used with an arbitrary predictor f and loss function $\ell := \ell(f(X), Y)$. We do not even require access to the original predictor f. Both methods here simply treat ℓ as a random variable in $\mathbb{P}$, for which we have samples from the training distribution.

6.3.1 Modelling shifted losses using reweighting

The shifts defined in Section 6.2 share common support, with the following density ratio.

Proposition 6.1. *For any $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ that satisfy Definition 6.4,* $\operatorname{supp}(\mathbb{P}) = \operatorname{supp}(\mathbb{P}_\delta)$ *and the density ratio $w_\delta := \mathbb{P}_\delta/\mathbb{P}$ is given by*

$$w_\delta(\mathbf{V}) = \exp\left(\sum_{i=1}^{m} s_i(Z_i;\delta_i)^\top T_i(W_i)\right) \exp\left(\sum_{i=1}^{m} h(\eta_i(Z_i)) - h(\eta(Z_i) + s_i(Z_i;\delta_i))\right).$$

The proof can be found in Appendix D.7, along with all proofs for all other claims.

Example 6.1 (Continued)**.** Suppose we perturb the probability of ordering a test O given age A and disease Y with shift function $s(Y;\delta) = \delta_0(1-Y)+\delta_1 Y$, independently changing the conditional probability of testing for healthy and sick patients. Here, the density ratio is given by

$$w_\delta(O, A, Y) = \exp(s(Y;\delta)\cdot O)\frac{1+\exp(\eta(A,Y))}{1+\exp(\eta(A,Y)+s(Y;\delta))}. \tag{6.5}$$

To model the loss $\mathbb{E}_\delta[\ell]$ using data from $\mathbb{P}$, we can consider an importance sampling (IS) estimator (Horvitz and Thompson, 1952; Shimodaira, 2000), observing that $\mathbb{E}_\delta[\ell] = \mathbb{E}[w_\delta(\mathbf{V})\cdot\ell]$. This requires estimation of the density ratio $w_\delta(\mathbf{V})$, and (given a sample $\{\mathbf{V}^j\}_{j=1}^n$ from $\mathbb{P}$) yields the estimator

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\text{IS}} := \frac{1}{n}\sum_{j=1}^{n}\hat{w}_\delta(\mathbf{V}^j)\ell(\mathbf{V}^j). \tag{6.6}$$

Equation (6.6) can have high variance when density ratios are large, and maximizing this equation with respect to δ is a general non-convex optimization problem, which is generally NP-hard to solve.

6.3.2 Approximating the shifted loss for exponential family models

We now propose an alternative approach for approximating the loss $\mathbb{E}_\delta[\ell]$. Recalling that $\mathbb{P}_{\delta=0} = \mathbb{P}$, we use a second-order Taylor expansion around the training distribution

$$\mathbb{E}_\delta[\ell] \approx \mathbb{E}[\ell] + \delta^\top \,\mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \,\mathrm{SG}^2\,\delta, \tag{6.7}$$

where $\mathbb{E}[\ell]$ denotes the loss in the training distribution and $\mathrm{SG}^1, \mathrm{SG}^2$ are defined as follows.

Definition 6.5 (Shift gradient and Hessian)**.** For a parametric shift satisfying Definition 6.1 where $\delta \mapsto \mathbb{E}_\delta[\ell]$ is twice-differentiable, we denote the *shift gradient* SG^1 and *shift Hessian* SG^2 as

$$\mathrm{SG}^1 := \nabla_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0} \qquad \text{and} \qquad \mathrm{SG}^2 := \nabla^2_\delta \mathbb{E}_\delta[\ell]\big|_{\delta=0}.$$

Equation (6.7) is a local approximation of the loss, whose approximation error we bound in Theorem 6.2, with smaller approximation error for smaller shifts.[2] For $\mathbb{P}_\delta$ satisfying Definition 6.4, SG^1 and SG^2 can be computed as expectations in the training distribution, without estimation of density ratios. Recall that the conditional covariance is given by $\mathrm{Cov}(A, B|C) := \mathbb{E}[(A - \mathbb{E}[A|C])(B - \mathbb{E}[B|C])|C]$.

Theorem 6.1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 6.4, with intervened variables $\mathbf{W} = \{W_1, \ldots, W_m\}$ and shift functions $s_i(Z_i; \delta_i)$, where $\delta = (\delta_1, \ldots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \ldots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$\mathrm{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top Cov\Big(\ell,\, T_i(W_i)\Big| Z_i\Big)\right],$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$, where the (i,j)th block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$\{\mathrm{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top Cov\left(\ell,\, \epsilon_{T_i|Z_i}\epsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \epsilon_{T|Z}\right] & i = j \\ Cov(\ell,\; D_{i,1}^\top \epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i; \delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here, $T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Theorem 6.1 handles arbitrary parametric shift functions in multiple variables, but for simple shift functions in a single variable, the notation simplifies substantially, as we show in Corollary 6.1.

Corollary 6.1 (Simple shift in a single variable)**.** *Assume the setup of Theorem 6.1, restricted to a shift in a single variable W, and that $s(Z; \delta) = \delta$. Then $D_1 = 1$, $D_2 = 0$,*

[2] In Appendix D.3.3, we give an example of a linear-Gaussian generative model where this second-order expansion is exact, corresponding to the setting of Anchor Regression (Rothenhäusler et al., 2021).

and

$$\mathrm{SG}^1 = \mathbb{E}\left[Cov\Big(\ell, T(W)\Big|Z\Big)\right] \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathbb{E}\left[Cov\Big(\ell, \epsilon_{T|Z}\epsilon_{T|Z}^\top\Big|Z\Big)\right],$$

where $T(W)$ is the sufficient statistic of W and $\epsilon_{T|Z} := T(W) - \mathbb{E}[T(W)|Z]$.

Example 6.1 (Continued)**.** Suppose that age (A) follows a normal distribution with mean μ and variance σ^2, and consider a shift in the mean (without changing lab testing). We can parameterize $\mathbb{P}(A)$ as an exponential family with parameter $\eta = \mu/\sigma$ and sufficient statistic $T(A) = A/\sigma$. Here, $s(\delta) = \delta$ implies a shift in the mean of δ standard deviations $\eta_\delta = \eta + s(\delta) = (\mu + \sigma\delta)/\sigma$, and we can write that $\mathrm{SG}^1 = \mathrm{Cov}\left(\ell, A\right)/\sigma$ and $\mathrm{SG}^2 = \mathrm{Cov}\left(\ell, (A - \mathbb{E}[A])^2\right)/\sigma^2$.

To estimate the shift gradient and Hessian from a sample from $\mathbb{P}$, for each $i = 1, \ldots, m$ we fit models $\hat{\mu}_\ell(Z_i) \approx \mathbb{E}[\ell|Z_i]$ and $\hat{\mu}_{W_i}(Z_i) \approx \mathbb{E}[T_i(W_i)|Z_i]$ and compute residuals on these predictions, which permits estimation of the gradient/Hessian as a sample average of residuals. A detailed treatment is given in Appendix D.3.1. Using estimates of the gradient and Hessian, we estimate the expected loss as

$$\mathbb{E}_\delta[\ell] \approx \hat{E}_{\delta,\mathrm{Taylor}} := \hat{\mathbb{E}}[\ell] + \delta^\top \hat{\mathrm{SG}}^1 + \frac{1}{2}\delta^\top \hat{\mathrm{SG}}^2 \delta. \tag{6.8}$$

Here, there are two sources of error: Finite-sample error, due to the estimates of $\mathrm{SG}^1, \mathrm{SG}^2$, as well as approximation error. The latter is bounded by the norm of δ and a term that depends on the covariance between the loss and the deviations of the sufficient statistic from its shifted mean.

Theorem 6.2. *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy the conditions of Theorem 6.1, with a shift in a single variable W, where $s(Z;\delta) = \delta$. Let $E_{\delta,Taylor}$ be the population Taylor estimate (Equation (6.7)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,Taylor}\right| \leq \tfrac{1}{2}\sup_{t\in[0,1]} \sigma\Big(Cov_{t\cdot\delta}(\ell, \epsilon_{t\cdot\delta,T|Z}\epsilon_{t\cdot\delta,T|Z}^\top) - Cov(\ell, \epsilon_{0,T|Z}\epsilon_{0,T|Z}^\top)\Big)\cdot\|\delta\|^2,$$

where $T(W)$ is the sufficient statistic of $W|Z$ and $\epsilon_{t\cdot\delta,T|Z} = T(W|Z) - \mathbb{E}_{t\cdot\delta}[T(W|Z)]$.

To build intuition, in Appendix D.3.8 we give a scenario where this bound can be simplified. In particular, we consider a "covariate shift" setting (Quiñonero-Candela et al., 2008) where X is standard Gaussian, $Y = f_0(X)+\epsilon$ with a noise term independent of X and we consider a shift δ in the mean of X. When evaluating a predictor $f(X)$ with the loss ℓ being the squared error, the bound in Theorem 6.2 depends on how the modelling error $g(X) = f_0(X) - f(X)$ behaves over the domain. In particular, the bound scales as the supremum (over $t \in [0,1]$) of $\sqrt{\mathrm{Var}(g(X+t\cdot\delta)^2 - g(X)^2)}$. As a simple corollary, if our predictor is off by an additive constant factor, $f = f_0 + C$, then the bound is zero, and the approximation is exact for any δ. On the other hand, if the squared modelling error $g(X)^2$ at one point X tends to be a poor predictor of the squared modelling error at another point $X + t\cdot\delta$, then this variance will be large, and the approximation will be loose.

In exchange for considering a second-order approximation of the loss, we gain two benefits: Variance reduction and tractable optimization. First, the variance of $\hat{E}_{\delta,\mathrm{Taylor}}$ is $O(\|\delta\|^4)$ for large $\|\delta\|$, while the variance of $\hat{E}_{\delta,\mathrm{IS}}$ can be much larger: We give a simple case in Appendix D.3.6 where $\mathrm{Var}(\hat{E}_{\delta,\mathrm{Taylor}}) = O(\delta^4)$ while $\mathrm{Var}(\hat{E}_{\delta,\mathrm{IS}}) = O(\delta^2\exp(\delta^2))$. Second, maximizing $\hat{E}_{\delta,\mathrm{Taylor}}$ over the set $\|\delta\| \leq \lambda$ can be solved in polynomial time by exploiting the quadratic structure, while maximizing $\hat{E}_{\delta,\mathrm{IS}}$ over the constraints is generally hard, and may be infeasible in high dimensions.

6.3.3 Identifying worst-case parametric shifts

For $\lambda > 0$, we can locally approximate the worst-case loss over all distributions $\mathbb{P}_\delta$ where $\|\delta\|_2 \leq \lambda$ by finding the worst-case loss in the Taylor approximation

$$\sup_{\|\delta\|_2\leq\lambda} \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\,\delta. \tag{6.9}$$

Since SG^2 is generally not negative definite, the maximization objective is non-concave. However, this particular problem is an instance of the 'trust region problem'[3] which is well-studied in the optimization literature (Conn et al., 2000), and can be solved in polynomial time by specialized algorithms (see Pólik and Terlaky (2007, Section 8.1) for an example). This follows from the fact that strong duality holds, so that the optimal solution δ^* can be characterized in terms of the Karush-Kuhn-Tucker conditions (Boyd and Vandenberghe, 2004, Section 5.2). For this problem, we use the `trsapp` routine from NEWUOA (Powell, 2006), as implemented in the python package `trustregion`. Depending on the application and prior knowledge, one may choose constraint sets that differ from $\|\delta\| \leq \lambda$. In particular, the strong duality of Equation (6.9) also holds when $\|\delta\|_2 \leq \lambda$ is replaced by any single quadratic constraint $\delta^\top A \delta + \delta^\top b \leq \lambda$, allowing for e.g., larger shifts in some directions than in others.

6.4 Experiments

6.4.1 Illustrative example: Laboratory testing

To build intuition, we illustrate our method in a simple generative model, similar to Example 6.1, where lab tests are more likely to be ordered (O) for sick patients (Y), and lab values (L) are predictive of Y.

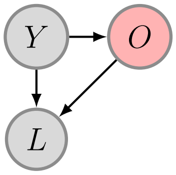
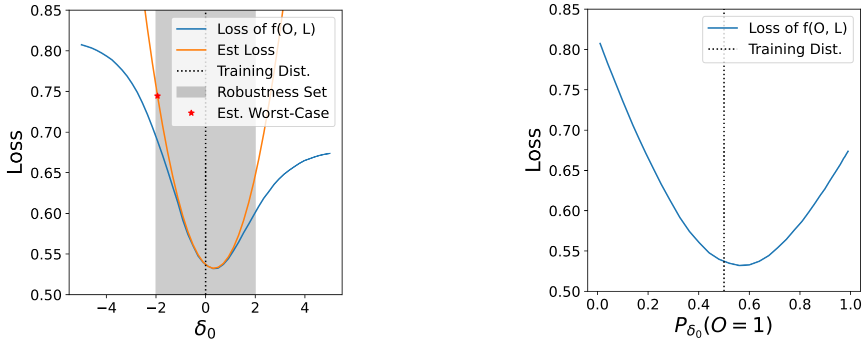


Figure 6-2

$$Y \sim \mathsf{Bern}(0.5) \quad O|Y \sim \mathsf{Bern}(\sigma(\alpha + \beta Y)) \quad L|(Y, O=1) \sim \mathcal{N}(\mu_y, 1)$$

where $\mu_1 = 0.5, \mu_0 = -0.5$, and we initialize with $\alpha = -1$, $\beta = 2$, so that $\mathbb{P}(O = 1|Y = 0) \approx 0.27$ and $\mathbb{P}(O = 1|Y = 1) \approx 0.73$, and the marginal probability of test ordering is $\mathbb{P}(O = 1) = 0.5$. When $O = 0$, we set L to a dummy value of $L = 0$. The underlying causal graph is given in Figure 6-2. The predictive model $f(O, L)$ is trained

[3] Not to be confused with the 'trust region *method*', which repeatedly solves the trust region *problem*.



(a) **(b)**

Figure 6-3: *The blue line gives the (unobserved) cross-entropy loss under parametric shifts, plotted with respect to the parameter δ_0 (a) and the resulting change in the marginal laboratory testing rate (b). We also provide the quadratic approximation (orange line), estimated using validation data, and the predicted worst-case shift (red star) for $|\delta_0| < 2$ (region in grey).*

on data from $\mathbb{P}$ to predict Y using all available features. If lab tests are not available ($O = 0$), this model predicts Y based on the observed likelihood of Y given $O = 0$, and otherwise uses a logistic regression model trained on cases where $O = 1$ in the training data.

Defining a shift function: $\mathbb{P}(O|Y)$ is a conditional exponential family with $\eta(Y) = \alpha + \beta Y$. We consider the shift function $s(Y; \delta) = \delta_0 + \delta_1 Y$, where δ_0 models an overall change in testing rate, and δ_1 models an additional change in the likelihood of testing sick ($Y = 1$) patients.

Estimating the impact of shift using quadratic approximation: To start, we keep $\delta_1 = 0$ fixed and vary only δ_0, which uniformly increases or decreases testing. In Figure 6-3, we show the ground-truth cross-entropy loss of $f(O, L)$ under perturbed distributions $\mathbb{P}_{\delta_0}$. We observe that the **direction** of the shift matters: In Figure 6-3, the model performance slightly increases under a small increase in testing rates, but degrades if testing increases too much; moreover, the loss under shift is generally asymmetric, as a decrease hurts more than an increase in testing. In Figure 6-3a, we demonstrate the use of the quadratic approximation described in Section 6.3.2. For illustration, we consider a robustness set of $\delta_0 \in [-2, 2]$, and see that the predicted worst-case shift coincides with the actual worst-case shift, and that the quadratic approximation is

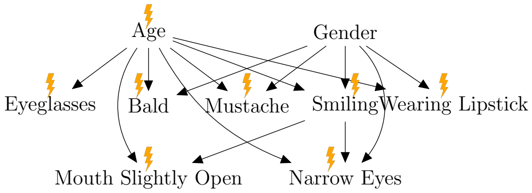


Figure 6-4: *Causal graph over attributes in the synthetic CelebA dataset, where lightning bolts indicate changes in mechanisms. All of these attributes are causal parents of the image X (not shown here), which is generated by a GAN conditioned on these attributes.*

accurate for smaller values of δ.

In Appendix D.4, we allow both δ_0 and δ_1 to vary, and compare our approach to that of worst-case $(1-\alpha)$ conditional subpopulation shifts (Subbaswamy et al., 2021). In the context of this example, we demonstrate that for any $1-\alpha < 0.27$, the worst-case conditional subpopulation loss is achieved by having all healthy patients get tested, and no sick patients get tested. We contrast this with an iterative approach to designing constraints that is made possible by considering parametric shifts, where end-users can restrict the degree to which the shift differs across sick and healthy populations.

6.4.2 Detecting sensitivity to non-causal correlations

A predictive model may pick up on various problematic dependencies in the data that may not remain stable under dataset shift. To understand the impact of these dependencies, a model user may wish to understand which changes in distribution pose the greatest threats to model performance, and to measure the impact of these changes. To illustrate this use-case, we make use of the CelebA dataset (Liu et al., 2015), which contains images of faces and binary attributes (e.g., glasses, beard, etc.) encoding several features whose correlations may be unstable (e.g., the relation between gender and being bald). We consider the task of predicting gender (Y) from images of faces (X), and assess sensitivity to a shift in the distributions of attributes $(\mathbf{W})$.[4]

[4] We do not endorse gender classification as an inherently worthwhile task. Nonetheless, gender classification is commonly studied in the context of understanding the implicit biases of machine learning models (Buolamwini and Gebru, 2018; Schwemmer et al., 2020), and we consider the task with that context in mind.

Setup: To obtain ground-truth shifts in distribution, we generate synthetic datasets of faces using CausalGAN (Kocaoglu et al., 2018), trained on the CelebA data. We simulate attributes following the causal graph in Figure 6-4, and then simulate images from the GAN conditioned on those attributes. We draw a training sample from this distribution $\mathbb{P}$, and fit a gender classifier $f(X)$ using the image data alone, by finetuning a pretrained ResNet50 classifier (Hu et al., 2018). Each attribute W_i is binary, so we consider shifts in the log-odds $\eta_i(Z_i)$ of each attribute W_i given parents Z_i. Here, we use a maximally flexible shift function $s_i(Z_i; \delta_i) = \sum_{z \in \mathcal{Z}_i} \delta_{i,z} \mathbf{1}\{Z_i = z\}$, such that for $Z_i \in \{0,1\}^k$ there are 2^k parameters. Across all intervened variables, $\delta \in \mathbb{R}^{31}$. Due to the synthetic nature of our setup, we can simulate from $\mathbb{P}_\delta(X, \mathbf{W}, Y)$ to evaluate the ground-truth impact of this shift, simulating first from the shifted attribute distribution, and then simulating images from the GAN conditional on those attributes. We use the 0/1 loss $\ell = \mathbf{1}\{f(X) \neq Y\}$, and constrain δ by $\|\delta\|_2 \leq \lambda = 2$.

Comparing importance sampling and Taylor across multiple simulations: We simulate $K = 100$ validation sets from $\mathbb{P}$, in each estimating the worst-case shifts δ_{Taylor} (via the approach in Section 6.3.3) and δ_{IS}, where the latter corresponds to minimizing $\hat{E}_{\delta,\text{IS}}$ using a standard non-convex solver from the `scipy` library (Virtanen et al., 2020). We simulate ground truth data from $\mathbb{P}_{\delta_{\text{IS}}}$ and $\mathbb{P}_{\delta_{\text{Taylor}}}$, to compare the two shifts. First, we demonstrate that the Taylor approach finds more impactful shifts, when searching over the space of small, bounded shifts considered here. In Table 6.1b, we compare the average drop in accuracy using the Taylor shifts (3.8%) and the IS shifts (2.2%). In Figure 6-5b we plot the differences in test accuracy $\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}] - \mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$, where the Taylor approach finds a more impactful shift in 96% of cases. Second, the Taylor approach has an average run-time of $0.01s$, versus $2.14s$ for the IS approach. Third, when **only** used to evaluate the shift δ_{Taylor}, the IS estimator is comparable to the Taylor estimator, with a near-identical average bias (shown in Table 6.1b) and RMSE (0.0191 and 0.0192 respectively). Finally, however, in Table 6.1b we observe that $\hat{E}_{\delta_{\text{IS}},\text{IS}}$ is strongly biased in predicting $\mathbb{E}_{\delta_{\text{IS}}}$, yielding a mean absolute prediction error (MAPE) of 0.069 (not shown in the table). This can be contrasted with a MAPE of 0.015 when using $\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$ to predict $\mathbb{E}_{\delta_{\text{Taylor}}}$.

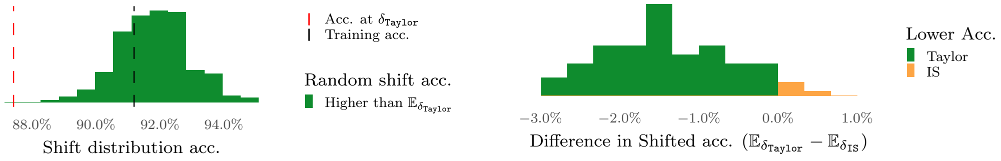
Table 6.1: *(a) Top 5 components (by magnitude) of the example shift vector $\delta \in \mathbb{R}^{31}$ where $\mathbb{P}$ and $\mathbb{P}_\delta$ denote conditional probabilities. The full example shift vector can be found in Appendix D.5.2. (b) Taylor and IS estimates vs. true accuracy for the δ_{Taylor} found by the Taylor approach, and IS estimate vs. true accuracy for the δ_{IS} found by the IS approach. Averages are taken over 100 simulations.*

Conditional		δ_i	$\mathbb{P}$	$\mathbb{P}_\delta$
Bald	| Female, Old	0.899	0.047	0.109
Bald	| Male, Young	-0.800	0.378	0.214
Bald	| Male, Old	-0.680	0.622	0.455
Wearing Lipstick	| Female, Young	-0.618	0.924	0.868
Wearing Lipstick	| Female, Old	-0.543	0.953	0.921

(a)

Metric	Example δ	Avg.
Original acc. $(\mathbb{E}[\mathbf{1}\{f(X) = Y\}])$		0.912
Acc. under Taylor shift $(\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}])$	0.874	0.874
IS est. of acc. under Taylor shift $(\hat{E}_{\delta_{\text{Taylor}},\text{IS}})$	0.829	0.863
Taylor est. of acc. under Taylor shift $(\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}})$	0.844	0.863
Acc. under IS shift $(\mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}])$		0.889
IS est. of acc. under IS shift $(\hat{E}_{\delta_{\text{IS}},\text{IS}})$		0.821

(b)



(a) **(b)**

Figure 6-5: *(a) Model accuracy at randomly drawn shifts. (b) Difference in accuracy in the worst-case shifts identified by Taylor and importance sampling approaches. The Taylor method identifies a more adversarial shift than importance sampling in 96% of simulations (green).*

This may suggest that optimizing the IS objective is prone to "overfitting", choosing a sub-optimal δ from a region of the search space that has high variance. Here, where $\lambda = 2$, the drop in accuracy is relatively mild for the shifts found by both approaches. In Appendix D.5.4 we show that larger values of λ correspond to more substantial drops in accuracy (e.g., an average drop of 23% for $\lambda = 8$ using the Taylor approach).

Examining a single shift: To illustrate the type of shift found by our approach, we consider the δ_{Taylor} (over the K runs) which yields the $\mathbb{P}_\delta$ with median test accuracy. We display the largest components of that δ in Table 6.1a. Among others, this shift entails a 5% increase in the probability of an older woman being bald, and a 5% decrease in the probability of a young woman wearing lipstick. This suggests that the learned classifier f relies on these associations in the images for prediction. We validate that this shift leads to a decrease in accuracy of around 3.8%, using simulated data from $\mathbb{P}_\delta$. To validate that this drop in accuracy is a non-trivial occurrence, we simulate $K = 400$ random shifts δ_k where $\|\delta_k\| = \lambda$ and evaluate the model accuracy in $\mathbb{P}_{\delta_k}$ (Figure 6-5a). As expected, the chosen δ yields a lower accuracy (red line) than all of the random shifts.

6.5 Conclusion

We argue for considering parametric shifts in distribution, to evaluate model performance under a set of changes that are interpretable and controllable. For parametric shifts in conditional exponential family distributions, we derive a local second-order approximation to the loss under shift. This approximation enables the use of efficient optimization algorithms (to find the worst-case shift), and empirically provides realistic estimates of the resulting loss. In a computer vision task, this approach finds more impactful shifts (in far less time) than optimizing a reweighted objective, and the estimates of shifted accuracy under the chosen shift are substantially more reliable.

Of course, our method is not without limitations. Our definition of parametric shifts and resulting approximation relies on the relevant mechanisms $\mathbb{P}(W|Z)$ being a conditional exponential family, and that the relevant variables are observed. As illustrated in our experiments, this can be used to model changes in the causal relationships **between** attributes of an image, but does not immediately extend to modelling changes in the distribution of images given a fixed set of attributes. As with any method that provides worst-case evaluation, there is potential for misuse and false confidence: If the specified shifts fail to capture important real-world changes, the resulting worst-case loss may be overly optimistic and misleading. Even if used correctly, our approach examines a narrow measure of model performance, and a small worst-case error should not be used to claim that a model is free of problematic behavior. For example, implicit dependence on certain attributes (e.g., race in medical imaging (Banerjee et al., 2021)) may be problematic based on ethical grounds, even if it does not lead to major issues with predictive performance under small shifts in distribution.

Chapter 7

Auditing and Prompt Design for Large Language-Image Models

7.1 Background and Motivation

CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) is a self-supervised model of image-text pairs, which has demonstrated remarkable zero-shot performance on a variety of computer vision benchmarks. For instance, it outperforms a fully supervised linear classifier (fit on ResNet-50 features) on ImageNet. To perform zero-shot classification, CLIP performs matching between a set of fixed prompts (short strings of text), and a given image. To perform zero-shot classification, it typically suffices to use prompts of the form "a photo of a **label**." (one for each label in the dataset), and choose the prompt with the highest similarity to the image. We refer to the selection of this set of strings as the problem of "prompt design".

CLIP also appears to generalize well across so-called "natural distribution shifts" represented by a variety of ImageNet-like datasets collected from diverse sources (see Section 3.3 of Radford et al. (2021)). In particular, zero-shot CLIP has much higher robustness to shift than other models, and this robustness decreases markedly for a linear classifier, built on CLIP features, which is trained in a fully supervised fashion on ImageNet. The authors ask

> *. . . is training or adapting to the ImageNet dataset distribution the cause of the observed robustness gap? Intuitively, a zero-shot model should not be able to exploit spurious correlations or patterns that hold only on a specific distribution, since it is not trained on that distribution.*

In this chapter, we further probe the robustness of CLIP to structured shifts in distribution, and explore the impact of prompt design on the robustness of the resulting zero-shot classifier. We do so to illustrate the application of the approach given in Chapter 6 for probing the robustness of models to structured shifts in distribution.

7.2 Defining a set of bounded, structured shifts

In many computer vision tasks, meta-data is available (or is able to be inferred) for images. For instance, the CelebA dataset (Liu et al., 2015) contains 40 attributes (e.g., gender, hair color, makeup worn, etc) for each image, one of which is typically taken as the label. The availability of meta-data allows us to model how the performance of a predictive model changes under structured changes in the data distribution. Here, the goal is to use images X to classify a label Y, where other image attributes Z are available during training.

A plausible change in this case might involve changes in the joint distribution $P(Y, Z)$ of the label Y and attributes Z, while the distribution of images X conditioned on the label / attributes $P(X \mid Y, Z)$ remains fixed. We discuss the plausibility of this assumption later on from the viewpoint of changes in causal mechanisms, and cases where this assumption may be violated. For now, we note that we do not expect models to be robust to arbitrary changes in the joint label/attribute distribution $P(Y, Z)$, but instead seek to understand how models perform under bounded and structured shifts in this joint distribution.

7.2.1 Defining a set of structured shifts

In this section, we consider shifts of the following form,

$$P_\delta(Z, Y, X) = P_\delta(Z, Y) P(X \mid Z, Y), \tag{7.1}$$

where we use red to indicate portions of the factorization which change, and we use P to denote the probability mass function for discrete variables, and the probability density function for continuous variables. Assuming that all attributes (including the label Y) are binary, we denote set of all attributes by

$$A \coloneqq (Y, Z), \tag{7.2}$$

we consider a **structured shift** in the distribution of $P(A)$, which we further factorize as follows, without assuming any conditional independences, and where $A_{1:i-1} \coloneqq (A_1, \dots, A_{i-1})$, and where we adopt the convention that $A_{1:0} = \varnothing$

$$P_\delta(A) = \prod_{i=1}^{d_A} P_\delta(A_i \mid A_{1:i-1}). \tag{7.3}$$

For each conditional distribution $P(A_i \mid A_{1:i})$, we define a set of possible conditional distributions indexed by a parameter $\delta \in \mathbb{R}^{|A|}$, where

$$P_\delta(A_i = 1 \mid A_{1:i}) = \sigma(\eta_i(A_{1:i}) + \delta_i). \tag{7.4}$$

where $\eta_i(A_{1:i-1})$ denotes the conditional log-odds $\log(p/(1-p))$ for $p \coloneqq P(A_i = 1 \mid A_{1:i-1})$. This in turn defines a set of joint distributions

$$P_\delta(A, X) = P_\delta(A) P(X \mid A) \tag{7.5}$$

where $P_\delta(A)$ is equal to the product of the shifted distributions. For a fixed predictive model, we then seek to estimate the worst-case zero-one loss under a set of distributions

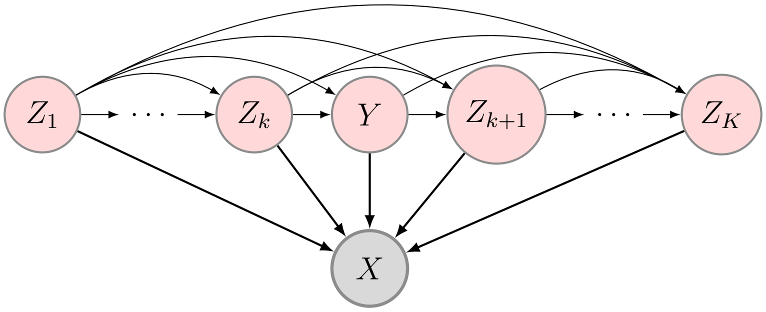


Figure 7-1: *An image X is a function of binary attributes Z and label Y. Some components of Z cause Y, while others are caused by Y, and all of their distributions are subject to change.*

defined by δ, recalling that Y is just one of the attributes

$$\sup_{\|\delta\|_2 \leq \lambda} \mathbb{E}_{X,Y \sim P_\delta}[f(X) \neq Y]. \tag{7.6}$$

Given an ordering over attributes A, each dimension of δ corresponds to a "uniform" shift in the probability of observing each attribute[1], given the attributes which come before it. This factorization is reflected in the directed acyclic graph (DAG) given in Figure 7-1, which we note imposes no statistical restrictions on the original distribution, as it implies no conditional independences between variables.

While our approach can handle more complex shifts (i.e., replacing δ_i with any parametric function $s(A_{1:i}; \delta_i)$), we choose this parametric form of shift because it has a straightforward interpretation: These shifts capture a "monotonic" change in each attribute that otherwise preserves existing correlations. On the appropriate scale, the tendency of every person to (for example) have blond hair decreases by the same amount. However, women are still more likely to have blond hair then men, and so on.

7.2.2 When do these shifts reflect changes in causal mechanisms?

If the DAG in Figure 7-1 reflects the causal data-generating process, then the changes in distribution we describe can be considered changes in causal mechanisms. We

[1] More precisely, an additive change in the conditional log-odds

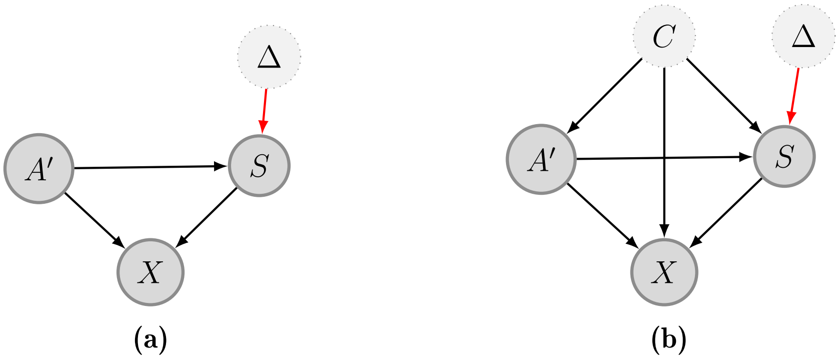


Figure 7-2: *S denotes "smiling", and $A' \coloneqq (Y, Z) \setminus S$ denotes all other attributes, including the label. (a) In this graph, a change in the causal mechanisms of S is reflected in a change in $P(S \mid A')$, but $P(X \mid A', S)$ is unchanged. (b) In this graph, a change in the same causal mechanism (of S) results in a change in $P(X \mid A', S)$. Here, C is a confounder denoting context (e.g., "on the red carpet"). Here, a change in the causal mechanism of smiling implies a change in the conditional distribution of images given all attributes $P(X \mid A', S)$.*

view this causal viewpoint as useful primarily in determining how to determine an appropriate factorization from an interpretability standpoint. For instance, consider the attributes of "Young/Old" and "Has Grey Hair". Intuitively, we might expect a change in age to be reflected by a change in hair color, and we might similarly expect that hair color could change without a corresponding change in age (e.g., due to changes in personal presentation like a change in the use of hair-coloring products). From a causal perspective, hair color is "downstream" of age, where an intervention to one's age could cause a change in hair color, but not vice-versa.

However, the DAG in Figure 7-1 encodes more assumptions than simply the correct causal ordering of variables: it assumes a lack of unobserved variables, which has implications for the impact of changes in causal mechanisms. For instance, the structure of the graph implies that soft interventions on individual attributes would not result in a change to the distribution of images given attributes $P(X \mid A)$. In Figure 7-2 we illustrate that this invariance would not necessarily hold under other causal structures.

In Figure 7-2a, we consider a shift in the causal mechanism of `Smiling` (S), and suppose that `Smiling` is not a causal parent of any other attribute. Let A' denote all other attributes: If these are equal to the causal parents of S in the underlying causal graph, then a shift in causal mechanisms would correspond to an **isolated** change in $P(S \mid A')$, without impacting any other observed conditional distribution. In this sense, we can interpret a change in $P(S \mid A')$ alone as a change in distribution arising from manipulation of this causal mechanism.

In Figure 7-2b we consider an alternative causal data-generating process, where this invariance of $P(X \mid A)$ does not hold. Considering the nature of the CelebA dataset (photos of celebrities), we might imagine that some confounders exist. For instance, consider a latent variable for the "context" in which a photo was taken, e.g., "at a red-carpet screening". If this context variable has a direct effect on the image X, as well as the attribute for smiling S, then a change in the causal mechanism of smiling implies a **non-isolated change** in **both** $P(S \mid A')$ and the conditional distribution of images given attributes $P(X \mid A', S)$.

This can be formalized by placing a new node Δ on the graph, which captures a change in mechanisms, and observing that $X \not\perp\!\!\!\perp \Delta \mid A, S$, though the back-door path $X \leftarrow C \rightarrow S \leftarrow \Delta$, where S serves as a collider. Here we consider some informal intuition for why this occurs: Consider two interventions on S, and consider the distribution of images given $S = 1, A' = a$. Suppose the first intervention corresponds to celebrities becoming more likely to smile at a red-carpet event, all else being equal (e.g., for all values of A'). In the second, celebrities are less likely to smile at a red-carpet event. Under the first intervention, knowing that a celebrity is smiling **increases** the posterior probability of a red-carpet event, which is then reflected in the distribution of images. Under the second intervention, this pattern is reversed, and knowing that a celebrity is smiling **decreases** the posterior probability of a red-carpet event. Here, we note that the problem lies in the fact that the observed attributes may not block the path from C to X.

We discuss the challenge of causal interpretation further in Section 7.5.1, but for now we utilize the parameterization of shift discussed above.

7.3 Comparing zero-shot CLIP against a fine-tuned model

In this section, we begin by probing the robustness of zero-shot CLIP versus a fine-tuned ResNet-50 model. In this section, we use the same synthetic gender classification task as in Section 6.4 where we have ground truth information allowing us to simulate performance under different shifts. For the purpose of this section, we use the same factorization of shift discussed in Section 6.4.

In the following, we use a publicly-available version of the CLIP model (Radford et al., 2021).[2] We first consider a basic prompt for binary classification of "Gender", using the strings (i) "a photo of a man", and (ii) "a photo of a woman", where the zero-shot CLIP model corresponds to picking the string (of these two options) whose embedding is most similar to the image. Following the same experimental setup as in Section 6.4, we repeat the following procedure 10 times:

(i) Generate a synthetic "validation" set of 1000 images, sampled from a fixed distribution $P(X, Y, Z)$, and evaluate the performance of zero-shot CLIP.

(ii) Using the factorization of $P(Y, Z)$ considered in Section 6.4, specify a flexible set of parametric perturbations to each factor[3], and solve for the worst-case shift in a bounded set of perturbations.

(iii) Validate the predicted drop in performance by simulating new data under the chosen shift.

The main results are given in Tables 7.1 and 7.2. In Table 7.1 we compare an illustrative worst-case shift found for the zero-shot CLIP model, as well as the illustrative shift given in Section 6.4 for a fine-tuned ResNet-50 model, trained on a separate training dataset also generated from this simulator. In Table 7.2, we give the training accuracy of each model, along with the average predicted (and actual) drop in performance under shift. From this we make a few observations:

[2] We use the `openai/clip-vit-base-patch32` model available at https://huggingface.co/

[3] Note that in the generative model of Section 6.4, $P(Y)$ does not change, while $P(Z \mid Y)$ changes. We follow the same set of shifts considered in that section.

Table 7.1: *Top 5 components of $\delta \in \mathbb{R}^{31}$ yielding median drop in performance.*

(a) ***Zero-Shot CLIP***

Conditional	δ_i	$\mathbb{P}$	$\mathbb{P}_\delta$
Mustache | Female, Old	1.238	0.076	0.221
Bald | Female, Old	0.972	0.047	0.116
Eyeglasses | Young	0.488	0.401	0.522
Bald | Male, Young	-0.451	0.378	0.279
Eyeglasses | Old	-0.430	0.500	0.394

(b) ***Fine-tuned ResNet-50***

Conditional	δ_i	$\mathbb{P}$	$\mathbb{P}_\delta$
Bald | Female, Old	0.899	0.047	0.109
Bald | Male, Young	-0.800	0.378	0.214
Bald | Male, Old	-0.680	0.622	0.455
Lipstick | Female, Young	-0.618	0.924	0.868
Lipstick | Female, Old	-0.543	0.953	0.921

Table 7.2: *Impact of the example shift, as well as average changes in accuracy over multiple simulations.*

(a) ***Zero-Shot CLIP***

	Example	Avg.
Acc. pre-shift		0.804
Acc. post-shift (Taylor)	0.757	0.757
IS estimate of Taylor Shift	0.780	0.738
Taylor estimate of Taylor Shift	0.762	0.743

(b) ***Fine-tuned ResNet-50***

	Example	Avg.
Acc. pre-shift		0.912
Acc. post-shift (Taylor)	0.874	0.874
IS estimate of Taylor Shift	0.829	0.863
Taylor estimate of Taylor Shift	0.844	0.863

Zero-Shot CLIP has a similar robustness gap to the fine-tuned model: In both cases, model performance drops by around 4% between the training distribution and the shifted distribution. The lower in-distribution performance of zero-shot CLIP does not correlate, in this case, with any measurable difference in robustness.

Zero-Shot CLIP is sensitive to different "directions" of shift: In the synthetic distribution of attributes that make up the training distribution in Section 6.4, there is a strong correlation between gender and wearing lipstick, which is reflected in the shifts that the fine-tuned ResNet-50 model is sensitive to. Zero-shot CLIP shares some common sensitivities to shift (e.g., in the distribution of baldness given age and gender), but the worst-case shift does not appear to be in a direction that emphasizes changes in lipstick wearing.

Conclusions: Based on this cursory evaluation, the zero-shot CLIP model (and prompt) used here does not immediately appear to be more robust than a fine-tuned model in this particular setting. In the next section we consider how robustness might be improved by using a different set of prompts.

7.4 Understanding the impact of prompt design on the robustness of zero-shot CLIP

In this section, we consider evaluation of different zero-shot CLIP models on the original CelebA dataset,[4] using the task of classifying a particular binary attribute (Blond Hair, in this section). Similar to the previous section, we use a publicly-available version of the CLIP model (Radford et al., 2021). We first consider a basic prompt for binary classification of "Blond Hair", using the strings (i) "a photo of a person", and (ii) "a photo of a person with blond hair".

We first allow for shifts in the causal mechanisms of **all attributes, including the label**, as described in Section 7.2. When evaluating the original zero-shot CLIP model using a simple prompt, we find simple shifts in distribution (corresponding primarily to interventions on gender, age, and hair color) that lead to substantial increases in classification error from 22% to 39%, with the latter estimated via importance sampling on the validation set.

We then investigate the use of different prompts (e.g., those that explicitly attempt to disentangle age and gender from hair color), which demonstrate an improvement both on in-distribution performance, and under the particular shift found for the original prompt. However, we find that the worst-case error of these alternative prompts is not generally improved: They are merely susceptible to different directions of shift.

We then demonstrate that the **class of shifts matters**. When we restrict to shifts only in **selected attributes, leaving the marginal distribution of the label unchanged**, for instance, we find that these prompts exhibit more robust performance than the original prompt. Overall, we find that looking at the chosen worst-case shift can give some insight into the vulnerabilities of specific models, and the robustness of different prompts depends on the type of shift that we allow.

[4] As opposed to the synthetic dataset used in the previous section.

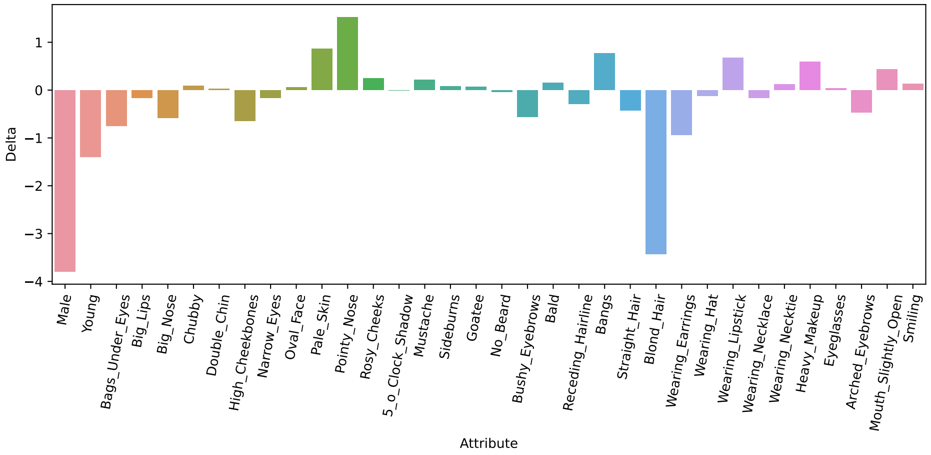


Figure 7-3: *Candidate worst-case δ found for the original prompt in the Blond Hair classification task, where attributes are given in the order used to generate shifts. In Table E.1, we give the precise values, as well as the estimated (via importance sampling) marginal proportions of each attribute in the shifted distribution.*

7.4.1 Finding a worst-case shift

We use the approach we developed in Chapter 6 to find a worst-case shift using the training data, constraining $\|\delta\| \leq 6$, and use importance sampling on the validation set to evaluate the shift. Further details of how we find and evaluate the shift are given in Appendix E.1.

For the original prompt, the worst-case shift involves (among other things) a substantial reduction in the prevalence of men, young persons, and individuals with blond hair. The full shift (in terms of δ) is given in Figure 7-3 and Table E.1, where attributes are ordered in the same way that we construct the shifts. In Table E.1 we also show the resulting (estimated) changes in prevalence for each attribute, which builds some useful additional intuition: As part of a general increase in the prevalence of women, there is also a dramatic decrease in facial hair of all kinds, an increase (from 15% to 40%) in the prevalence of bangs (a hair style), and an increase in lipstick and heavy makeup.

7.4.2 Further interpreting the chosen shift

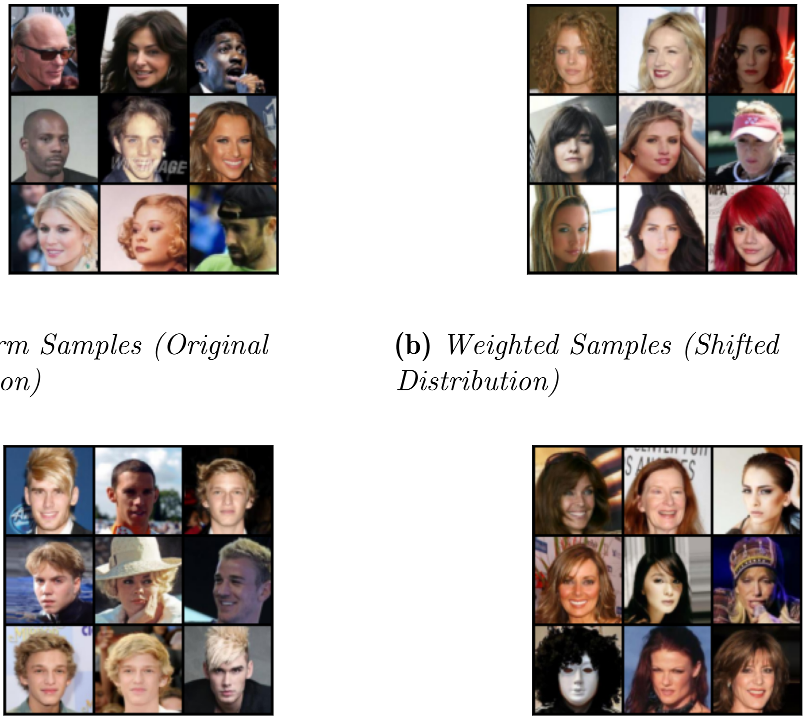
For each conditional distribution, the shift shown in Figure 7-3 is simple to describe on a technical level, e.g., there is a uniform decrease in the log-odds of having blond hair, given all the preceding variables. However, this shift can still feel difficult to interpret, because the shift occurs in multiple conditional distributions, and there are downstream impacts of each change (e.g., a change in proportion of genders implies other changes in the proportion of lipstick, etc). With that in mind, we briefly discuss a few other ways of interpreting the chosen shift.

Highlighting differences via illustrative images In Figure 7-4, we adopt a simple approach to visualizing the new distribution over images, as well as the shift itself, in terms of examples. In Figure 7-4a, we show nine randomly selected images chosen uniformly from the original validation set, and in Figure 7-4b we show nine randomly selected images where we sample in proportion to the importance weights. In Figures 7-4c and 7-4d, we visualize the change by comparing the most down-weighted images (smallest importance weights) with the most up-weighted images (largest importance weights). This primarily illustrates the shift away from young men with blond hair, towards older women without blond hair.

Examining marginal and lower-order changes In an ideal world, we would have simple explanations or descriptions of the shifts we find. Even when presenting this particular shift, one is tempted to reach for a simple description like "more women and less blond hair", even though the shifts we consider are a bit more complex.

Here, we briefly discuss how we might describe shifts in a post-doc way. Notably, this is still non-trivial, even when (a) we have a set of known concepts, and (b) we know the shift only occurs with respect to these concepts. Even if we had a good way of extracting many concepts from images (e.g., using nodes in a neural network that correspond to specific concepts (Bau et al., 2017)), we would have a similar challenge.

Our general approach is to look at lower-dimensional projections of the change, i.e.,



(a) *Uniform Samples (Original Distribution)*

(b) *Weighted Samples (Shifted Distribution)*

(c) *Most Down-Weighted Samples*

(d) *Most Up-Weighted Samples*

Figure 7-4

looking at changes in the prevalence of subgroups defined by one or two variables. This approach is inspired by prior work that studies a similar problem using CelebA (describing hard examples, rather than a new distribution), and which handles the complexity problem implicitly by only considering "simple" low-order descriptions. For instance, Jain et al. (2022) uses CLIP embeddings to find captions (from a pre-specified list) that are the closest match to a set of difficult examples. The resulting descriptions are compellingly simple (e.g., "a photo of a young woman who has an oval face") but this is by construction: All of the captions are single-attribute captions of the form "a photo of a `<adjective>` `<gender>` who has `<attribute>`".

Here we give two examples of how we might do low-dimensional projections, though one might argue that it is easier to understand the shift in the original form that was given. In particular, the original shift is more parsimonious, with four major changes (in gender, age, pointy noses, and blond hair), while looking at the post-doc differences

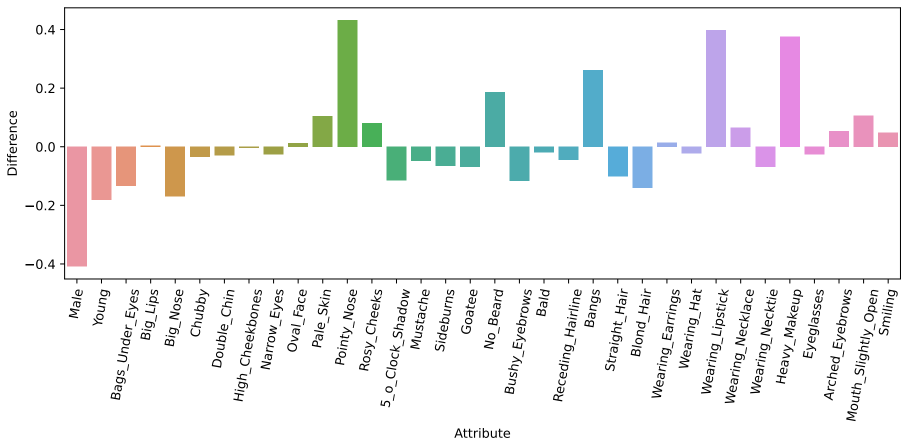


Figure 7-5: *Similar to Figure 7-3, except that we visualize using the difference in the marginal proportion of each binary feature under the new distribution, estimated via importance sampling. We can see, for instance, a large increase in makeup and lipstick, likely as a downstream result of fewer men. In Table E.1, we give the precise values.*

just highlights many of the downstream impacts of those changes.

Examining marginal changes: One approach is to simply look at differences in the marginal proportion of each attribute. In Figure 7-5 and Table E.1 we show the resulting (estimated) changes in prevalence for each attribute, which demonstrates several down-stream effects of the increase in the prevalence of women, including a decrease in facial hair of all kinds, an increase (from 15% to 40%) in the prevalence of bangs (a hair style), and an increase in lipstick and heavy makeup.

Finding low-order interactions with large changes: In Table 7.3, we look at all pairs of binary variables, filtering to those subgroups that have a prevalence of at least 5% in the original distribution, and pull out the subgroups that have the largest absolute increase or decrease in prevalence. As we can observe, most of these reflect the broader increase in the proportion of women, people with pointy noses, and people without blond hair.

Table 7.3: *Changes in prevalence when examining subgroups defined by two binary variables. Loss gives the validation loss on each subgroup.*

(a) *Subgroups with the largest decrease in prevalence.*

Var 1	Var 2	Pre	Post	Diff	Loss
Pointy Nose (No)	Bangs (No)	0.612	0.173	-0.439	0.193
Pointy Nose (No)	Pale Skin (No)	0.683	0.250	-0.434	0.201
Rosy Cheeks (No)	Pointy Nose (No)	0.686	0.271	-0.415	0.204
Male (Yes)	Heavy Makeup (No)	0.425	0.016	-0.409	0.159
Rosy Cheeks (No)	Male (Yes)	0.424	0.016	-0.408	0.160

(b) *Subgroups with the largest increase in prevalence.*

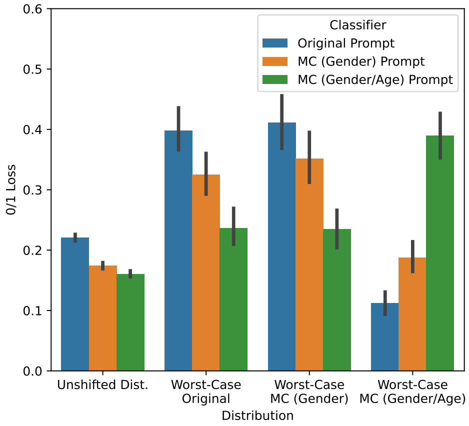
Var 1	Var 2	Pre	Post	Diff	Loss
Heavy Makeup (Yes)	Blond Hair (No)	0.284	0.742	0.458	0.341
Pointy Nose (Yes)	Blond Hair (No)	0.221	0.694	0.473	0.311
Pointy Nose (Yes)	Male (No)	0.217	0.697	0.480	0.288
Wearing Lipstick (Yes)	Blond Hair (No)	0.325	0.818	0.492	0.339
Male (No)	Blond Hair (No)	0.430	0.971	0.541	0.325

7.4.3 Evaluation of alternative prompts

Evaluation of shifted performance and prompt tuning: Based on the discovered shift, we consider alternative prompts, designed to help the model disentangle blond hair from some of the relevant shifted attributes, focusing on age and gender in particular. To this end, we consider two alternative "multi-class" prompts: (i) four strings mapping to two classes: "a photo of a {man/woman}", and "a photo of a {man/woman} with blond hair", and (ii) eight strings mapping to two classes: "a photo of a {younger/older} {man / woman}" and "a photo of a {younger/older} {man / woman} with blond hair". We refer to these prompts as the multi-class (gender) and multi-class (gender/age) prompts, respectively.

We observe that *while the multi-class (gender/age) prompt improves performance on some distributions, the estimated worst-case loss is similar.* In Figure 7-6, we compare the loss of each prompt, both on the validation dataset and on the worst-case shifted distributions for each prompt.[5] The multi-class (gender/age) prompt yields better

[5] Reported performance on all shifted distributions is estimated via importance sampling on the



(a)

	Prompt		
Distribution	Original	MC (G)	MC (G/A)
Unshifted Dist.	0.2206	0.1742	0.1606
Worst-Case [Orig.]	0.3983	0.3253	0.2369
Worst-Case [MC (G)]	0.4116	0.3516	0.2347
Worst-Case [MC (G/A)]	0.1126	0.1879	0.3898

(b)

Figure 7-6: *Comparison of the average loss of each zero-shot CLIP model on the validation dataset, and the estimated loss on the worst-case shift found for each prompt. The worst-case shifts for the original prompt and the MC (Gender/Age) prompt are shown in Figure 7-7. (a) Uncertainty estimates are 95% confidence intervals derived via bootstrapping the validation dataset, with the estimated importance weights fixed. (b) Estimates using the entire validation set.*

performance than the original prompt both on the validation dataset, and under the particular shift which maximizes the loss of the original prompt (*Worst-Case Original*). However, if we consider a worst-case shift for the multi-class (gender/age) prompt itself (*Worst-Case MC (Gender/Age)*), the estimated worst-case loss is similar. A similar pattern occurs for the multi-class (gender) prompt, whose worst-case loss does not appear to significantly differ from that of the original prompt.

To build intuition for the differences in model sensitivity, we can also *compare the estimated worst-case shifts* for the original and multi-class (gender/age) prompts. This is shown in Figure 7-7, where we observe that the major difference is in the prevalence of blond hair: The worst-case shift for the original prompt involves a substantial decrease in blond hair, while the worst-case shift for the multi-class (gender/age) prompt involves a substantial increase in blond hair.

validation dataset.

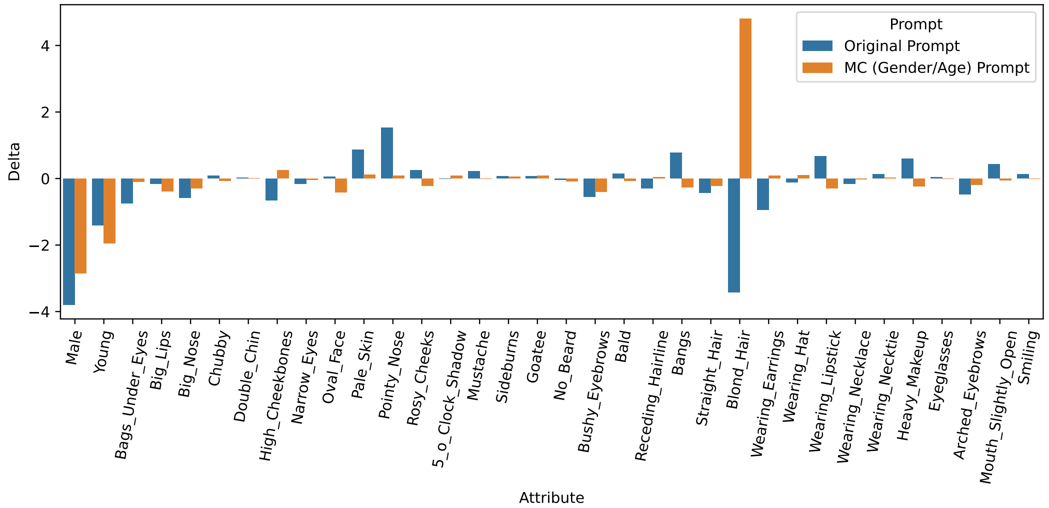
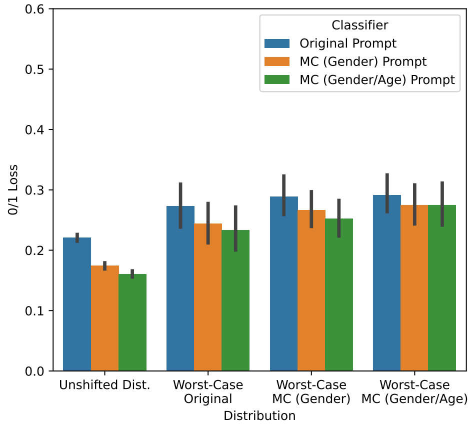


Figure 7-7: *Comparison between the δ found for the original prompt, and for the multi-class prompt including gender and age.*



(a)

	Prompt		
Distribution	Original	MC (G)	MC (G/A)
Unshifted Dist.	0.2206	0.1742	0.1606
Worst-Case [Orig.]	0.2732	0.2440	0.2329
Worst-Case [MC (G)]	0.2885	0.2666	0.2526
Worst-Case [MC (G/A)]	0.2910	0.2747	0.2745

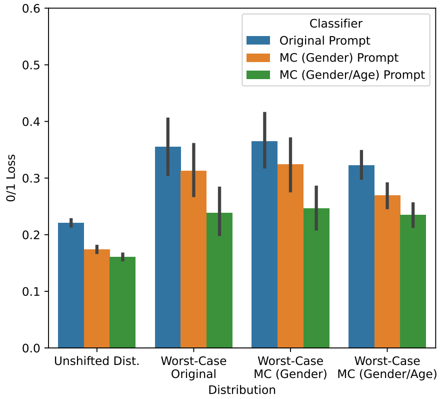
(b)

Figure 7-8: *Blond Hair Classification under restricted shifts: Similar to Figure 7-6, except where we restrict the type of loss that we consider to avoid shifts in hair-related features.*

7.4.4 Evaluation of alternative prompts under restricted shifts

Here, we consider a restricted class of shifts, where we do not permit shifts in certain hair-related features (e.g., a direct intervention on blond hair is not allowed).

Shifts that do not change the causal mechanisms of the label In Figure 7-8, we give the results analogous to Figure 7-6, except where we restrict the shifts that we consider. In particular, we do not allow for shifts in hair-related features. We show



(a)

	Prompt		
Distribution	Original	MC (G)	MC (G/A)
Unshifted Dist.	0.2206	0.1742	0.1606
Worst-Case [Orig.]	0.3553	0.3126	0.2385
Worst-Case [MC (G)]	0.3649	0.3237	0.2462
Worst-Case [MC (G/A)]	0.3221	0.2694	0.2344

(b)

Figure 7-9: *Blond Hair Classification under restricted shifts: Similar to Figure 7-6, except where we use a different causal ordering where hair-related features come first, and are not subject to shift*

the corresponding shift (for the original prompt) in Table E.2. Table E.4 gives the full order of the factorization we consider, along with a specification of which attributes are allowed to shift. For all other attributes, we fix $\delta_i = 0$. Here, the magnitude of the worst-case loss decreases substantially (from around 40% to around 27%). However, the same general conclusions hold, that the multi-class (gender/age) prompt has similar worst-case performance to the original prompt, despite having substantially better in-distribution performance.

Shifts that do not change the marginal distribution of the label In Figure 7-9, we consider a different causal order (shown in Table E.3) where hair color (including the label) comes first, and is not directly intervened upon. In effect, we consider interventions of the form $P(A \setminus H \mid H)P(H)$ where H corresponds to the attributes describing hair color, and where $P(H)$ remains fixed while $P(A \setminus H \mid H)$ is allowed to change. In this case, we do see more reliably robust performance from the multi-class prompt, whose worst-case miss-classification error is around 24%, compared to a worst-case loss of around 36% for the original prompt.

7.5 Challenges and open directions

Unlike the other chapters in this thesis, the work presented here consists primarily of exploratory work, working through the challenges that can arise when applying the method developed in Chapter 6 to the problem of model design (in this case, prompt design) on a real imaging dataset.

With that in mind, we discuss some of those challenges here, and potential solutions. The first challenge is the relative complexity of interpreting the shifts themselves, particularly in the absence of a clear causal structure over all variables. The second is the correct performance metric: While worst-case performance is intuitive in some ways, it is lacking in other ways, particularly when the magnitude of the change is not clear a-priori.

7.5.1 Generating simpler, easier to interpret shifts

The first challenge is the relative complexity of the shifts that we consider. Here we discuss two near-term ways of reducing the complexity, while still maintaining some degree of flexibility and interpretability. These include (i) Finding shifts on fewer attributes (e.g., single attribute interventions) (ii) Finding shifts with fewer degrees of freedom (e.g., hard interventions instead of soft interventions)

Searching over single-target hard interventions: In the medical setting, one could imagine the utility of understanding performance across simple scenarios like “if all patients were tested on scanner brand X” or “if all patients received this lab test”. These scenarios are less expressive shifts than the ones we currently consider, but may be a more natural starting point for interpretable, causality-motivated shifts. For a given causal graph, and discrete variables, this would yield a straightforward brute-force approach to searching for single-target interventions: For each variable, simply estimate the causal effect (on the loss) of the each possible interventions (one for each value of the variable), and then report the most impactful interventions.

The downside of this approach is the potential for inadequate overlap: To guard against this on a first pass, one could assess the effective sample size for each intervention[6], or assess overlap via propensity scores, and eliminate from consideration those interventions that have insufficient overlap.

Relaxing the assumption of no unmeasured variables, and pre-specifying a causal ordering: To avoid making claims about the causal ordering of individual attributes (except in relation to Y), one could make an alternative assumption that would still allow for estimating the impact of hard interventions. An illustrative causal structure is given in Figure 7-10, where each attribute is correlated due to an unmeasured confounder, but this confounder does **not** have any direct impact on the image X. In this case, a hard intervention on any attribute Z_i can be estimated by using the remaining attributes (and Y) to block all backdoor paths to the loss.

Moreover, under this graph, the average loss under hard interventions on any given Z_i should be equivalent to the **counterfactual loss** if we performed full counterfactual inference on each image to infer what it would have looked like under intervention on Z_i. This could be viewed as a computationally cheaper way (relative to using a generative model that can generate counterfactual images) to seek out potential counterfactual interventions where the model would perform poorly.

Sparse (soft) interventions using our existing approach: Finally, one could also attempt to find shift vectors in a more computationally efficient way, which only impact a smaller set of variables. Currently we find shifts using an ℓ_2 constraint on δ, but we could also have considered an ℓ_1 constraint to encourage sparsity. This does not take advantage of the Taylor approximation approach, but could be easily implemented using a generic non-convex solver, optimizing over the importance weights.

[6] The effective sample size (ESS) is often used in offline RL to assess our ability to assess a policy, see http://www.nowozin.net/sebastian/blog/effective-sample-size-in-importance-sampling.html for some details of ESS calculation.

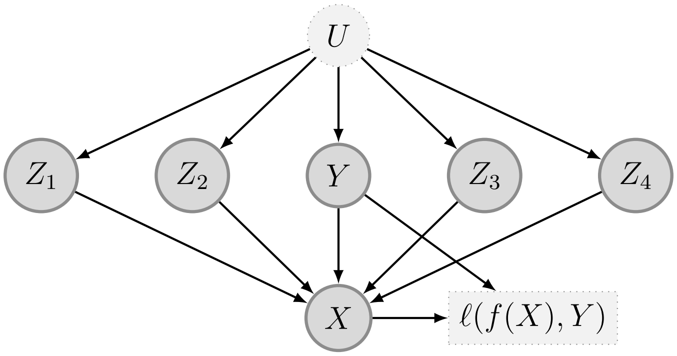


Figure 7-10: *An illustrative graph where the impact (on the loss) of hard interventions on each Z_i can be estimated, even though we cannot apply our existing approach due to the unmeasured confounder U. Note that this requires that U has no direct effect on X, which is a limitation.*

7.5.2 Choosing the right performance measure

In Example 6.1 (the lab testing example, also discussed in Chapter 6), the causal framing of shifts and use of worst-case loss have a clear justification. We care about the **worst-case loss** because we think it plausible that our future data will be drawn from **any of the distributions in the robustness set**. In this context, the object of interest is as much the estimated worst-case loss itself, as the shift that we find.

However, this motivation is harder to transport to the CelebA task, where (1) we have less domain knowledge about the size and type of shifts that are plausible, which motivates (in other work, e.g., Makar et al. (2022)) the assumption that certain distributions like $P(Z \mid Y)$ could change arbitrarily, and (2) given the presence of arbitrary shifts, the actual worst-case loss seems less useful as a metric, since we don't actually expect to see a true worst-case change.

Once the worst-case loss loses its meaning as "the loss under a plausible future distribution", the goal behind finding and interpreting adversarial shifts becomes more nebulous. One might reasonably ask the question

What "should" the performance of our model look like, under these shifts?
How should these results impact the way we build or decide to use models?

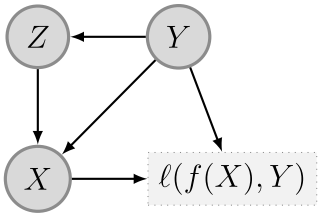


Figure 7-11: *We do not necessarily expect that models can have invariant performance under interventions on Z, even in this anti-causal setting: Interventions on Z do not influence Y, but may influence X. In general, there may not exist a model with invariant performance across interventions on Z. Indeed, some values of Z could make the learning problem more difficult, if e.g., Z encodes some level of noise / blur in the image.*

As we illustrate in Figure 7-11, we should not necessarily expect models to have invariant performance, even when the shifts occur in attributes that do not influence Y.

Distribution-specific regret as an alternative performance metric: An alternative formulation would be to look at regret, defined as

$$\sup_{P \in \mathcal{P}} \{\mathbb{E}_P[\ell(f(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_P[\ell(f(X), Y)]\} \tag{7.7}$$

where we compare our loss against the best predictor in hindsight for each distribution P. Supposing we could compute this, it would tell us "the biggest gap between your model and a fine-tuned model is XX%". This perspective has two benefits: First, it helps us understand how much of the excess loss is due to intrinsic difficulty, versus something that we could adapt to in principle. Second, it is perhaps useful from an adaptation perspective, if low regret is taken to imply that we could adapt quickly to find the optimal model.

However, this perspective is not without its own drawbacks. For instance, while comparing f and the optimal model in hindsight f^*, we must note that any environment-specific f^* is unlikely to be a plausible choice of predictor in the first place (before the shift is observed). To illustrate, consider the illustrative example below, where there are two environments in the uncertainty set.

$$P_1(X = 1) = 0.5 \qquad\qquad P_2(X = 1) = 0.5$$
$$P_1(Y = 1 \mid X = 1) = 1 \qquad\qquad P_2(Y = 1 \mid X = 1) = 0$$

Here, X is always a perfect predictor of Y, and the optimal predictor in P_1 is $f_1^*(X) = 1$, and in P_2 the optimal predictor is $f_0^*(X) = 0$, as the correlation is reversed. A model that ignores X has invariant $0/1$ loss of 0.5, but if we choose either f_1 or f_2 as a predictor, they will always be incorrect in the opposing environment.

Given a fixed shift P', one could nonetheless seek to use the idea of regret to characterize what makes the distribution P' difficult. This would also be more tractable to compute. That is, for a fixed P', compute

$$\mathbb{E}_{P'}[\ell(f(X), Y) - \ell(f^*(X), Y)] \tag{7.8}$$

for a fixed model $f(X)$, and a model f^* learned on P'. One way to train f^* is to draw samples from P' in proportion to the importance weights.

Min-max regret as an alternative performance metric: A different alternative would be to ask about the regret versus the worst-case optimal model, which would tell us about the excess worst-case risk we are taking on in exchange for in-distribution performance. This formulation would look something like

$$\sup_{P \in \mathcal{P}} \{\mathbb{E}_P[\ell(f(X), Y)] - \mathbb{E}_P[\ell(f^*(X), Y)]\} \tag{7.9}$$

where $f^* = \arg\min_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(f(X), Y)]$. Of course, actually computing this would require solving the problem of **learning** such a predictor in the first place. For some uncertainty sets, this may actually be plausible (like the worst-case subpopulations over X, Y considered in joint DRO, or the worst-case subpopulations over X considered in marginal DRO (Duchi and Namkoong, 2021; Duchi et al., 2020a)).

Comparing model weaknesses: An alternative goal is to focus on understanding shifts that yield the largest difference in performance between two models. For instance, we could straightforwardly optimize

$$\sup_{P \in \mathcal{P}} \{\mathbb{E}_P[\ell(f_1(X), Y) - \ell(f_2(X), Y)]\} \tag{7.10}$$

to find scenarios where f_1 performs poorly relative to f_2, and vice versa.

Chapter 8

Conclusion

In this thesis, we have presented partial progress towards answering the question

> *How do we make machine learning as rigorously tested and reliable as any medication or diagnostic test?*

Taking a broader view of this question, we consider the drug development process, shown schematically in the top of Figure 8-1. Given a promising candidate drug, an exhaustive process of validation and iteration is required before approval and widespread use. Pre-clinical testing occurs before drugs are ever used in humans, including assessments of potential toxicity and testing in animals. Clinical trials are conducted first in smaller populations to establish reasonable dosages, before going to larger populations to demonstrate effectiveness. Even after a drug is approved, post-marketing surveillance is conducted to monitor for adverse events, long-term negative effects of a drug, and so on.

Machine learning models are not drugs — but we can still learn from the process of drug development. Showing initial promise for a candidate model (e.g., high predictive accuracy on historical datasets) is only the first step. On the other hand, rushing to deploy models without sufficient oversight can cause more harm than good. Mitigating the risk of faulty and biased predictions, while realizing the potential of machine learning in health, requires us to consider the entire process of development.

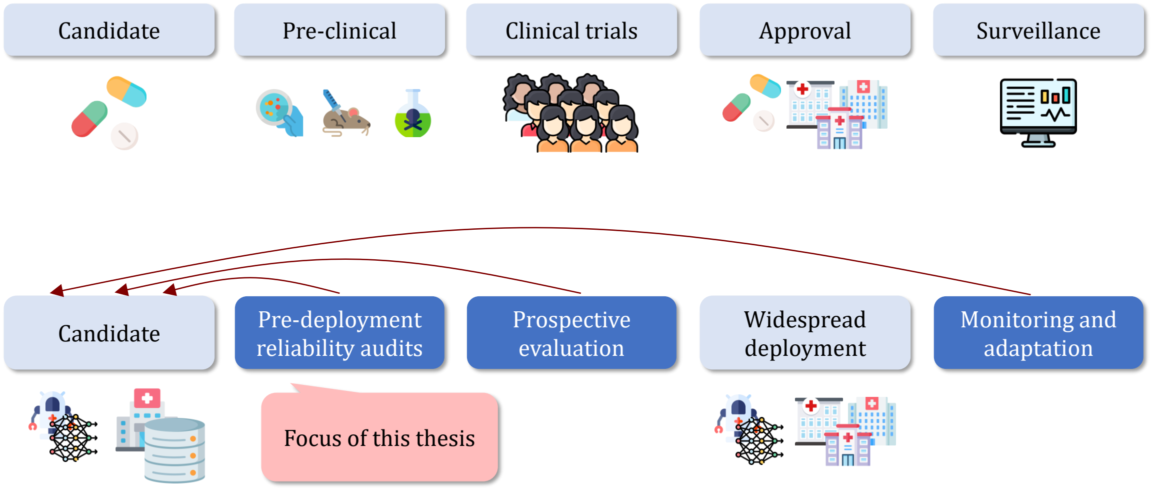


Figure 8-1: *Comparing the machine learning development process (bottom) to the drug development process (top). The focus of this thesis has been primarily on the "pre-clinical" stage of assessing and improving the reliability of models prior to deployment.*

This thesis has focused largely on assessing and improving the reliability of predictions prior to deployment, analogous to pre-clinical testing in drug development. However, deploying machine learning with confidence requires better tools for all parts of the model deployment process. How should we design clinical trials for machine learning models, taking into account the fact that models only impact patients via human decision-makers, and will need to evolve over time? How do we appropriately monitor the impact of deployed systems, not only in terms of predictive accuracy, but their impact on patients and providers? How should we adapt models to new scenarios, without compromising on safety guarantees? In keeping with the core themes of this thesis, doing all of the above will require us to think carefully about the context in which models are deployed, the process by which data is generated in healthcare, and the careful balance between optimism and pessimism.

Ultimately, our goal is to see machine learning become an "unremarkable" part of the standard of care, as reliable and trustworthy as any drug or diagnostic test we use today. Our hope is that the research presented in this thesis provides a small step towards bringing that future to pass.

Part III

Appendix

Appendix A

Appendix for Chapter 3

The supplement is structured as follows:

- **Guidance on hyperparameter selection**: We take a deeper dive into the impact of hyperparameter selection on support and overlap estimation, including an in-depth empirical evaluation with concrete recommendations on how to set hyperparameters for support estimation given an a-priori belief that higher-order intersections of variables may be excluded from the cohort.

- **Application to Policy Evaluation**: We discuss in more depth how the OverRule algorithm can be applied to finding areas of sufficient coverage for policy evaluation tasks.

- **Additional experimental results**: In addition to providing additional detail on the experiments presented in the corresponding chapter of this thesis, we also present several results that were only alluded to in the corresponding chapter of this thesis. This includes the detailed results for the policy evaluation task (antibiotic prescription), as well as additional rules learned for the opioids prescription task.

- **Theoretical results**: We include proofs for our theoretical results, as well as an additional Theorem bounding the generalization error of our two-stage estimator in terms of the error of the base estimators.

In addition, to build further intuition for Boolean rules, we illustrate a Boolean rule in the DNF form in a 2D example in Figure A-1.

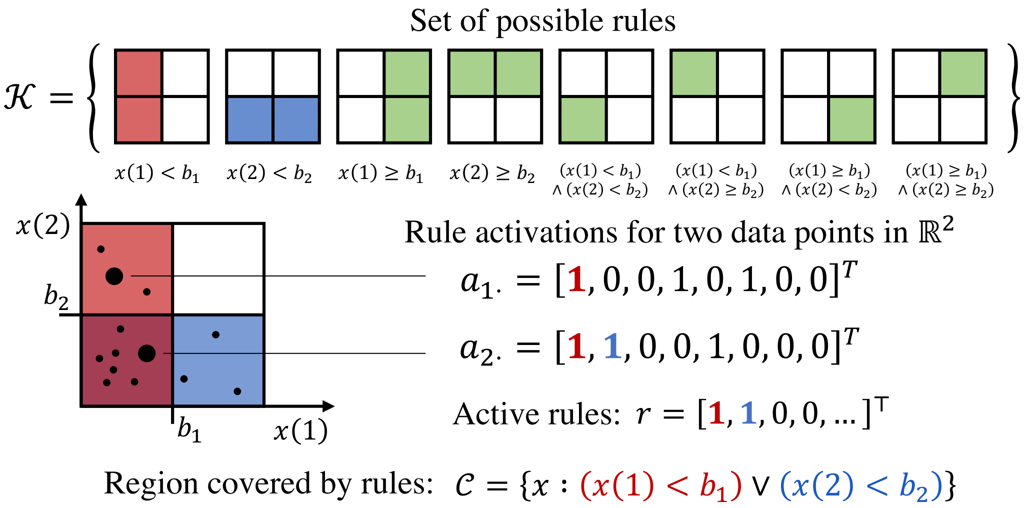

Set of possible rules
$\mathcal{K} =$
$x(1) < b_1$
$x(2) < b_2$
$x(1) \geq b_1$
$x(2) \geq b_2$
$(x(1) < b_1) \wedge (x(2) < b_2)$
$(x(1) < b_1) \wedge (x(2) \geq b_2)$
$(x(1) \geq b_1) \wedge (x(2) < b_2)$
$(x(1) \geq b_1) \wedge (x(2) \geq b_2)$
$x(2)$
b_2
b_1
$x(1)$
Rule activations for two data points in $\mathbb{R}^2$
$a_{1\cdot} = [1, 0, 0, 1, 0, 1, 0, 0]^T$
$a_{2\cdot} = [1, 1, 0, 0, 1, 0, 0, 0]^T$
Active rules: $r = [1, 1, 0, 0, \ldots]^\top$
Region covered by rules: $\mathcal{C} = \{x : (x(1) < b_1) \vee (x(2) < b_2)\}$

Figure A-1: *Boolean rules on disjunctive normal form (DNF). We highlight data points represented by their activations, $a_{1\cdot}, a_{2\cdot}$ of rules from the set $\mathcal{K}$ of all possible rules. $\mathcal{C}$ is the region described by the rule set and r indicators for the rules.*

Code for this chapter can be found at https://github.com/clinicalml/overlap-code

A.1 Choosing Hyperparameters

A.1.1 Overview

Considering OverRule along with the base estimator, there are a few distinct sets of hyperparameters to choose

- **Support Rules**: The support rule estimation task requires a specification of DNF versus CNF form, a specification of $\alpha, \lambda_0, \lambda_1$ used in the objective, and the number of samples to draw from the reference measure.

- **Base Estimator and Overlap Labels**: In addition to the hyperparameters of the base estimator itself, a threshold ϵ must be chosen to generate overlap labels

- **Overlap Rules**: These rules similarly require a specification of DNF or CNF form, and specification of $\beta, \lambda_0, \lambda_1$.

For the base estimator itself, the hyperparameters can be tuned in the usual way using cross-validation using a metric of interest (e.g., AUC). The choice of ϵ is studied in the existing literature (Crump et al., 2009) and ultimately depends on the downstream causal inference task, though $\epsilon = 0.1$ is sometimes considered as a rule of thumb. For the support rules, we typically set the number of reference measure samples to be as large as computationally feasible.

For the overlap and support rules, the remaining hyperparameters can be chosen (1) by using cross-validation to optimize for balanced accuracy (or some other metric, like false positive rates) with respect to the overlap labels or uniform background samples, (2) with some other objective in mind, e.g., setting the λ parameters to be large to discourage many rules, even if more rules would increase accuracy, or (3) with the goal in mind of choosing values (or exploring a range of values) most likely to discover "interesting patterns" in the cohort.

We expand upon a concrete instance of this latter goal in the remainder of this section, particularly as regards hyperparameter selection for support estimation, where extremely high accuracy is particularly easy to achieve and is thus less informative for the purposes of hyperparameter selection.

A.1.2 Choosing Support Hyperparameters to highlight exclusions

Motivation: In the context of our motivating applications, the primary purpose of support estimation is to identify regions where we do not have any (or have very few) observations. For instance, if there are no men in our dataset who also have cardiac arrhythmia[1], then this would be a clinically relevant fact that should be highlighted. Thus, we would like to select hyperparameters which minimize our risk of overlooking these types of exclusions.

[1] This would be surprising, as men with arrhythmia are fairly common in the general population

In this section we give some guidance on how to select hyper-parameters for support estimation with this particular goal in mind, based on synthetic and real-data experiments. To recap, these hyper-parameters include (i) α, the support level, and (ii) λ_0, λ_1, regularization parameters for learning support rules. There are also relevant hyperparameters in the underlying algorithm of Wei et al. (2019), primarily the width of the beam search used during column generation.

Summary: For this purpose, we recommend setting $\alpha \approx 1$, and in particular we consistently observed best results for $\alpha \geq 0.98$. We observe that for α sufficiently close to 1, the results are less sensitive to different values of λ_0, λ_1. In addition, we recommend setting the width of the beam search in the algorithm of Wei et al. (2019) to be on the same order of magnitude as the number of binary features.

These recommendations have the effect of encouraging the algorithm to consider higher-order interactions between variables that describe regions with little or no support in the data (e.g., "there are no men with cardiac arrhythmia"), and we verify this through experiments where we selectively remove regions of the data, and verify whether or not the algorithm can recover these regions.

Concretely, we use both a synthetic and semi-synthetic case where we manually exclude all points which satisfy a simple boolean rule, and look to identify that exclusion automatically. That is to say, in both cases we take a dataset and **remove** data points $\mathbf{x} \in \{0, 1\}^d$ which satisfy a rule of the form $x_i = 1 \wedge x_j = 1$ for two features x_i, x_j, and then check if our algorithm incorporates this into the learned rule set.

- **Synthetic Case:** In this setting, we generate data comprised of 22 independent binary features, such that 10 features are rare (binomial with $p = 0.01$), 12 features are common ($p = 0.5$), and we remove all data points which satisfy a conjunction of the last two common features.

- **Semi-Synthetic Case (Antibiotic Prescription):** In this setting, we used the medical records dataset described in Section 3.5.4, and removed all men with cardiac arrhythmia, which compromised 5% of the total population.

This particular type of exclusion benefits from a CNF formulation (AND of ORs) of the support task. This is because the exclusion can be described in a parsimonious way (independently of other aspects of support) as a single additional rule. As discussed in Section 3.4.1, it is straightforward to convert the CNF formulation to a DNF formulation and vice versa. However, we note that the CNF formulation (for a fixed number of reference samples) can be more computationally intensive than the DNF formulation.

Synthetic Experiments

For the synthetic case, our goal is to build intuition that we can validate in the semi-synthetic setting. We will first describe our data-generating process in more detail, and then describe the results and conclusions from an exhaustive hyperparameter search.

Synthetic Data Generation: We generate data as follows. Note that we are only concerned (for the moment) with estimating support, so we do not include any notion of treatment groups.

- We sample 10,000 data points $x \in \{0,1\}^d$ where $d = 22$, by sampling (for each data point):
  - 10 "rare" binary features $r_1, \ldots, r_{10}$, generated independently with $p = 0.01$
  - 12 "common" binary features $c_1, \ldots, c_{12}$, generated independently with $p = 0.01$
  - Thus, each data point is given by $\mathbf{x} = [r_1, \ldots, r_{10}, c_1, \ldots, c_{12}]$
- We remove all data points which satisfy $c_{11} = 1 \wedge c_{12} = 1$, which is approximately 25% of all data points. Our goal is to recover the corresponding **inclusion rule** as part of the final rule set of $c_{11} = 0 \vee c_{12} = 0$.

Hyperparameter Search & Outcomes: With this setup, we estimate support using the algorithm given in the corresponding chapter of this thesis, using every combination of the following hyperparameters

- $\alpha \in \{0.95, 0.96, 0.97, 0.98, 0.99\}$, the constraint on covering our data.
- $\lambda_0 \in \{0, 10^{-6}, 10^{-4}, 10^{-2}\}$, and $\lambda_1 \in \{10^{-6}, 10^{-4}, 10^{-2}\}$, the regularization terms.
- $B \in \{10, 15, 20, 25, 30\}$, the width of the beam search used in Wei et al. (2019)

For each combination of hyperparameters, we run the experiment three times, generating a new set of fake data with each run. The same three random seeds are used across all hyperparameter combinations. We recorded a number of relevant outcomes, including

- Does the final rule set include the inclusion rule $c_{11} = 0 \vee c_{12} = 0$?
- How many rules are considered in the final rule set, and how long (on average) are these rules?
- How many "perfect" rules are found, which exclude none of the generated data points?

Observations: The full results of the hyperparameter search are given in Table A.5, but we summarize our observations and recommendations here.

- *Recovery by LP → recovery by rounded rules:* Across all hyperparameter settings, if the desired inclusion rule was found during column generation (and thus considered by the LP), it was uniformly included in the final rounded rule.[2] Thus, our goal is to ensure that the desired inclusion is picked up by the LP during column generation.
- *Beam Search Width should be higher than # features:* Recall that the LP relaxation with column generation starts by considering only rules with a single

[2] This is not a general rule; While it holds in the synthetic case, it will not hold exactly in the semi-synthetic case with real data, as demonstrated in the next section.

literal, and beam search is used to select additional rules for consideration, with a maximum width of B. If B is lower than the number of rare features, then the first B rules considered will tend to be rules on single rare features. This prevents the beam search from exploring interactions between more prevalent features. Setting the beam-search width to a sufficiently high number ($\approx$ total features) forces the column generation to explore all rules with two literals, helpful for recovery of our desired inclusion rule. This is demonstrated in Table A.1.

- *Higher values of α produce more stable results across λ.* Higher values of α tends to render the results less sensitive to choice of regularization λ, and tends to produce more reliable results in terms of recovery of our desired rule. As demonstrated in Tables (A.2a-A.2c), lower values of α are more sensitive to λ_1 in terms of both recovering the desired exclusion, as well as the number of rules found. At higher values of α, there is more consistent recovery of "perfect" rules, which exclude none of the sample points (and hence do not contribute to the constraint).

Table A.1: *Beam Search Width and proportion of runs (across all other hyperparameter settings of $\alpha, \lambda_0, \lambda_1$) in which the synthetic region was correctly identified by the final rule set ("Rounded"). Once the beam search width is sufficiently high (larger than the number of rare features), further increasing it does not appear to help.*

Beam Width	10	15	20	25	30
Recovered	0.07	0.87	0.87	0.87	0.87

Discussion / Intuition: Due to the greedy nature of the column generation procedure, a common failure mode is to only consider rules that include rare features, because those singleton rules exclude a significant amount of reference measure, and excluding rare features does not violate the α-constraint. For instance, a support rule of the form "not one of these K rare features" will (roughly speaking) exclude K percent of the samples (if each rare feature has 1% prevalence), while producing a volume of 2^{-K}. Thus, an overly greedy approach can obtain an objective value that is exponentially small in the number of rare features excluded, as long as it does not hit the α constraint. This has the effect of "crowding out" more complex rules.

(a) *Recovery of inclusion rule*

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	1.0	1.0	0
$\alpha = 0.96$	1.0	1.0	0
$\alpha = 0.97$	1.0	1.0	1.0
$\alpha = 0.98$	1.0	1.0	1.0
$\alpha = 0.99$	1.0	1.0	1.0

(b) *Avg. # of rules*

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	23.67	15.75	5.0
$\alpha = 0.96$	35.58	33.33	4.0
$\alpha = 0.97$	39.83	31.92	4.0
$\alpha = 0.98$	44.17	47.17	23.83
$\alpha = 0.99$	31.42	31.25	27.67

(c) *Avg. # of Perfect Rules*

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	12.5	9.25	0.0
$\alpha = 0.96$	20.75	18.67	0.0
$\alpha = 0.97$	24.67	24.92	1.0
$\alpha = 0.98$	30.17	28.33	14.0
$\alpha = 0.99$	23.0	24.08	20.42

Table A.2: *Value of α parameter and λ_1 parameters, for a fixed beam search width ($B = 15$), along with (a) the proportion of runs (across all other hyperparameter settings) in which the synthetic region was correctly identified by the final rule set, (b) the number of rules in the final solution, and (c) the number of perfect rules, defined as those which exclude none of the samples but which exclude some number of reference points. Note that these results marginalize over λ_0, and (b-c) are averaged across all runs.*

Take a concrete example in Table A.2b to build intuition for how the greedy set covering algorithm can fail in this case: Suppose $\lambda_0 = 0$, $\lambda_1 = 0.01$, and $\alpha = 0.95$, and suppose that our current solution excludes 5 rare features before hitting the α constraint, then the reference volume is given by $2^{-5} \approx 0.03$. In this case, adding the desired inclusion rule will reduce the volume by $1/4$ (a reduction in absolute terms which is < 0.01) while increasing the regularization penalty by 0.02. Thus, it will not be included.

To avoid this failure mode, we can increase α, which has the effect of reducing the number of singleton rules K that can be added before violating the constraint.

Semi-Synthetic Experiments

In the semi-synthetic experiment, our goal is to verify that the intuition from the synthetic setting carries over to a real dataset.

Semi-Synthetic Data Generation: We generated the dataset for this experiment as follows.

1. *Subsampling:* We randomly sample 5000 patients from the full cohort of 65k patients, due to computational constraints. In this subset, there were 185 binary features, and 5 continuous features.

2. *Synthetic Exclusion:* We remove all male patients with cardiac arrhythmia, which was around 5% of the total population.

3. *Pre-Processing:* Given the prevalence of very rare binary features, we removed all binary features with a prevalence of less than 1%, as well as all samples that had any of these features, resulting in the removal of 118 binary features and 850 samples. This was done both for computational reasons (to reduce the number of features) as well as to condition the problem such that it is more realistic for the support estimation to recover higher-order interactions.

4. *Final Dataframe*: The final dataset had 66 binary features and 5 continuous features, with the latter being converted into binary features via the use of deciles.

Hyperparameter Search: We then followed a similar approach to the synthetic experiment, using every combination of the following hyperparameters. For each combination, we ran the algorithm three times, inducing randomness over the data by taking a random 80% of the data with each iteration.

- $\alpha \in \{0.95, 0.96, 0.97, 0.98, 0.99\}$

- $\lambda_0 \in \{10^{-6}, 10^{-4}, 10^{-2}\}, \lambda_1 \in \{10^{-6}, 10^{-4}, 10^{-2}\}$

In this case, we fixed the width of the beam search at $B = 1000$ (which encourages a more thorough search during column generation, as discussed above), and also found that we needed to adjust the value of K, another hyperparameter from the column generation algorithm, to be roughly on the same order as B. The parameter K controls how many rules get added to the LP at each iteration. We also fixed the maximum number of iterations at 10. We recorded all the same outcomes as were used in the synthetic case.

Observations: We observed that a number of patterns from the synthetic case carried over to the semi-synthetic case.

- *Inclusion in LP (mostly) implies inclusion in final rules:* When the desired inclusion rule appears among the rules considered during column generation, it mostly appears in the final rounded rules, in 80% of runs. We conjecture that this is due to a large number of “perfect” rules existing in this dataset, which are also two-variable interactions, though many of these appear to be noise (see example inclusion rules below).

- *Increasing α leads to more consistent recovery in the LP* of the desired inclusion rule. However, as discussed, this does not always translate into the desired inclusion rule showing up in the final rounded rule set. See Table A.3

- *Higher values of α are less sensitive to choice of λ*: In Tables (A.4a-A.4b) we demonstrate that, similar to the synthetic case, the number of rules and the number of “perfect” rules is highly sensitive to λ_1 when α is lower, but for $\alpha \geq 0.98$ it yields consistent results across different values of λ.

Example “Perfect” Rules: These rules exclude none of the samples in our data, while excluding reference points. While occasional rules appear to be based on reasonable exclusions (such as a lack of pregnant veterans, given that 80% of veterans are male in our data), most appear to be combinations of rare features (such as rare medications) that simply do not appear together in our data. These are three representative rules from one run (where $\alpha = 0.99, \lambda_0 = \lambda_1 = 1e - 6$, resulting in 23 rules, of which 17 were "perfect"):

- not (Pregnant and Veteran)
- not (Complicated Hypertension and Previous Medication of Cephalexin)
- not (Previous Medication of Doxycycline and Norfloxacin)

Table A.3: *Values of α and the proportion of runs in which the desired inclusion rule was included in the LP during column generation, as well as included in the final rule set. Results are averaged over values of λ_0, λ_1, with the exception of $\lambda_0 = \lambda_1 = 1\mathrm{e} - 2$, because this did not run for $\alpha = 0.97$*

	LP	Final Rule Set
$\alpha = 0.95$	0.50	0.50
$\alpha = 0.96$	0.75	0.71
$\alpha = 0.97$	1.00	0.88
$\alpha = 0.98$	1.00	0.62
$\alpha = 0.99$	1.00	0.62

(a) *Recovery of inclusion rule*

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	0.7	1.0	0.0
$\alpha = 0.96$	1.0	0.8	0.0
$\alpha = 0.97$	0.8	0.7	1.0
$\alpha = 0.98$	0.7	0.5	0.7
$\alpha = 0.99$	0.8	0.7	0.3

(b) *Avg. # of rules*

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	210.2	115.8	6.0
$\alpha = 0.96$	334.3	148.0	5.0
$\alpha = 0.97$	25.2	75.2	49.8
$\alpha = 0.98$	25.0	24.7	24.3
$\alpha = 0.99$	23.3	23.3	23.7

(c) *Avg. # of Perfect Rules*

	λ_1 =1e-6	λ_1 =1e-4	λ_1 =1e-2
$\alpha = 0.95$	200.2	105.8	0.0
$\alpha = 0.96$	326.0	140.0	0.0
$\alpha = 0.97$	19.5	69.0	42.2
$\alpha = 0.98$	21.3	21.0	20.7
$\alpha = 0.99$	19.0	18.7	19.7

Table A.4: *Value of α parameter and λ_1 parameters, along with (a) the proportion of runs (across all other hyperparameter settings) in which the synthetic region was correctly identified by the final rule set, (b) the number of rules in the final solution, and (c) the number of perfect rules, defined as those which exclude none of the samples but which exclude some number of reference points. Note that these results marginalize over $\lambda_0 \in \{1e-6, 1e-4\}$ because $\lambda_0 = \lambda_1 = 1e-2$ did not run for $\alpha = 0.97$, and (b-c) are averaged across all runs.*

Table A.5: ***Rec***: *Proportion of runs where synthetic exclusion was recovered.* **# *R***: *Number of rules in final output.* **# *PR***: *Number of "perfect" rules which exclude zero data points.* ***Length***: *Average length of rules. Each entry is the average of three independent runs with different random seeds, and run with* $B = 15$

α	λ_0	λ_1	Rec	# R	# PR	Length
0.95	0	1e-06	1.00	31.00	17.00	2.36
		1e-04	1.00	19.33	12.00	2.25
		1e-02	0.00	5.00	0.00	1.00
	1e-06	1e-06	1.00	30.67	17.00	2.37
		1e-04	1.00	19.33	12.00	2.25
		1e-02	0.00	5.00	0.00	1.00
	1e-04	1e-06	1.00	27.00	15.00	2.36
		1e-04	1.00	18.33	12.00	2.23
		1e-02	0.00	5.00	0.00	1.00
	1e-02	1e-06	1.00	6.00	1.00	1.17
		1e-04	1.00	6.00	1.00	1.17
		1e-02	0.00	5.00	0.00	1.00
0.96	0	1e-06	1.00	46.33	28.33	2.69
		1e-04	1.00	43.67	25.00	2.43
		1e-02	0.00	4.00	0.00	1.00
	1e-06	1e-06	1.00	45.33	27.67	2.70
		1e-04	1.00	43.67	25.67	2.41
		1e-02	0.00	4.00	0.00	1.00
	1e-04	1e-06	1.00	45.67	26.00	2.67
		1e-04	1.00	41.00	23.00	2.41
		1e-02	0.00	4.00	0.00	1.00
	1e-02	1e-06	1.00	5.00	1.00	1.20
		1e-04	1.00	5.00	1.00	1.20
		1e-02	0.00	4.00	0.00	1.00
0.97	0	1e-06	1.00	49.67	31.00	2.74
		1e-04	1.00	38.00	30.00	2.51
		1e-02	1.00	4.00	1.00	1.25
	1e-06	1e-06	1.00	49.67	31.00	2.73
		1e-04	1.00	38.00	30.00	2.51
		1e-02	1.00	4.00	1.00	1.25
	1e-04	1e-06	1.00	48.33	29.00	2.71
		1e-04	1.00	37.33	29.33	2.55
		1e-02	1.00	4.00	1.00	1.25
	1e-02	1e-06	1.00	11.67	7.67	2.27
		1e-04	1.00	14.33	10.33	2.43
		1e-02	1.00	4.00	1.00	1.25
0.98	0	1e-06	1.00	47.00	33.67	2.82
		1e-04	1.00	50.67	30.33	2.74
		1e-02	1.00	27.33	16.00	1.97
	1e-06	1e-06	1.00	46.67	33.33	2.81
		1e-04	1.00	50.67	30.33	2.74
		1e-02	1.00	27.00	15.67	1.97
	1e-04	1e-06	1.00	46.00	31.33	2.74
		1e-04	1.00	50.67	31.00	2.74
		1e-02	1.00	28.00	16.33	1.99
	1e-02	1e-06	1.00	37.00	22.33	2.29
		1e-04	1.00	36.67	21.67	2.26
		1e-02	1.00	13.00	8.00	1.95
0.99	0	1e-06	1.00	33.00	23.33	2.33
		1e-04	1.00	33.00	27.33	2.33
		1e-02	1.00	28.33	21.00	1.96
	1e-06	1e-06	1.00	33.00	21.67	2.36
		1e-04	1.00	34.33	24.67	2.30
		1e-02	1.00	28.33	21.00	1.96
	1e-04	1e-06	1.00	31.33	25.67	2.34
		1e-04	1.00	27.00	20.67	2.17
		1e-02	1.00	28.33	21.00	1.96
	1e-02	1e-06	1.00	28.33	21.33	2.08
		1e-04	1.00	30.67	23.67	2.11
		1e-02	1.00	25.67	18.67	1.96

A.2 Application of OverRule to Policy Evaluation

In this section we give the detailed algorithm for applying OverRule to policy evaluation, as described in the corresponding chapter of this thesis. In this context, we wish to evaluate not a specific treatment decision (e.g., the average treatment effect of giving a drug vs. withholding it), but rather a conditional *policy* representing a personalized treatment regime, which we will refer to as the *target* policy. This problem falls under the setting of off-policy policy evaluation when this target policy π differs from the policy which generated the data, which we observe in the observational data as $p(T = t \mid x)$.

Rationale for $\mathcal{B}^\epsilon(\pi)$: In the corresponding chapter of this thesis, we drew a connection between the set $\mathcal{B}^\epsilon$ and the following set, a function of the target policy π, $\mathcal{B}^\epsilon(\pi) := \{x \in \mathcal{X}; \forall t : \pi(t \mid x) > 0 : p(T = t \mid x) > \epsilon\}$. In this section, we recall the theoretical rationale for why we are restricted to this set, if we wish to evaluate the policy π given samples generated according to $p(T = t \mid x)$.

Following similar notation to Kallus and Zhou (2018b), we will let $X \in \mathcal{X}$ correspond to covariates, $Y \in \mathcal{Y}$ to an outcome of interest, $T \in \mathcal{T}$ to a treatment decision. We write $\pi(t|x_i)$ as the probability of each treatment under the policy, which may be stochastic. We write $Y(t)$ to represent the potential outcome under treatment t. In this setting, we wish to evaluate the expected value of Y under the target policy, which we denote as $\mathbb{E}[Y(\pi)]$.

Proposition A.1 (Informal)**.** *The expectation $\mathbb{E}[Y(\pi)]$ is only defined w.r.t. the observed distribution $p(X, T, Y)$ for the subset $B \in \mathcal{X}$ such that $\forall x \in B,\ \pi(T = t \mid X = x) > 0 \implies p(T = t \mid X = x) > 0$*

Proof. Under the assumption that ignorability (Pearl, 2009) holds, we can write out our desired quantity as follows in terms of observed distribution $p(X, T, Y)$. For brevity, let $p(t \mid x) = p(T = t \mid X = x), p(x) = p(X = x)$, et cetera.

$$
\begin{aligned}
& \mathbb{E}[Y(\pi)] && \text{(A.1)} \\
={} & \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x)\pi(t \mid x) \cdot p(Y(t) = y \mid x, t) dx dt dy \\
={} & \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x) \frac{\pi(t \mid x)}{p(t \mid x)} \\
& \cdot p(Y(t) = y \mid x, t) p(t \mid x) dx dt dy && \text{(A.2)} \\
={} & \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x) p(t \mid x) \\
& \cdot p(Y = y \mid x, t) \frac{\pi(t \mid x)}{p(t \mid x)} dx dt dy && \text{(A.3)} \\
={} & \int_{\mathcal{X},\mathcal{T},\mathcal{Y}} y \cdot p(x, t, y) \cdot \frac{\pi(t \mid x)}{p(t \mid x)} dx dt dy && \text{(A.4)}
\end{aligned}
$$

Where in Equation (A.2) we multiply by one, in Equation (A.3) we use the assumption of ignorability to write $p(Y(t) = y \mid X = x, T = t) = p(Y = y \mid X = x, T = t)$ and rearrange terms, and in Equation (A.4) we collect the terms which represent the observed distribution. For our purposes, it is sufficient to look at the integral in Equation (A.4) to see that it requires the condition that for all $(x, t) \in \mathcal{X} \times \mathcal{T}$, the relationship $\pi(T = t \mid X = x) > 0 \implies p(T = t \mid X = x) > 0$ must hold. □

The condition given in Proposition A.1 is sometimes referred to as the condition of *coverage* (see Sutton and Barto, 2017, Section 5.5) in off-policy evaluation. Rewriting Equation (A.4) as an expectation over the observed distribution, we can see that this leads naturally to the importance sampling (Kahn, 1955) estimator

$$
\mathbb{E}\left[Y \frac{\pi(T = t \mid X = x)}{p(T = t \mid X = x)} \right] \approx \frac{1}{n} \sum_{i=1}^{n} y_i \frac{\pi(t_i \mid x_i)}{p(t_i \mid x_i)}, \qquad \text{(A.5)}
$$

which approximates our desired quantity. If $\epsilon > p(t|x) > 0$ for some small value of ϵ, then the variance of the importance sampling estimator increases dramatically. This motivates our notion of "strict" coverage, that for each value of $x \in \mathcal{B}^{\epsilon}(\pi)$, we require that for all actions t such that $\pi(t|x) > 0$, the condition $p(t|x) > \epsilon$ must hold.

Note that this differs conceptually from the binary treatment case in an important respect: Since we are not seeking to contrast all treatments, we do not require that $\mu(t|x) > \epsilon, \forall t \in \mathcal{T}$, but rather just for those treatments which have positive probability of being taken under the target policy.

Algorithmic Details As described in the corresponding chapter of this thesis, applying OverRule to the policy evaluation setting only requires a single change to the procedure, which is that the set $\hat{B}^{\epsilon}(\pi)$ is used in place of the set $\hat{B}^{\epsilon}$ in Equation (3.9) in Section 3.4.2. Nonetheless, we provide an explicit self-contained sketch of the procedure here to avoid any confusion:

1. Given a dataset, find an α-MV set $\mathcal{S}^{\alpha}$ using the approach given in the corresponding chapter of this thesis.

2. Using this set, learn the conditional probabilities of each possible treatment $t \in \mathcal{T}$, resulting in estimated propensities $\hat{p}(T = t \mid X = x)$

3. For each data point in the support set $\mathcal{S}^{\alpha}$, assign the label

$$\hat{b}_i(\pi) = \prod_{t \in \pi(x_i)} \mathbb{1}[\hat{p}(T = t \mid X = x_i) \geq \epsilon],$$

   where $\pi(x_i) \coloneqq \{t : \pi(t|x_i) > 0\}$. The set $\hat{B}^{\epsilon}(\pi)$ is the collection of data points such that $\hat{b}_i(\pi) = 1$. Note that we know the target policy π that we are evaluating, so we can evaluate $\pi(t|x_i)$ for each data point.

4. Solve the following Neyman-Pearson-like classification problem, using the techniques discussed in the corresponding chapter of this thesis. Note that this is identical to solving Equation (3.9) in Section 3.4.2, with the substitution of

$\hat{B}^{\epsilon}(\pi)$ for $\hat{B}^{\epsilon}$:

$$
\begin{aligned}
\hat{\mathcal{B}}(\pi) := \arg\min_{C} \quad & \frac{1}{|\hat{\mathcal{S}} \setminus \hat{B}|} \sum_{i \in \hat{\mathcal{S}} \setminus \hat{B}^{\epsilon}(\pi)} \mathbb{1}[x_i \in \mathcal{C}] + R(\mathcal{C}) \\
\text{s.t.} \quad & \sum_{i \in \hat{\mathcal{S}} \cap \hat{B}^{\epsilon}(\pi)} \mathbb{1}[x_i \in \mathcal{C}] \geq \beta |\hat{\mathcal{S}} \cap \hat{B}^{\epsilon}(\pi)| \ .
\end{aligned}
$$

A.3 Additional Experimental Results

As a general note across all experiments: When estimating support in OverRule, we use $m_R = c \cdot m \cdot d$ uniform reference samples where $c > 0$ is some constant, m is the number of data samples and d their dimension. Continuous features were binarized by deciles unless otherwise specified. Finally, for propensity-based base estimators, we use the standard threshold $\epsilon = 0.1$ (Crump et al., 2009) throughout.

A.3.1 Iris

For the results given in the corresponding chapter of this thesis, we fit OverRule using a k-NN base estimator ($k = 8$) and DNF Boolean rules for both support and overlap rules, with $\alpha = 0.9$ and regularization $\lambda_0 = 2 \cdot 10^{-2}, \lambda_1 = 0$ for support rules, a cutoff of $\epsilon = 0.1$, and $\beta = 0.9, \lambda_0 = 10^{-2}, \lambda_1 = 0$ for overlap rules.

A.3.2 Jobs

For the results given in the corresponding chapter of this thesis, we use the following hyperparameters:

1. **Support Rules**: CNF formulation, along with hyperparamters $\alpha = 0.98, \lambda_0 = 10^{-2}, \lambda_1 = 10^{-3}$.

2. **Base Estimators**: For CBB we used $\alpha = 0.1$, for the logistic regression propensity estimator we used $C = 1$ in `LogisticRegression` in scikit-learn, and other

hyperparameters were chosen based on cross-validation: For k-NN, we selected $k \in \{2, 4, \ldots, 20\}$ based on held-out accuracy in predicting group membership and used $1/k$ as threshold. For OSVM, we use a Gaussian RBF-kernel with bandwidth $\gamma \in [10^{-2}, 10^{2}]$, selected based on the held-out likelihood of kernel density estimation.

3. **Overlap Rules**: We use a DNF formulation with $\beta = 0.9$ and select $\lambda_0 \in [10^{-4}, 10^{-1}]$ and $\lambda_1 \in [10^{-4}, 10^{-2}]$. Within each class of base estimators, we choose these parameters based on average *training* performance over 5-fold CV, choosing the setting in each class that achieves a balanced accuracy (with respect to the base-estimator overlap labels) within 1% of the best performing model in the class, while minimizing the number of rules.

Note that the reported results are using the held-out portions of each 5-fold CV run, and using the ground-truth overlap labels, which are at no point used during the hyperparameter tuning process. This reflects a real-world scenario where ground-truth is unknown and only the base-estimator derived labels are given. The reported rules in the figure were selected from one of the five cross-validation runs for the same hyperparameter setting chosen using the above procedure. In Figure A-2 we see the correlation between held-out balanced accuracy for the rule set w.r.t. the experimental label, and the balanced accuracy for the rule set in approximating the base estimator. Note that AUC is equal to balanced accuracy for binary predictions.

A.3.3 Opioids

For the results in the corresponding chapter of this thesis, we fit an OverRule model (OR) to a random forest base estimator with $\beta = 0.8$ for $\mathcal{B}$ and $\alpha = 0.9$ for $\mathcal{S}$ picked a priori. The hyperparameter λ_0 was set to $\lambda_0 =$1e $- 3$ for $\mathcal{B}$, and $\lambda_0 =$1e $- 5$ for $\mathcal{S}$, and $\lambda_1 = 0$ for both.

For a full table of covariate statistics for the Opioids dataset, see Table A.6. For a illustration of the rules learned by OverRule to describe the complement of the overlap

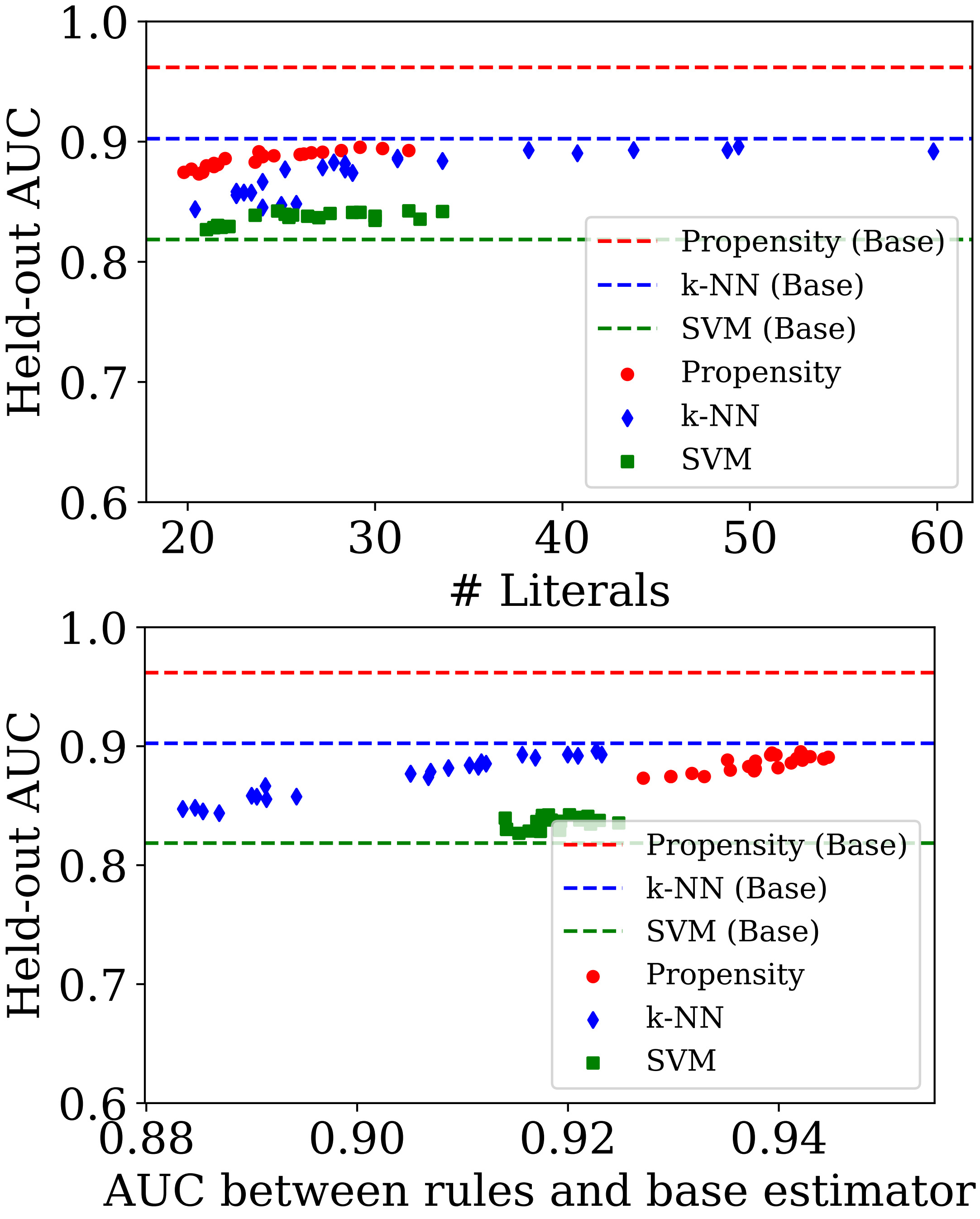
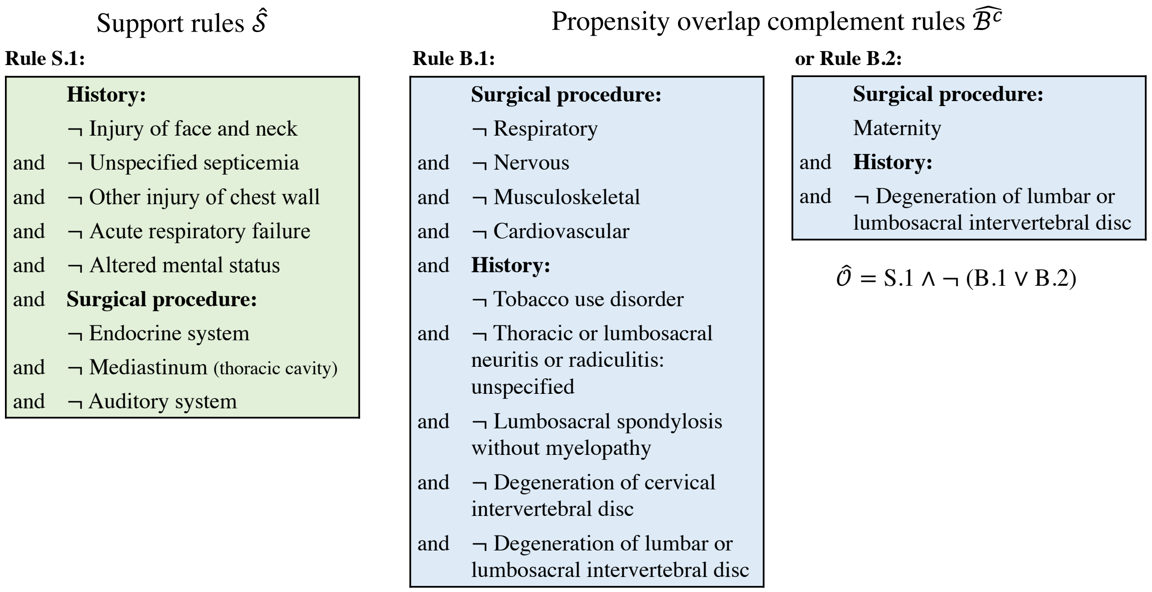


Figure A-2: *Results from the Jobs datasets for OverRule approximations of different base estimators, sweeping λ_0, λ_1. AUC (i.e., balanced accuracy) is measured with respect to the experimental indicator. The dotted line 'Propensity (base)' refers to the logistic regression base estimator, 'k-NN (base)' refers to the k-NN base estimator, and 'SVM (base)' refers to the one-class SVM. The colored points refer to performance of OverRule using the respective base estimator, for different values of λ_0, λ_1*

set, see Figure A-3.



Support rules $\hat{\mathcal{S}}$

Rule S.1:

History:
¬ Injury of face and neck
and ¬ Unspecified septicemia
and ¬ Other injury of chest wall
and ¬ Acute respiratory failure
and ¬ Altered mental status
and **Surgical procedure:**
¬ Endocrine system
and ¬ Mediastinum (thoracic cavity)
and ¬ Auditory system

Propensity overlap complement rules $\widehat{\mathcal{B}^c}$

Rule B.1:

Surgical procedure:
¬ Respiratory
and ¬ Nervous
and ¬ Musculoskeletal
and ¬ Cardiovascular
and **History:**
¬ Tobacco use disorder
and ¬ Thoracic or lumbosacral neuritis or radiculitis: unspecified
and ¬ Lumbosacral spondylosis without myelopathy
and ¬ Degeneration of cervical intervertebral disc
and ¬ Degeneration of lumbar or lumbosacral intervertebral disc

or Rule B.2:

Surgical procedure:
Maternity
and **History:**
and ¬ Degeneration of lumbar or lumbosacral intervertebral disc

$\hat{O}$ = S.1 ∧ ¬ (B.1 ∨ B.2)

Figure A-3: *OverRule description of the* complement *of the overlap between post-surgical patients with higher and lower opioid prescriptions. If the support rule (left) applies and* neither *propensity overlap rule (right) applies, a patient is consider to be in the overlap set. ¬ indicates a negation. The rules cover 36% of patients with balanced accuracy 0.92 w.r.t. the base estimator (random forest). Procedures are not mutually exclusive.*

Supplemental Rules: We learned an additional set of rules, motivated by our experiments in Section 3.5.3, where we noted that the support rules did not capture certain combinations of surgery types or conditions that should be rare or non-existent. This motivated the empirical investigation in Section A.1.2, and this vignette represents the result of re-running our procedure with this goal in mind.

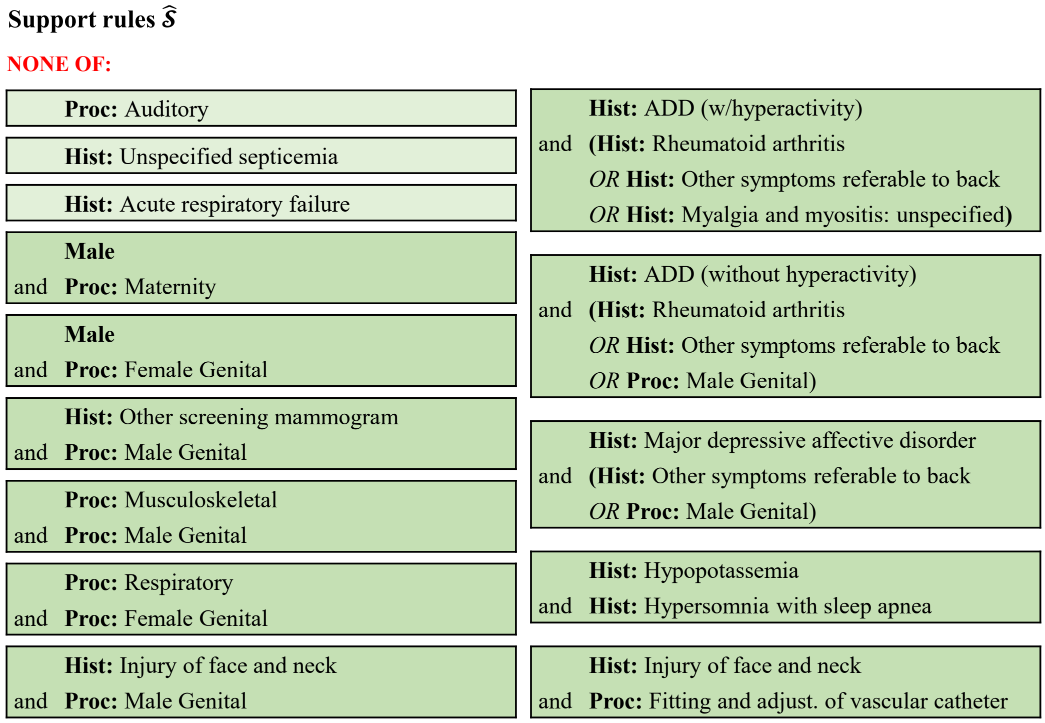
For support rules, we followed the recommendations laid out in Section A.1.2, choosing to use a CNF formulation with $\alpha = 0.98, \lambda_0 = 0, \lambda_1 = 0.01$. Continuous features were binarized using deciles. For our base estimator, we used a random forest classifier with 100 trees and 20 minimum samples per leaf, and we used $\epsilon = 0.1$ as our cutoff. For the overlap rules, we searched over the following grid of hyperparameters, with the goal of maximizing balanced accuracy with respect to the overlap labels on a validation set: $\beta \in \{0.8, 0.9, 0.95\}$ and then a set where $\lambda_0 = 0$ and $\lambda_1 \in \{10^{-3}, 2 \cdot 10^{-3}, 10^{-2}\}$, and a set where $\lambda_1 = 0$ and $\lambda_0 \in \{10^{-3}, 2 \cdot 10^{-3}, 10^{-2}\}$. The selected hyperparameters were $\beta = 0.95, \lambda_0 = 0, \lambda_1 = 10^{-3}$. The support rules cover 98.5% of the test samples, and the overlap rules achieved a balanced accuracy of 0.96 on a held-out test set (with respect to the overlap labels) and covered 36% of the test samples. The chosen ruleset is given in given in Figures A-4-A-5.

We note that the resulting support rules, in line with the findings in Section A.1.2, include a large number of rules that exclude zero training data points, by identifying rare interactions of features. For instance, the rules identify that there are *no men in our dataset who have maternity surgery*, an intuitive exclusion.

We shared this rule set with one of the participants of the original user study, who made the following observations: First, the support rules in Figure A-4 generally made sense as excluding combinations that are intuitively absent from the data (e.g., men w/maternity surgery) or that are just combinations of features that are themselves rare. Regarding the overlap rules in Figure A-5, they observed that B.1 and B.2 were consistent with clinical intuition, where B.2 likely serves to exclude C-section patients with epidurals. B.3 and B.4 were intuitive with the exception of the negations, e.g., it is unclear what the role of abdominal pain is in B.3, although it could be correlated with generalized pain syndromes. B.5-B.7 correspond to individuals with lower back pain (Lumbago) and neck pain (Cervicalgia) which are intuitive indicates for higher doses of opioids. B.8 corresponds to plastic surgery, and the broad category of respiratory surgery in B.9 could correspond to thoracic surgery, one of the main surgical categories associated with opioid misuse. B.10-B.12 relate to back pain, which is associated with higher opioid dosages.

A.3.4 Observational Study: Policy Evaluation of Antibiotic Prescription Guidelines

Antibiotic resistance is a growing problem in the treatment of urinary tract infections (UTI) (Sanchez et al., 2016), a common infection for which more than 1.6 million prescriptions are given annually in the United States (Shapiro et al., 2013). With this in mind, we are interested in the following clinical problem: When a patient presents



Support rules $\hat{\mathcal{S}}$

NONE OF:

Proc: Auditory

Hist: Unspecified septicemia

Hist: Acute respiratory failure

Male
and **Proc:** Maternity

Male
and **Proc:** Female Genital

Hist: Other screening mammogram
and **Proc:** Male Genital

Proc: Musculoskeletal
and **Proc:** Male Genital

Proc: Respiratory
and **Proc:** Female Genital

Hist: Injury of face and neck
and **Proc:** Male Genital

Hist: ADD (w/hyperactivity)
and **(Hist:** Rheumatoid arthritis
OR **Hist:** Other symptoms referable to back
OR **Hist:** Myalgia and myositis: unspecified**)**

Hist: ADD (without hyperactivity)
and **(Hist:** Rheumatoid arthritis
OR **Hist:** Other symptoms referable to back
OR **Proc:** Male Genital)

Hist: Major depressive affective disorder
and **(Hist:** Other symptoms referable to back
OR **Proc:** Male Genital)

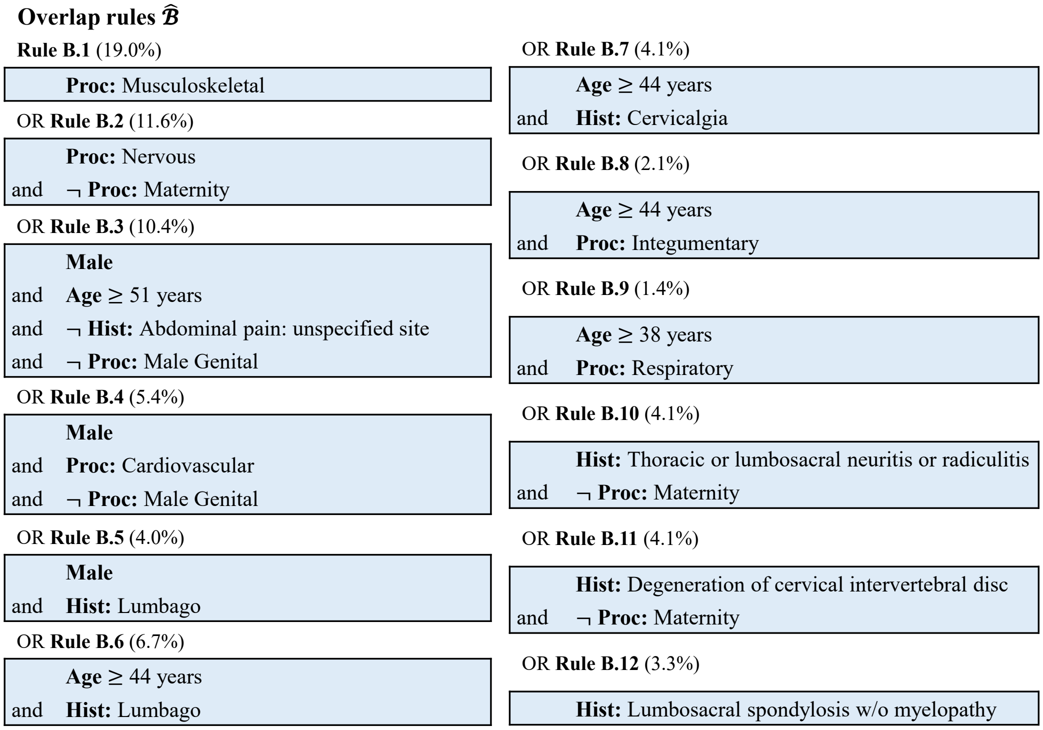
Hist: Hypopotassemia
and **Hist:** Hypersomnia with sleep apnea

Hist: Injury of face and neck
and **Proc:** Fitting and adjust. of vascular catheter

Figure A-4: *Support Rules using CNF formulation for the Opioids task.* ***Proc*** *indicates a procedure, and* ***Hist*** *indicates a history of a condition. A sample is considered in the support set if NONE of the above rules apply. Note that rules are negated for simplicity of presentation, as "AND NOT (X AND Y)" is equivalent to "AND (NOT X OR NOT Y)", and in some cases several rules are combined for simplicity of presentation (e.g., those related to Attention Deficit Disorder). Dark green rules are highlighted to indicate that they cover <4 training samples (and in many cases zero training samples) in line with our findings in Section A.1.2 for this setting of hyperparameters.*

with a UTI, the physician needs to choose between a range of antibiotics, with the dual goals of (a) treating the infection, and (b) minimizing the use of broad-spectrum antibiotics, which are more likely to select for drug-resistant strains of bacteria.

In this context, we might be interested in evaluating a range of potential treatment policies. For our purposes, we will use a pre-defined policy: The clinical guidelines published by the Infectious Disease Society of America (IDSA) for treatment of uncomplicated UTIs in female patients (Gupta et al., 2011). Using the policy evaluation formulation of $\mathcal{B}^{\epsilon}(\pi)$, we will apply OverRule to a conservative interpretation of the IDSA guidelines, using data curated from the Electronic Medical Record (EMR) of two academic medical centers.



Overlap rules $\widehat{\mathcal{B}}$

Rule B.1 (19.0%)

Proc: Musculoskeletal

OR **Rule B.2** (11.6%)

Proc: Nervous
and ¬ **Proc:** Maternity

OR **Rule B.3** (10.4%)

Male
and **Age** ≥ 51 years
and ¬ **Hist:** Abdominal pain: unspecified site
and ¬ **Proc:** Male Genital

OR **Rule B.4** (5.4%)

Male
and **Proc:** Cardiovascular
and ¬ **Proc:** Male Genital

OR **Rule B.5** (4.0%)

Male
and **Hist:** Lumbago

OR **Rule B.6** (6.7%)

Age ≥ 44 years
and **Hist:** Lumbago

OR **Rule B.7** (4.1%)

Age ≥ 44 years
and **Hist:** Cervicalgia

OR **Rule B.8** (2.1%)

Age ≥ 44 years
and **Proc:** Integumentary

OR **Rule B.9** (1.4%)

Age ≥ 38 years
and **Proc:** Respiratory

OR **Rule B.10** (4.1%)

Hist: Thoracic or lumbosacral neuritis or radiculitis
and ¬ **Proc:** Maternity

OR **Rule B.11** (4.1%)

Hist: Degeneration of cervical intervertebral disc
and ¬ **Proc:** Maternity

OR **Rule B.12** (3.3%)

Hist: Lumbosacral spondylosis w/o myelopathy

Figure A-5: *Overlap rules, where the percentage next to each rule indicates the percentage of the dataset that is covered by that rule. Collectively, these rules cover 36% of the held-out datapoints.*

The official guidelines discuss the importance of patient and population level risk factors in predicting resistance, and include some factors that we do not observe in our data (such as drug allergies). In order to characterize the guideline explicitly as a policy that we can evaluate in our dataset, we used the following interpretation:

- Choose the first-line agent, either Nitrofurantoin (NIT) or Trimethoprim/Sulfamethoxazole (SXT), to which the patient did not have previous antibiotic exposure or resistance in the prior 90 days. Additionally, if local rates of resistance to SXT are $\geq 20\%$ in the prior 30-90 days, then avoid prescription of SXT.

- If neither of the first-line agents are indicated, then prescribe Ciprofloxacin (CIP), a second-line agent.

Experimental details From our data set, we selected all patients from 2007–2017 which had a UTI, and were prescribed one of the four most common antibiotics: NIT, SXT, CIP, or Levofloxacin (LVX). Features include demographics (race, gender, age, and veteran status), comorbidities observed in the past 90 days, information about previous infections (organism, antibiotics given, and resistance profile), hospital ward (inpatient, outpatient, ER, and ICU), and indicators for pregnancy and nursing home residence in the past 90 days. The local rates of resistance (for each hospital ward) are given over the past 30–90 days, and used at the patient level as a feature, as well as an input to the decision of the guidelines.

We preprocess our data first, removing any binary feature with a prevalence of less than 0.1%, and any associated subject: This results in the removal of 48 binary features with less than 0.1% prevalence and 888 corresponding subjects. This leaves a total of 156 (150 binary, 6 continuous) features and 64593 subjects. Detail on all remaining features are given in Table A.7. For the purposes of running our algorithm, we convert all continuous variables into binary variables by using indicator functions for deciles.

We then characterize the support set $\mathcal{S}^\alpha$ as described in the corresponding chapter of this thesis, using a DNF formulation, along with $\alpha = 0.95, \lambda_0 = 0.01, \lambda_1 = 0$. Using the data points which fall into the support set, we then estimate the propensity $p(t|x)$ of prescribing each of the four drugs using a random forest classifier, with hyperparameter selection done using 5-fold cross-validation on 80% of the remaining cohort used as a training set, over the following parameter grid: Number of estimators $\in [100, 500]$, Minimum samples per leaf (as fraction of total) $\in [0.005, 0.01, 0.02]$. The resulting calibration curves for each antibiotic are given in Figure A-6, using the remaining held-out 20% of the data. Using these propensity scores, we apply the procedure described in Section A.2 to estimate the region of strict coverage, $\hat{\mathcal{B}}^\epsilon(\pi)$ using Boolean rules, and the resulting rules are given in Figure A-7. For this stage, we used a DNF formulation and hyperparameters of $\beta = 0.9, \lambda_0 = 0.03, \lambda_1 = 0$.

Clinical Validity / Interpretation Towards understanding the clinical validity of these rules, we interviewed a clinician who specialises in infectious diseases. First, we asked them, based on the available features, which they would expect to differentiate

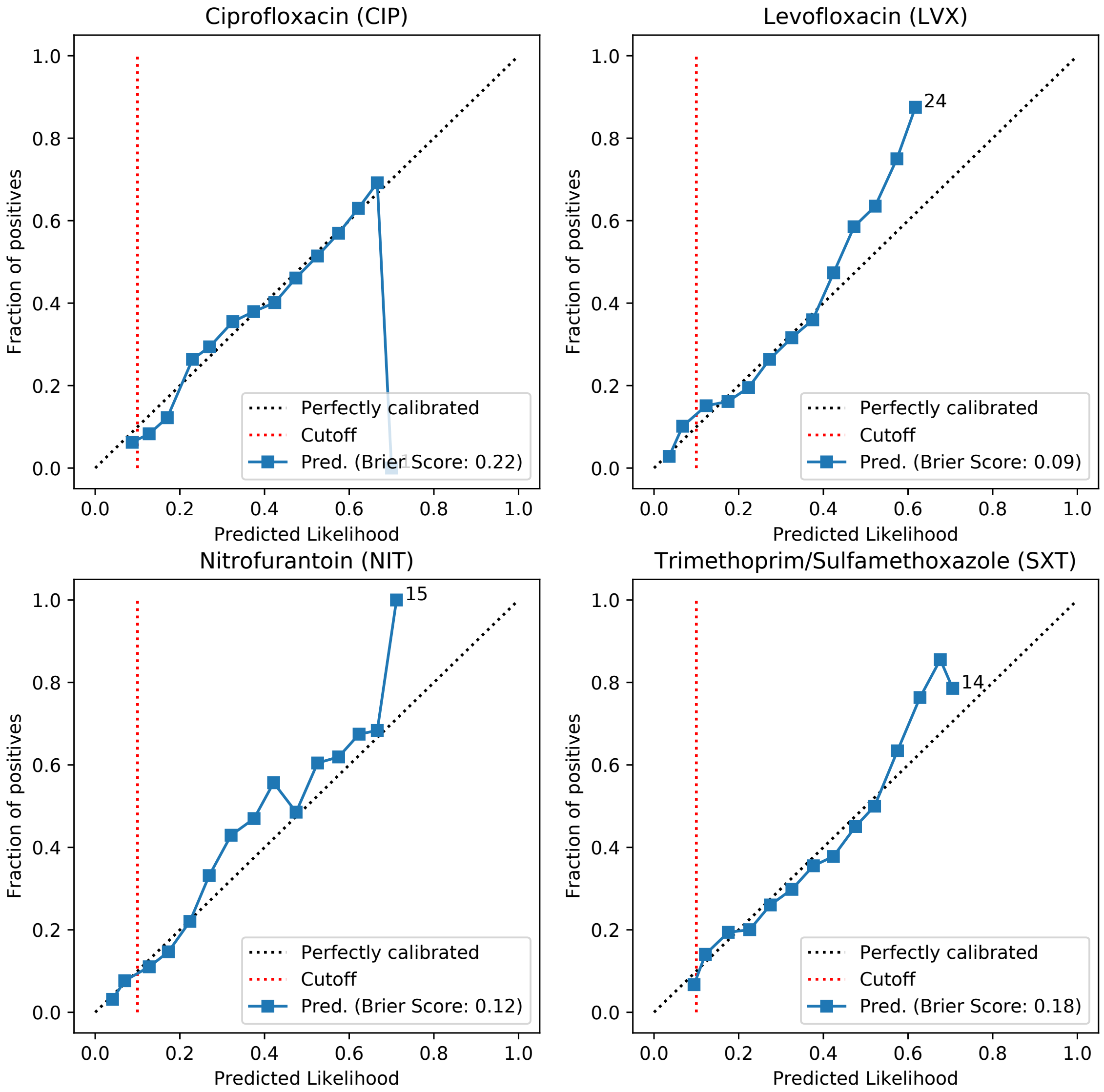
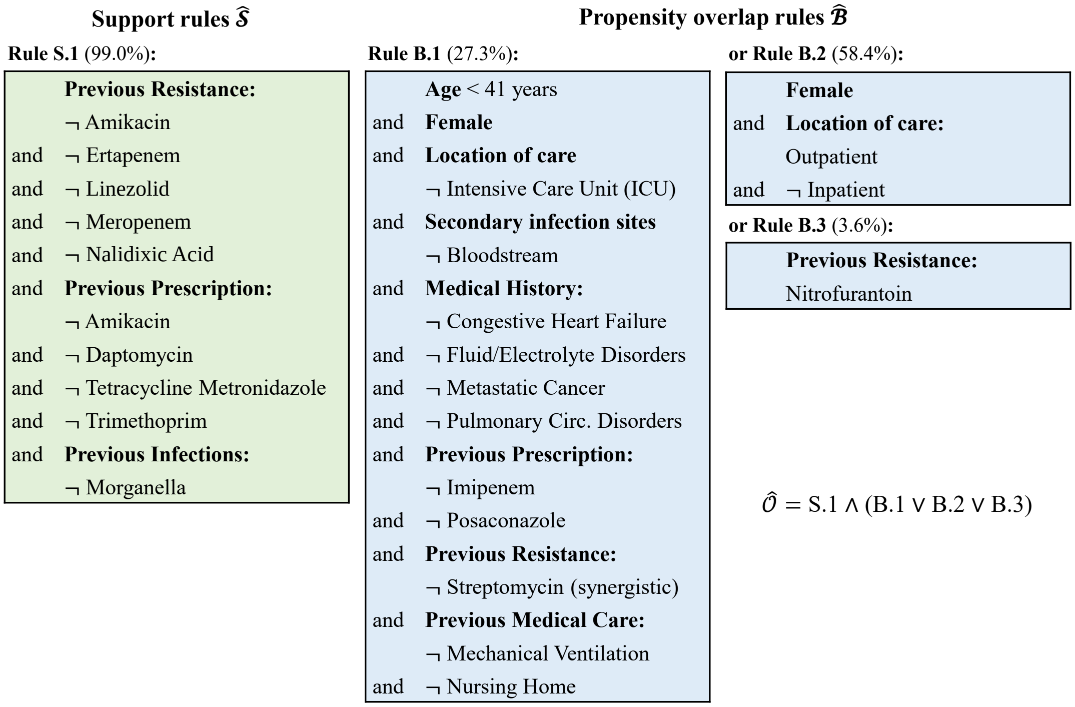


Figure A-6: *Calibration curves for each antibiotic, using 20 evenly spaced bins in the range* $[0, 1]$*. Numbers indicate the number of samples, and are given when when the number of samples in a bin is less than 0.5% of the total. The cutoff is a reminder that* $\epsilon = 0.1$ *in this experiment: For any subject with covariates* x*, the propensity must be above this cutoff for every treatment under the target policy (i.e., for all* t *such that* $\pi(t|x) > 0$*) for them to be included in the coverage region.*

between subjects for whom the policy is or is not followed. They noted that the guidelines are designed for uncomplicated cases: In particular, patients who have a Foley catheter (a catheter used to drain urine from the bladder) are not covered under these guidelines, because infections in these patients tend to be more complex (e.g., the infection could have been introduced by the catheter itself). The use of the Foley



Support rules $\hat{\mathcal{S}}$

Rule S.1 (99.0%):

Previous Resistance:
¬ Amikacin
and ¬ Ertapenem
and ¬ Linezolid
and ¬ Meropenem
and ¬ Nalidixic Acid
and **Previous Prescription:**
¬ Amikacin
and ¬ Daptomycin
and ¬ Tetracycline Metronidazole
and ¬ Trimethoprim
and **Previous Infections:**
¬ Morganella

Propensity overlap rules $\hat{\mathcal{B}}$

Rule B.1 (27.3%):

Age < 41 years
and **Female**
and **Location of care**
¬ Intensive Care Unit (ICU)
and **Secondary infection sites**
¬ Bloodstream
and **Medical History:**
¬ Congestive Heart Failure
and ¬ Fluid/Electrolyte Disorders
and ¬ Metastatic Cancer
and ¬ Pulmonary Circ. Disorders
and **Previous Prescription:**
¬ Imipenem
and ¬ Posaconazole
and **Previous Resistance:**
¬ Streptomycin (synergistic)
and **Previous Medical Care:**
¬ Mechanical Ventilation
and ¬ Nursing Home

or Rule B.2 (58.4%):

Female
and **Location of care:**
Outpatient
and ¬ Inpatient

or Rule B.3 (3.6%):

Previous Resistance:
Nitrofurantoin

$\hat{O}$ = S.1 ∧ (B.1 ∨ B.2 ∨ B.3)

Figure A-7: *OverRule description of the coverage region for policy evaluation of the clinical guidelines. Beside each rule we give the percentage of subjects that are covered by the rule in the test set. Overall, the rules for $\hat{B}$ cover 65.4% of the data points in the support region (compared to the 71% of points labelled by our base estimator), and they have an balanced accuracy of 0.96 versus the base estimator.*

catheter is common during intensive care (e.g., in the ICU), so complex hospitalized patients are less likely to be treated according to the guidelines.

With that in mind, they reviewed the available features and noted the following: (i) While UTIs are common for women, they are rare for men; Men with UTIs tend to be more complicated cases, because it is indicative of deeper abnormalities. Similarly, pregnant women are excluded from the guidelines. (ii) Of the comorbidities given, none of them should directly disqualify patients from the guidelines, except potentially for complicated diabetes. (iii) Prior organisms / resistance / prescriptions should not directly disqualify patients from the guidelines, though they will influence the type of antibiotic given. In particular, if a patient has had previous resistance to an antibiotic, they are unlikely to be prescribed it again. (iv) The previous procedures given (with the exception of surgery) are associated with ICU patients. For instance, mechanical ventilation and parenteral nutrition are exclusive to the ICU, and those patients likely have a Foley catheter as well. Surgery is too broad of a category to draw any conclusions. (v) In terms of locations besides the ICU, patients who are admitted to the hospital and who are on intravenous (IV) antibiotics already will be treated differently. The guidelines are focused on oral antibiotics, whereas if an IV already exists, additional IV antibiotics are likely to be given instead.

Having discussed these points first, we then showed them the rules learned by the OverRule algorithm, and asked for their interpretation, as well as for any critiques of the rules based on their clinical knowledge. Their reaction to each of the rules was as follows:

- **Rule B.1**: This appears to correspond to a relatively straightforward young inpatient female (given that Rule B.2 covers all outpatient females). In particular, it rules out ICU patients directly, as well as those with recent mechanical ventilation, which would indicate a recent ICU stay. It also rules out patients with current bloodstream infections, and those who had previously been tested for (and found to be) resistance to Streptomycin (synergistic): This is only tested for in the context of bloodstream infections by enterococcus, and would be an indicator of previous bloodstream infections. Imipenem is an IV antibiotic only given in inpatient settings, and posaconazole is an antifungal used in bone marrow transplant patients. Patients who are both young and in a nursing home tend to be more complex, e.g., they may be paralysed or otherwise unable to perform activities independently. Finally, the excluded comorbidities are less intuitive, because some of them (e.g., congestive heart failure) manifest with a range of severity: For patients with controlled congestive heart failure, this is not a contraindication for following the guidelines, but if they are fully decompensated, then they would likely be on a Foley catheter.

- **Rule B.2**: This concisely describes the most common manifestation of UTI and the set of patients who are most likely to be treated according to the guidelines[3].

[3] Note that outpatient and "not inpatient" can appear in the same rule without being redundant,

- **Rule B.3**: The conjecture is that this represents patients who have had an uncomplicated UTI in the past, since patients are usually tested for the antibiotics under consideration by a physician, and since nitrofurantoin is one of the first-line treatments for uncomplicated UTIs.

From a quantitative perspective, we compared the learned region with an explicitly constructed cohort of patients whose inclusion criteria were explicitly designed to make them eligible for application of the IDSA guidelines. In particular, we defined this cohort as including non-pregnant women between the ages of 18 to 55 years of age with no record of genitourinary surgery or instrumentation, immunosuppression, indwelling catheters, or neurologic dysfunction in the preceding 90 days. There were 14k of these patients, 21% of the total.

In relationship to this conservative subset, the learned region (covering 42k patients, 64% of total) covers 96% of the explicitly constructed cohort, while also demonstrating that a broader set of patients are treated according to these guidelines in practice.

A.4 Theoretical Results on Regularized Minimum-Volume Boolean Rules

A.4.1 Bounds on minimum volume

In this subsection, we derive lower bounds on the volume of optimal DNF Boolean rules in problem (3.5).

First we obtain an expression for the normalized volume of a clause in a DNF (we use the terms clause and conjunction interchangeably in the case of a DNF). We express the domain $\mathcal{X}$ as the Cartesian product $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. A DNF rule with K clauses a_k

because multiple specimens collected on the same day for the same patient are collapsed into a single subject.

is written as

$$r(x) = \bigvee_{k=1}^{K} a_k(x) = \bigvee_{k=1}^{K} \bigwedge_{j \in \mathcal{J}_k} (x_j \in \mathcal{S}_{jk}), \tag{A.6}$$

where $\mathcal{J}_k$ is the set of covariates participating in clause k, and each $x_j \in \mathcal{S}_{jk} \subseteq \mathcal{X}_j$ is a subset membership condition on an individual covariate. Examples of such conditions are (Age ≥ 30) for a continuous-valued covariate and (Sex = Female) for a discrete-valued one. For $j \notin \mathcal{J}_k$, it is understood that $x_j \in \mathcal{X}_j$, i.e. there is no restriction on x_j. The volume of clause a_k is then given by the product

$$V(a_k) = \prod_{j \in \mathcal{J}_k} |\mathcal{S}_{jk}| \prod_{j \notin \mathcal{J}_k} |\mathcal{X}_j|,$$

where $|\mathcal{S}_{jk}|$ is the length of subset $\mathcal{S}_{jk}$ for a continuous covariate j or the cardinality of $\mathcal{S}_{jk}$ for a discrete covariate, and similarly for $|\mathcal{X}_j|$. Likewise, the volume of $\mathcal{X}$ is $\prod_{j=1}^{d} |\mathcal{X}_j|$, and the normalized volume of a_k is therefore

$$\bar{V}(a_k) = \prod_{j \in \mathcal{J}_k} f_{jk}, \quad f_{jk} = \frac{|\mathcal{S}_{jk}|}{|\mathcal{X}_j|} \in [0,1]. \tag{A.7}$$

We define $p_k = |\mathcal{J}_k|$ to be the *degree* of conjunction k.

Proposition A.2. *Assume that the regularization $R(r)$ follows (3.6). Then in any optimal solution to (3.5), all clauses a_k of degree p_k have normalized volume satisfying $\bar{V}(a_k)^{(p_k-1)/p_k} - \bar{V}(a_k) \geq \lambda_1$.*

Proof. Suppose that rule r with corresponding set $\mathcal{C}$ is an optimal solution to (3.5). Recalling the expansion in (A.6), we consider modifications to r in which one condition $(x_j \in \mathcal{S}_{jk})$ is removed from a clause a_k. The modified rule satisfies the mass constraint $P(\mathcal{C}) \geq \alpha$ because it covers at least those points covered by r. From (A.7), the increase in volume is at most $\bar{V}(a_k)((1/f_{jk}) - 1)$, with equality if none of the additional volume is already covered by another clause in r, while the complexity penalty decreases by

λ_1. The change in objective value is thus bounded from above by

$$\bar{V}(a_k)\left(\frac{1}{f_{jk}} - 1\right) - \lambda_1.$$

This upper bound must be non-negative as otherwise r is not optimal. In particular, for $f_{jk} = \max_{j' \in \mathcal{J}_k} f_{j'k}$ and all k we have

$$\bar{V}(a_k)\left(\frac{1}{\max_{j \in \mathcal{J}_k} f_{jk}} - 1\right) \geq \lambda_1.$$

Since (A.7) implies that $\max_{j \in \mathcal{J}_k} f_{jk} \geq \bar{V}(a_k)^{1/p_k}$, the desired result follows. □

For $p > 1$, the function $\bar{V}^{(p-1)/p} - \bar{V}$ is positive and concave on $(0, 1)$ with roots at 0 and 1. For $\lambda_1 > 0$, the equation $\bar{V}^{(p-1)/p} - \bar{V} = \lambda_1$ therefore has either two roots, $0 < \bar{V}_L < \bar{V}_U < 1$, which define an interval where the inequality $\bar{V}^{(p-1)/p} - \bar{V} \geq \lambda_1$ is satisfied, or no roots if λ_1 is too large. We are interested primarily in the root $\bar{V}_L$ as a lower bound on volume. While $\bar{V}_L$ is not available in closed form for $p > 2$, the following corollary gives a simple expression that is a lower bound on $\bar{V}_L$.

Corollary A.1. *Under the assumption in Proposition A.2, in any optimal solution to (3.5), all clauses a_k of degree $p_k > 1$ have normalized volumes of at least $\lambda_1^{p_k/(p_k-1)}$.*

Proof. Proposition A.2 implies $\bar{V}(a_k)^{(p_k-1)/p_k} \geq \lambda_1$ after dropping $-\bar{V}(a_k)$ from the left-hand side. □

Lastly, since the volume of a DNF rule is at least that of any of its clauses, we have the following.

Corollary A.2. *Under the assumption in Proposition A.2, any optimal solution to (3.5) has normalized volume of at least $\lambda_1^{p_{\max}/(p_{\max}-1)}$, where $p_{\max} = \max_k p_k$ is the largest degree of its clauses.*

A.4.2 Bounds on the number of candidate DNF rules

The results in the previous subsection are necessary conditions of optimality for problem (3.5). The implication is that in searching for optimal solutions to (3.5), we may restrict the class $\mathcal{C}$ of DNF rules considered to those satisfying these necessary conditions. In this subsection, we develop the consequences of this restriction, culminating in a bound on $|\mathcal{C}|$, the number of candidate DNF rules (Lemma A.5).

For simplicity, we assume in the following that all variables X_j are binary-valued. An extension to non-binary categorical variables and continuous variables (discretized using interval conditions $l_j \leq x_j \leq u_j$) is likely possible with the additional complications of accounting for the cardinalities of categorical variables and bounding the fractions f_{jk} associated with continuous variables.

First, the simplified lower bound on volume in Corollary A.1 implies an upper bound on conjunction degree.

Lemma A.1. *Assume that the regularization $R(r)$ follows (3.6) and that all variables are binary. Then in any optimal solution to (3.5), the maximum degree of a conjunction is $p_{\max} := 1 + \lfloor \log_2(1/\lambda_1) \rfloor$.*

Proof. The normalized volume of a conjunction of degree p_k is 2^{-p_k}. Corollary A.1 then requires

$$2^{-p_k} \geq \lambda_1^{p_k/(p_k-1)}.$$

Taking logarithms and rearranging, we obtain

$$-1 \geq \frac{1}{p_k - 1} \log_2 \lambda_1,$$
$$p_k \leq 1 + \log_2(1/\lambda_1).$$

The right-hand side can be rounded down since p_k is integer. □

Given Lemma A.1, we may enumerate DNF rules satisfying the lemma according to the numbers of clauses of degree $p = 1, \ldots, p_{\max}$ that they possess. Denote by K_p the number of clauses of degree p and call $\mathbf{K} = (K_1, \ldots, K_{p_{\max}})$ the *signature* of a DNF rule. The signatures of optimal DNF rules obey the following constraint.

Lemma A.2. *Under the assumptions of Lemma A.1, the signature* $\mathbf{K} = (K_1, \ldots, K_{p_{\max}})$ *of an optimal solution to* (3.5) *must satisfy*

$$\sum_{p=1}^{p_{\max}} K_p(\lambda_0 + p\lambda_1) < 1. \tag{A.8}$$

Proof. From (3.6), the complexity penalty of a solution with K_p clauses of degree p, $p = 1, \ldots, p_{\max}$ is given by the left-hand side of (A.8). For a solution to be optimal, it must have lower cost than the trivial "all true" rule, which has a normalized volume of 1 and complexity penalty of 0. In particular, the complexity penalty must be less than 1. □

Let Δ denote the set of signatures that satisfy (A.8), and for $\mathbf{K} \in \Delta$, let $\mathcal{C}(\mathbf{K})$ be the set of DNF rules with signature $\mathbf{K}$. The number of DNF rules satisfying the necessary conditions of optimality in Lemmas A.1 and A.2 can be bounded as follows:

$$|\mathcal{C}| = \sum_{\mathbf{K} \in \Delta} |\mathcal{C}(\mathbf{K})| \leq |\Delta| \max_{\mathbf{K} \in \Delta} |\mathcal{C}(\mathbf{K})|. \tag{A.9}$$

The next two lemmas provide upper bounds on the two right-hand side factors in (A.9).

Lemma A.3. *The number of signatures satisfying* (A.8) *is bounded as*

$$|\Delta| \leq 2 \left(\frac{1}{\lambda_1} \right)^{p_{\max}}.$$

Proof. For simplicity, we consider a superset $\Delta_0 \supseteq \Delta$ obtained by dropping λ_0 from (A.8), i.e.

$$\sum_{p=1}^{p_{\max}} p\lambda_1 K_p \leq 1. \tag{A.10}$$

Condition (A.10) together with the implicit non-negativity constraints $K_p \geq 0$, $p = 1, \ldots, p_{\max}$ define a simplex in $p_{\max}$ dimensions. Bounding the number of signatures in Δ_0 is thus equivalent to bounding the number of non-negative integer points in this simplex. This problem has been studied extensively by mathematicians. Applying e.g. (Yau and Zhang, 2006, eq. (1.5)), we have

$$\begin{aligned}
|\Delta_0| &\le \frac{1}{p_{\max}!}\prod_{p=1}^{p_{\max}}\frac{1}{p\lambda_1}\left(1+\sum_{p=1}^{p_{\max}}p\lambda_1\right)^{p_{\max}}\\
&= \frac{1}{\left(p_{\max}!\right)^2}\left(\frac{1}{\lambda_1}\right)^{p_{\max}}\left(1+\frac{p_{\max}(p_{\max}+1)\lambda_1}{2}\right)^{p_{\max}}\\
&\le \left(\frac{1}{\lambda_1}\right)^{p_{\max}}\underbrace{\frac{(1+p_{\max}(p_{\max}+1)2^{-p_{\max}})^{p_{\max}}}{\left(p_{\max}!\right)^2}}_{F(p_{\max})},
\end{aligned}$$

where the last inequality is obtained by using the definition of $p_{\max}$ in Lemma A.1 to bound $\lambda_1/2 \le 2^{-p_{\max}}$.

To complete the proof, we bound the function $F(p_{\max})$ from above. The numerator of $F(p_{\max})$ converges to 1 as $p_{\max}\to\infty$, as seen by taking its logarithm and bounding it:

$$\begin{aligned}
&p_{\max}\log\left(1+p_{\max}(p_{\max}+1)2^{-p_{\max}}\right)\\
&\qquad\qquad\le p_{\max}^2(p_{\max}+1)2^{-p_{\max}}\to 0 \quad \text{as } p_{\max}\to\infty.
\end{aligned}$$

Thus $F(p_{\max})$ decreases to zero as $p_{\max}$ increases. Numerical evaluation shows that $F(p_{\max})$ attains a maximum value of 2 at $p_{\max}=1$. □

Lemma A.4. *The maximum number of DNF rules with a given signature* $\mathbf{K}\in\Delta$ *is bounded as*

$$\max_{\mathbf{K}\in\Delta}|\mathcal{C}(\mathbf{K})| < (2d)^{1/\lambda_1}.$$

Proof. The number of conjunctions of degree p is $\binom{d}{p}2^p$, where the factor of 2^p is due to there being two choices of conditions on each of the p selected variables. The number of DNF rules with signature $\mathbf{K}$ is therefore

$$|\mathcal{C}(\mathbf{K})| = \prod_{p=1}^{p_{\max}}\binom{\binom{d}{p}2^p}{K_p} < \prod_{p=1}^{p_{\max}}\frac{\left(\binom{d}{p}2^p\right)^{K_p}}{K_p!}.$$

Taking logarithms, we obtain

$$\begin{aligned}
&\max_{\mathbf{K}\in\Delta} \log|\mathcal{C}(\mathbf{K})| < \\
&\quad \max_{\mathbf{K}} \sum_{p=1}^{p_{\max}} K_p \log\left(\binom{d}{p}2^p\right) - \log(K_p!) \\
&\quad \text{s.t.} \sum_{p=1}^{p_{\max}} K_p(\lambda_0 + p\lambda_1) \le 1.
\end{aligned} \tag{A.11}$$

For simplicity, we drop the nonlinear term $-\log(K_p!) \le 0$. The right-hand side of (A.11) then becomes a maximization of a linear function over a simplex. The maximum value is given by

$$\max_{p=1,\dots,p_{\max}} \frac{\log\left(\binom{d}{p}2^p\right)}{\lambda_0 + p\lambda_1} \tag{A.12}$$

(attained by setting $K_{p^*} = 1/(\lambda_0 + p^*\lambda_1)$ for a maximizing value p^* and $K_p = 0$ otherwise). Again for simplicity, we further bound (A.12) from above by dropping λ_0 from the denominator, resulting in

$$\max_{\mathbf{K}\in\Delta} \log|\mathcal{C}(\mathbf{K})| < \frac{1}{\lambda_1} \max_{p=1,\dots,p_{\max}} \frac{1}{p}\log\binom{d}{p} + \log 2$$

(otherwise (A.12) may require solving a transcendental equation). Since $\log\binom{d}{p}$ increases sublinearly with p, the maximum occurs at $p = 1$, yielding the desired result. □

By combining (A.9), Lemmas A.3 and A.4, we obtain the desired bound on the number of DNF rules satisfying the optimality conditions in Lemmas A.1 and A.2.

Lemma A.5. *Under the assumptions of Lemma A.1, the number of DNF rules satisfying the necessary conditions of optimality in Lemmas A.1 and A.2 is bounded as*

$$|\mathcal{C}| < 2(2d)^{1/\lambda_1}\left(\frac{1}{\lambda_1}\right)^{p_{\max}}.$$

A.4.3 Proof of Theorem 3.1

We prove the theorem in two steps, first relating the empirical estimator in (3.7) to a problem intermediate between (3.5) and (3.7),

$$\begin{aligned} \mathcal{S}^* := \arg\min_{\mathcal{C}} \;\; & Q(\mathcal{C}) := \bar{V}(\mathcal{C}) + R(\mathcal{C}) \\ \text{subject to} \;\; & \sum_{i\in\mathcal{I}} \mathbb{1}[x_i \in \mathcal{C}] \geq \alpha m, \end{aligned} \tag{A.13}$$

and then relating this intermediate problem (A.13) to (3.5). Problem (A.13) has the same regularized volume objective as (3.5) but with the empirical probability constraint of (3.7).

For the first step, let $\hat{V}(\mathcal{C})$ denote the empirical volume in (3.7) (i.e. the first term in the objective function). As noted in Section 3.4.1, $\hat{V}(\mathcal{C})$ is a scaled binomial random variable with n trials and mean $\bar{V}(\mathcal{C})$. Hoeffding's inequality thus provides the following tail bound:

$$\Pr\Big(\big|\hat{V}(\mathcal{C}) - \bar{V}(\mathcal{C})\big| > \epsilon_n\Big) \leq 2e^{-2n\epsilon_n^2}.$$

Defining $\hat{Q}(\mathcal{C}) = \hat{V}(\mathcal{C}) + R(\mathcal{C})$ and recalling that $Q(\mathcal{C}) = \bar{V}(\mathcal{C}) + R(\mathcal{C})$, the same bound holds for the difference $\hat{Q}(\mathcal{C}) - Q(\mathcal{C})$. Taking the union bound over the hypothesis class $\mathfrak{C}$ yields

$$\Pr\Big(\exists \mathcal{C} \in \mathfrak{C} : \big|\hat{Q}(\mathcal{C}) - Q(\mathcal{C})\big| > \epsilon_n\Big) \leq 2|\mathfrak{C}|e^{-2n\epsilon_n^2}. \tag{A.14}$$

Assuming that the event in (A.14) is not true, we obtain the following sequence of bounds, where the second inequality is due to the optimality of $\hat{\mathcal{S}}$ in (3.7):

$$Q(\hat{\mathcal{S}}) \leq \hat{Q}(\hat{\mathcal{S}}) + \epsilon_n \leq \hat{Q}(\mathcal{S}^*) + \epsilon_n \leq Q(\mathcal{S}^*) + 2\epsilon_n. \tag{A.15}$$

For this to hold with probability at least $1 - \delta$, we set δ equal to the right-hand side of (A.14) to obtain

$$\epsilon_n = \sqrt{\frac{\log(2|\mathfrak{C}|/\delta)}{2n}}. \tag{A.16}$$

For the second step, we observe that the empirical probability $\hat{P}(\mathcal{C}) = \frac{1}{m}\sum_{i\in\mathcal{I}} \mathbb{1}[x_i \in \mathcal{C}]$ is also a scaled binomial random variable, this time with m trials and mean $P(\mathcal{C})$. We thus have a similar bound as in (A.14),

$$\Pr\big(\exists \mathcal{C} \in \mathfrak{C} : \big|\hat{P}(\mathcal{C}) - P(\mathcal{C})\big| > \epsilon_m\big) \leq 2|\mathfrak{C}|e^{-2m\epsilon_m^2},$$

and setting the right-hand side equal to δ yields the same expression for ϵ_m as in (A.16) with n replaced by m. We then use Theorem 3 and Corollary 12 in (Scott and Nowak, 2006) to conclude that with probability at least $1-\delta$,

$$Q(\mathcal{S}^*) \leq q^*(\alpha + \epsilon_m) \quad \text{and} \quad P(\mathcal{S}^*) \geq \alpha - \epsilon_m.$$

Indeed, since $\hat{S} \in \mathfrak{C}$ and satisfies the constraint $\hat{P}(\hat{\mathcal{S}}) \geq \alpha$ as well, the above may be changed to

$$Q(\mathcal{S}^*) \leq q^*(\alpha + \epsilon_m) \quad \text{and} \quad P(\hat{\mathcal{S}}) \geq \alpha - \epsilon_m. \tag{A.17}$$

Combining (A.15) and (A.17) gives

$$Q(\hat{\mathcal{S}}) \leq q^*(\alpha + \epsilon_m) + 2\epsilon_n \quad \text{and} \quad P(\hat{\mathcal{S}}) \geq \alpha - \epsilon_m$$

with probability at least $1 - 2\delta$.

Lastly, we use Lemma A.5 to bound ϵ_n from above by

$$\sqrt{\frac{\lambda_1^{-1}\log(2d) + p_{\max}\log\lambda_1^{-1} + \log(4/\delta)}{2n}}$$

and similarly for ϵ_m.

A.5 Generalization of the product estimator

Below, we give a Theorem bounding the expected error of the two-stage estimate $\hat{\mathcal{O}} = \hat{\mathcal{S}} \cap \hat{\mathcal{B}}$ as a function of the error of the base estimators $\hat{\mathcal{S}}, \tilde{\mathcal{B}}$. This justifies the two-stage nature of our algorithm and motivates selecting hyperparameters for overlap rules $\hat{\mathcal{B}}$ based on the error with respect to the base estimator $\tilde{\mathcal{B}}$. Before we state the result, we give a Lemma bounding the error of an estimator of a product of functions in terms of estimators of the respective terms in the product.

Consider the task of predicting the binary deterministic label $g(X) = g_1(X)g_2(X)$ by approximating the product of estimators f_1, f_2 of g_1, g_2. Now, let $R_g(f)$ denote the expected zero-one loss of f with respect to g over p,

$$R_g(f) = \mathbb{E}_{X\sim p}[\mathbb{1}[f(x) \neq g(x)]] \ .$$

Lemma A.6. *For f_1 and f_2 such that $R_{g_1}(f_1) \leq A \leq \min\{p(f_2(X) = 1), p(g_2(X) = 1)\}$, $R_{g_2}(f_2) \leq B \leq \min\{p(f_1(X) = 1), p(g_1(X) = 1)\}$ and $\max\{A + B, C\} \leq 1/2$, let $f(X)$ approximate $f_1(X)f_2(X)$ and assume that $R_{f_1 f_2}(f) \leq C$. Then,*

$$R_g(f) \leq A + B + C$$

Proof. For convenience, let $f_1 = f_1(X), g_1 = g_1(X)$, et cetera, and let $\gamma = p(g(X) = 1)$.

$$\begin{aligned}
&R_g(f_1 f_2) = p(f_1 f_2 \neq g_1 g_2) \\
&= p(f_1 = f_2 = 1 \wedge (g_1 = 0 \vee g_2 = 0)) \\
&+ p((f_1 = 0 \vee f_2 = 0) \wedge g_1 = g_2 = 1) \\
&\leq p(f_1 = f_2 = 1 \wedge g_1 = 0) + p(f_1 = f_2 = 1 \wedge g_2 = 0) \\
&+ p(g_1 = g_2 = 1 \wedge f_1 = 0) + p(g_1 = g_2 = 1 \wedge f_2 = 0) \\
&\leq \min\{p(h_2 = 1), p(f_1 = 1 \wedge g_1 = 0)\} \\
&+ \min\{p(f_1 = 1), p(f_2 = 1 \wedge g_2 = 0)\} \\
&+ \min\{p(g_2 = 1), p(g_1 = 1 \wedge f_1 = 0)\} \\
&+ \min\{p(g_1 = 1), p(g_2 = 1 \wedge f_2 = 0)\} \\
&\leq A + B
\end{aligned}$$

In the first inequality, we use the standard Frechet inequalities. In the second and third, we use the assumptions in the statement. Alternatively, we could arrive at the same result by assuming that h_2 and (f_1, h_1) as well as h_1 and (f_2, h_2) are independent and decomposing the joint distributions. This could be guaranteed by sample splitting. We could then remove the assumption that the marginal probability of the label is larger than the error. In either case,

$$
\begin{aligned}
R_g(f) &= p(f = f_1 f_2 \wedge f_1 f_2 \neq g) \\
&\quad + p(f \neq f_1 f_2 \wedge f_1 f_2 = g) \\
&\leq \min\{p(f = f_1 f_2), p(f_1 f_2 \neq g)\} \\
&\quad + \min\{p(f \neq f_1 f_2), p(f_1 f_2 = g)\} \\
&= p(f_1 f_2 \neq g) + p(f \neq f_1 f_2) \\
&\leq A + B + C \ .
\end{aligned}
$$

□

We now state our result. First, we view membership in $\hat{\mathcal{O}} = \hat{\mathcal{S}} \cap \hat{\mathcal{B}}$ as given by an instance of the hypothesis class $\mathcal{F} = \{f(x) := \mathbb{1}[x \in \hat{\mathcal{S}}]h(x); h \in \mathcal{H}\}$, for some function family $\mathcal{H}$. Then, let $R_g(f) = \mathbb{E}_{X\sim p}[\mathbb{1}[f(x) \neq g(x)]]$ denote the expected risk of f with respect to g over p, and $\hat{R}_g(f) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[f(x_i) \neq g(x_i)]$ the empirical risk.

Theorem A.1. *Given are classifiers $\hat{s}, \tilde{b}$ of support membership s and propensity boundedness b, with overlap defined as $o(x) = s(x)b(x)$, such that for all $n > N$ it holds for $A_n, C_n \in \tilde{\mathcal{O}}(1/\sqrt{n})$ with $\max\{A_n, C_n\} \leq 1/4$ that $R_s(\hat{s}) \leq A_n, R_b(\tilde{b}) \leq C_n$. Then, for any function $\hat{o} \in \mathcal{H}$ approximating $\hat{s} \cdot \tilde{b}$, with probability larger than $1 - \delta$,*

$$
R_o(\hat{o}) \leq \hat{R}_{\hat{s}\cdot\tilde{b}}(\hat{o}) + \frac{D_{\mathcal{F},\delta,n}}{\sqrt{n}} + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right) \ ,
$$

where $D_{\mathcal{F},\delta,n} = \sqrt{8d(\log\frac{2m}{d} + 1) + 8\log\frac{4}{\delta}}$, with d the VC-dimension of $\mathcal{F}$ and $\tilde{O}$ hides logarithmic factors.

Proof. From Lemma A.6 and assumptions, we have that

$$R_o(\hat{o}) \leq R_{\hat{s}\cdot\tilde{b}}(\hat{o}) + R_s(\hat{s}) + R_b(\tilde{b}) \leq R_{\hat{s}\cdot\tilde{b}}(\hat{o}) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right) \ .$$

By applying standard VC-theory w.r.t. $\mathcal{F}$, we have our result. □

Theorem A.1 bounds the generalization error of (e.g., Boolean rule) approximations of $\sqrt{n}$-consistent base estimators. It may be generalized to other rates, but convergence at *some* rate is necessary for consistency of the final estimator. Critically, the bias incurred by the approximation is observable and may be traded off for interpretability.

Table A.6: *Population averages for covariates in Opioids in order of difference between the overlapping and non-overlapping set. DMME, MME and Duration are the medians of daily MME, total MME and prescription duration days in each group.*

	Total	**DMME**	**MME**	**Duration**
Total sample	35106	46	225	5
Male	9301	50	300	5
Female	25805	45	225	5
Age groups				
<15	847	20	100	5
15-24	3334	45	200	5
25-34	9994	45	210	4
35-44	6820	46	225	5
45-54	6196	50	250	5
55-64	7915	50	300	5
>=65	0	0	0	0
Surgery type				
Auditory	29	18	135	6
Cardiovascular	3633	45	270	5
Integumentary	1507	48	225	5
Mediastinum	54	47	300	5
Female genital	3913	48	225	5
Hemic	885	50	225	5
Respiratory	665	45	250	5

Endocrine	214	45	200	5
Nervous	4350	60	375	6
Urinary	1476	45	225	5
Musculoskeletal	6678	60	450	7
Maternity	13553	45	200	4
Male genital	585	45	225	5
Year				
2011	7547	45	225	5
2012	10743	46	225	5
2013	9651	50	225	5
2014	7165	45	225	5
Diagnosis history (until day before surgery)				
Other specified gastritis: without mention of hemorrhage	491	42	225	5
Other ascites	233	45	225	5
Lumbosacral spondylosis without myelopathy	1135	60	400	6
Nausea with vomiting	1914	45	225	5
Other respiratory abnormalities	1935	45	225	5
Vomiting alone	765	45	200	5
Myalgia and myositis: unspecified	1522	50	250	5
Attention deficit disorder with hyperactivity	370	45	225	5
Attention deficit disorder without mention of hyperactivity	444	45	225	5
Depressive disorder: not elsewhere classified	2221	50	225	5
Dysthymic disorder	752	50	225	5
Tachycardia: unspecified	631	45	225	5
Degeneration of cervical intervertebral disc	904	56	337	6
Flatulence: eructation: and gas pain	427	45	225	5
Generalized anxiety disorder	833	45	225	5
Other symptoms referable to back	368	50	300	5
Cellulitis and abscess of leg: except foot	450	45	225	5
Constipation: unspecified	1136	45	225	5
Thoracic or lumbosacral neuritis or radiculitis: unspecified	1676	60	326	6
Anxiety state: unspecified	2205	50	225	5
Lumbago	4559	50	250	5
Abdominal pain: generalized	1607	45	225	5
Degeneration of lumbar or lumbosacral intervertebral disc	1542	60	388	6

Other and unspecified noninfectious gastroenteritis and colitis	1254	45	225	5
Major depressive affective disorder: recurrent episode: moderate	507	45	225	5
Asthma: unspecified type: unspecified	2044	45	225	5
Arthrodesis status	178	60	450	7
Chest pain: unspecified	4701	45	225	5
Routine general medical examination at a health care facility	9529	50	225	5
Diarrhea	1714	50	225	5
Fitting and adjustment of vascular catheter	318	45	225	5
Hypopotassemia	721	45	225	5
Bariatric surgery status	302	40	200	5
Sprain of neck	816	50	225	5
Unspecified gastritis and gastroduodenitis: w/o mention of hemorrhage	960	45	225	5
Injury of face and neck	271	46	300	5
Backache: unspecified	2471	50	225	5
Unspecified septicemia	222	45	225	5
Acute pharyngitis	4219	45	225	5
Acute bronchitis	3311	46	225	5
Abdominal pain: other specified site	2890	45	225	5
Atrophic gastritis: without mention of hemorrhage	537	45	225	5
Cough	3946	45	225	5
Altered mental status	202	45	225	5
Cervicalgia	2758	50	250	5
Abdominal pain: unspecified site	6339	45	225	5
Other chronic pain	346	56	300	6
Headache	3514	45	225	5
Tobacco use disorder	1834	50	225	5
Other screening mammogram	5722	50	240	5
Observation and evaluation for other specified suspected conditions	337	45	225	5
Unspecified sinusitis (chronic)	1624	46	225	5
Rheumatoid arthritis	353	50	300	5
Brachial neuritis or radiculitis NOS	1147	50	300	5
Loss of weight	455	46	225	5
Hypersomnia with sleep apnea: unspecified	424	42	225	5
Insomnia: unspecified	968	50	225	5
Other malaise and fatigue	5178	46	225	5
Other injury of chest wall	210	50	300	5

Dehydration	841	45	225	5
Acute respiratory failure	120	40	225	5

Table A.7: *Population averages for the 156 features in the UTI cohort. Mean values and total (for binary features) are given, and there are 64593 subjects in total.*

	Mean	**Total**
Demographics		
Age	55.1	
Male	16.53%	10685
White	72.17%	46662
Veteran	4.61%	2981
Current Location		
Outpatient	64.89%	41957
Emergency Room	15.69%	10142
Inpatient	17.26%	11159
Intensive Care Unit (ICU)	2.69%	1736
Local Resistance Rates (Past 30-90 days, at this location)		
Trimethoprim/Sulfamethoxazole	18.61%	
Nitrofurantoin	19.85%	
Ciprofloxacin	22.70%	
Levofloxacin	24.19%	
Secondary Site of Infection		
Skin / Soft Tissue	0.20%	132
Blood	1.59%	1031
Respiratory Tract	0.53%	341
Nasal or Rectal Swab	0.19%	124
Medical History (Past 90 Days)		
Alcohol abuse	1.66%	1074
Deficiency anemia	2.84%	1837

Cardiac arrhythmias	17.08%	11041
Blood loss anemia	0.49%	315
Congestive heart failure	10.16%	6571
Coagulopathy	3.81%	2466
Diabetes, uncomplicated	14.13%	9135
Diabetes, complicated	5.00%	3232
Depression	11.80%	7627
Drug abuse	1.72%	1114
Fluid and electrolyte disorders	13.84%	8946
AIDS/HIV	0.43%	281
Hypertension, uncomplicated	32.51%	21017
Hypertension, complicated	5.43%	3513
Hypothyroidism	7.86%	5085
Liver disease	4.36%	2822
Lymphoma	1.63%	1051
Metastatic cancer	5.50%	3559
Other neurological disorders	6.68%	4319
Obesity	6.70%	4332
Pulmonary circulation disorders	3.13%	2025
Peptic ulcer disease, excluding bleeding	0.61%	393
Peripheral vascular disorders	5.68%	3672
Paralysis	3.08%	1992
Psychoses	2.42%	1563
Chronic pulmonary disease	11.29%	7299
Renal	8.87%	5735
Rheumatoid arthritis / collagen vascular diseases	3.76%	2428
Solid tumor without metastasis	12.00%	7760
Valvular disease	7.79%	5034
Weight loss	3.59%	2319
Preganant	3.08%	1989
Previous Care (Past 90 days)		

Inpatient Stay	18.38%	11882
Nursing Home Stay	1.20%	779
Previous Procedures (Past 90 days)		
Central Venous Catheder	5.27%	3410
Hemodialysis	0.66%	427
Mechanical Ventilation	5.74%	3714
Parenteral Nutrition	0.67%	434
Surgery	59.84%	38689
Previous Organisms (Past 90 days)		
Citrobacter species	0.42%	270
Coagulate negative Staphylococcus species	1.15%	741
Enterobacter aerogenes	0.15%	95
Escherichia coli	7.82%	5057
Enterococcus species	2.66%	1718
Enterobacter cloacae	0.29%	186
Group B Streptococcus	0.17%	109
Klebsiella pneumoniae	2.02%	1307
Morganella species	0.11%	73
Pseudomonas aeruginosa	0.92%	594
Proteus species	0.69%	445
Staph aureus	1.55%	1003
Serratia species	0.22%	145
Previous Resistance, measured by culture (Last 90 Days)		
Amoxicillin Clavulanate	2.34%	1511
Amikacin	0.10%	67
Ampicillin	7.44%	4808
Aztreonam	0.95%	616
Ceftazidime	0.30%	197
Cefazolin	9.22%	5962
Chlorampenicol	0.17%	111
Ciprofloxacin	4.62%	2984

Clindamycin	0.97%	624
Ceftriaxone	1.24%	804
Doxycycline	0.39%	249
Ertapenem	0.14%	88
Erythromycin	3.71%	2399
Cefepime	0.54%	351
Cefoxitin	0.49%	319
Gentamicin	1.65%	1066
Gentamicin (Synergistic)	0.47%	307
Imipenem	0.47%	303
Levofloxacin	5.32%	3439
Linezolid	0.09%	58
Meropenem	0.13%	85
Moxifloxacin	0.86%	556
Nalidixic Acid	0.09%	60
Nitrofurantoin	4.06%	2628
Oxacillin	1.79%	1158
Penicillin	2.41%	1559
Piperacillin	0.62%	402
Polymyxin B	1.22%	790
Rifampin	0.80%	518
Ampicillin Sulbactam	1.63%	1056
Streptomycin (Synergistic)	0.23%	150
Trimethoprim Sulfamethoxazole	3.10%	2006
Tetracycline	5.33%	3443
Ticarcillin	0.24%	153
Tobramycin	0.31%	203
Piperacillin Tazobactam	0.53%	341
Vancomycin	0.92%	598
Previous Antibiotic Prescription (Last 90 Days)		
Amikacin	0.09%	60

Amoxicillin	2.47%	1596
Amoxicillin/Clavulanate	2.15%	1388
Amphotericin B	0.16%	102
Ampicillin/Sulbactam	0.34%	217
Azithromycin	2.86%	1847
Aztreonam	0.25%	159
Cefadroxil	0.15%	96
Cefazolin	4.87%	3150
Cefepime	2.30%	1489
Cefixime	0.26%	166
Cefotetan	0.18%	114
Cefoxitin	0.25%	161
Cefpodoxime	0.88%	570
Ceftazidime	0.73%	475
Ceftriaxone	2.75%	1775
Cefuroxime	0.24%	156
Cephalexin	2.31%	1496
Ciprofloxacin	11.09%	7170
Clarithromycin	0.35%	226
Clindamycin	1.84%	1187
Daptomycin	0.10%	63
Dicloxacillin	0.19%	126
Doxycycline	1.73%	1119
Ertapenem	0.22%	140
Erythromycin	0.39%	249
Fluconazole	3.56%	2301
Fosfomycin	0.36%	232
Gentamicin	0.94%	607
Imipenem	0.33%	216
Levofloxacin	5.94%	3838
Linezolid	0.73%	470
Meropenem	0.40%	256

Metronidazole	4.49%	2906
Micafungin	0.24%	154
Minocycline	0.20%	129
Moxifloxacin	0.27%	174
Nafcillin	0.24%	157
Nitrofurantoin	2.73%	1767
Norfloxacin	4.25%	2749
Penicillin	0.31%	199
Piperacillin	0.41%	268
Piperacillin/Tazobactam	0.23%	148
Polymyxin B	0.52%	333
Posaconazole	0.18%	118
Tetracycline Metronidazole	0.09%	59
Trimethoprim	0.12%	79
Trimethoprim/Sulfamethoxazole	3.96%	2558
Vancomycin	8.80%	5690
Vancomycin Gentamicin	3.35%	2165

Appendix B

Appendix for Chapter 4

B.1 When can biased estimators be falsified?

As discussed in Examples 4.2 and 4.3, we imagine that observational estimators differ in a few possible ways. They may represent the same identification strategy applied to different datasets, different identification strategies applied to the same dataset (e.g., different choices of confounders), or some combination of the two.

Assumption 4.3 states that there exists a consistent and asymptotically normal observational estimator for τ, as defined in Def. 4.1. This is a fundamental assumption in our work, and so we build additional intuition for when we might expect this condition to hold, and when we might be able to falsify this assumption. In this section, we give basic intuition regarding patterns of confounding, and in Section B.2, we discuss issues of transportability.

In Example B.1.1, we give a simple example where the causal graph is consistent across two subgroups, and where an estimator must control for all confounders to get consistent estimates of the GATE in either subgroup. In this setting, falsification is possible. On the other hand, in Example B.1.2, we give a counterexample, where there are multiple estimators that can deliver consistent estimates of the GATE on the RCT subpopulation, but only one provides consistent estimates across all subpopulations.

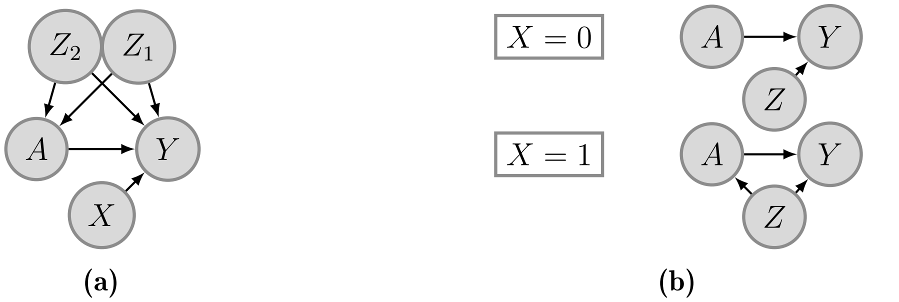


Figure B-1: *Example B.1.1 is depicted in (a), and Example B.1.2 in (b)*

Example B.1.1 (Consistent confounding across subgroups)**.** In the causal graph shown in Figure B-1a, there are two sets of confounders, Z_1, Z_2, a binary treatment variable A, a binary subgroup variable X, and the outcome Y. We assume a linear outcome model, whereby $E[Y|X, Z_1, Z_2, A] = \alpha + \beta X + \gamma_1 AX + \gamma_2 A(1 - X) + \delta_1 Z_1 + \delta_2 Z_2$. Note that the true group average treatment effect (GATE) for the two subgroups are, $\text{GATE}(X = 0) = \gamma_2; \text{GATE}(X = 1) = \gamma_1$. It is straightforward to show that not conditioning on the full set of confounders will lead to biased GATE estimates for both subgroups, whereas conditioning on both Z_1 and Z_2 will lead to consistent estimates for both subgroups.

Example B.1.2 (Selective confounding by subgroup)**.** Let there be two subgroups, $X = 0$ and $X = 1$, with the former having support in both RCT and observational studies and the latter having support in only observational data. Now, suppose we had the following treatment assignment mechanism, $p(A = 1|X, Z) = f(Z) \cdot \mathbf{1}(X = 1) + c \cdot \mathbf{1}(X = 0)$, where Z is a set of confounders, f is a nonlinear function of Z, and c is a constant. A candidate estimator that does not condition on Z would be able to get consistent estimates for the validation effect but not the extrapolated effect. On the other hand, conditioning on Z would allow for consistent estimates on both validation and extrapolated effects.

B.2 Conditions for valid observational / randomized comparisons

Recall that we had defined the group average treatment effect (GATE) as follows in Equation (4.1)

$$\tau_i \coloneqq \begin{cases} \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0], & \text{if } i \in \mathcal{I}_R \\ \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 1], & \text{if } i \in \mathcal{I}_O \end{cases}, \tag{B.1}$$

and refer to τ_i for $i \in \mathcal{I}_R$ as a validation effect, and τ_i for $i \in \mathcal{I}_O$ as an extrapolated effect. In this section, we discuss sufficient conditions under which these causal effects are identifiable from observational data drawn from a distribution $D = k$, and give examples of doubly-robust estimators of these quantities. These assumptions cover both comparisons of the observational studies to the randomized trial (used for validation), as well as the normalization of observational estimates (used for confidence intervals on the extrapolated effects).

Our goal in presenting these results is to build intuition in this setting for when we might expect a consistent observational estimator to exist across all groups. This is a well-studied topic, often in the context of generalizing effect estimates from randomized trials to other supported populations (e.g., all trial-eligible individuals). We primarily make use of results in that literature to build intuition here, pointing the reader to Degtiar and Rose (2021) for a recent review whose presentation we largely mirror, with modifications to account for our notation.

B.2.1 Identification

First, we state standard assumptions under which the GATE in the observational population for $D = k$,

$$\mathbb{E}[Y_1 - Y_0 \mid G = i, D = k], \tag{B.2}$$

is identifiable from data in the dataset $D = k$, with notation adapted to our setting.

Assumption B.2.1. The following conditions hold for the distribution $\mathbb{P}(\cdot \mid D = k)$:

1. *Conditional Exchangeability over* A: $Y_a \perp\!\!\!\perp A \mid X, D = k$ for all treatments a.
2. *Positivity of Treatment Assignment*: $\mathbb{P}(X = x \mid D = k) > 0 \implies \mathbb{P}(A = a \mid X = x, D = k) > 0$ for all a.
3. *Consistency*: $A = a \implies Y_a = Y$

These causal assumptions ensure that the ATE and CATE can be identified from observational data for the observational population and are standard in the causal inference literature (Imbens and Rubin, 2015). In order to transport these estimates to the RCT population (or from one observational dataset to another), we require additional assumptions. Next, we give assumptions under which these estimates can be transported to another population $D = k'$, where in our case $k' \in \{0, 1\}$.

Assumption B.2.2. Let k correspond to a source population, and k' correspond to the target population. Conditioned on the event $D \in \{k, k'\}$, define the random variable $S = 1$ if $D = k$ and $S = 0$ otherwise. Then let the following hold, on the distribution $\mathbb{P}(\cdot \mid D \in \{k, k'\})$.

1. *Conditional Exchangeability over* S: $Y_a \perp\!\!\!\perp S \mid X$ for all treatments a.
2. *Positivity of Selection*: $\mathbb{P}(X = x) > 0 \implies \mathbb{P}(S = 1 \mid X = x) > 0$ almost surely over X for all a.
3. *Consistency*: $S = s$ and $A = a \implies Y_a = Y$

Here, we note that this introduces non-trivial additional assumptions. Most notably, we require that the potential outcomes are independent of the dataset, given X. This would be violated, for instance, if the distribution of unobservable effect modifiers differs between different observational studies. As a result, we note that it is possible for an observational study to fail to replicate the RCT results due to failures of transportability (failure of Assumption B.2.2) even if it has "internal validity", allowing for identification of the causal effect in the population $D = k$. There also exists a large body of work on identifying transportable causal effects via causal graphs (Pearl and Bareinboim, 2011, 2014; Pearl, 2015).

B.2.2 Estimation of the ATE in the target population

Regarding estimation, Dahabreh et al. (2020) consider the problem of transporting average treatment effects from randomized trials to observational studies, under Assumption B.2.1 with $k = 0$ and Assumption B.2.2 with $k = 0, k' = 1$. These assumptions admit identification of the potential outcomes means as follows (see Section 4.2 of Dahabreh et al. (2020))

$$\mathbb{E}[Y_a \mid S = 0] = \mathbb{E}[\mathbb{E}[Y \mid X, S = 1, A = a] \mid S = 0] \tag{B.3}$$

where the outer expectations are over $\mathbb{P}(X \mid S = 0)$, i.e., the covariate distribution of the target population. Dahabreh et al. (2019) give a doubly robust estimator for the statistical quantity on the right-hand side as the empirical expectation of the following pseudo-outcome (see Equation A.13 of Dahabreh et al. (2020))

$$\hat{\mu}(a) = \frac{1}{n} \sum_{i=1}^{n} Y_i^a(\hat{\eta}, \hat{\pi}) \tag{B.4}$$

where n is the total samples in both the source $S = 1$ and target $S = 0$ samples, and where

$$Y_i^a(\hat{\eta}, \hat{\pi}) \coloneqq \frac{1}{\hat{\pi}} \left(\mathbf{1}\left\{S_i = 1, A_i = a\right\} \cdot \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i)\hat{e}_a(X_i)} \cdot \{Y_i - \hat{g}_a(X_i)\} + (1 - S_i)\hat{g}_a(X_i) \right). \tag{B.5}$$

In Equation (B.5), $\hat{\eta} \coloneqq (\hat{g}_a, \hat{e}_a, \hat{p})$, and $\hat{\pi} \coloneqq n^{-1} \sum_{i=1}^{n} \mathbf{1}\left\{S_i = 0\right\}$ is an estimate of $\mathbb{P}(S = 0)$, $\hat{g}_a(X)$ is an estimate of the mean conditional outcome $\mathbb{E}[Y \mid A = a, S = 1, X]$, $\hat{p}(X)$ is an estimate of the selection probability $\mathbb{P}(S = 1 \mid X)$, and $\hat{e}_a(X)$ is an estimate of the propensity score $\mathbb{P}(A = a \mid S = 1, X)$. Dahabreh et al. (2019) derives precise asymptotic properties of this estimator, which is asymptotically normal and consistent for the observational quantity on the right-hand side of Equation (B.3). In particular, this estimator is doubly-robust in the sense that it is consistent if either $\hat{p}(X)$ or $\hat{g}_a(X)$ is consistent, but requires consistency of $\hat{e}_a(X)$. It also enjoys the rate double-robustness property, retaining consistency and asymptotic normality even if the estimators for $\hat{p}, \hat{g}$ converge at slower than parametric rates, and allows for the same cross-fitting schemes used in the Double ML (Chernozhukov et al., 2018) literature for relaxing Donsker conditions.

Note that the average treatment effect in this setting can be estimated by the following contrast, which is similarly an empirical expectation of a pseudo-outcome

$$\hat{\mu}(1) - \hat{\mu}(0) = \frac{1}{n}\sum_{i=1}^{n} Y_i^1(\hat{\eta}, \hat{\pi}) - Y_i^0(\hat{\eta}, \hat{\pi}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{\eta}, \hat{\pi}), \tag{B.6}$$

where $Y_i(\hat{\eta}, \hat{\pi}) \coloneqq Y_i^1(\hat{\eta}, \hat{\pi}) - Y_i^0(\hat{\eta}, \hat{\pi})$. Furthermore, the variance of these estimates can be estimated using either sandwich estimators from M-estimation theory (Stefanski and Boos, 2002), or via bootstrap methods. We refer the reader to Sections 5.3, 5.4 and Appendix A.4 of (Dahabreh et al., 2020) for more details.

B.3 Estimation and comparison of GATE in semi-synthetic experiments

In Sections 4.2.2 and B.2, we discuss several estimators for average treatment effects (ATEs) that are known to be asymptotically normal, such as the double ML estimator discussed in Example 4.2 or the doubly-robust estimator in Section B.2.

Given a fixed set of discrete subgroups, one could analyze each subgroup independently and apply such estimators directly, since the ATE in each subgroup is precisely the GATE. This would be a straightforward way to ensure that the same formal guarantees hold regarding asymptotic normality. While this approach would be feasible in our experimental setting, due to the small number of groups, it is less practical in general, especially with a larger number of groups, since information cannot be shared across nuisance models such as $\hat{g}_a, \hat{e}_a, \hat{p}$ discussed in Section B.2.

In an effort to emulate a more realistic setting, we take a slightly different approach in the semi-synthetic experiments. We draw inspiration from the double ML approach given in Semenova and Chernozhukov (2021) for GATE estimation, while taking into consideration the transportation of causal effects in the sense of Section B.2. Note that in Semenova and Chernozhukov (2021), the required assumptions and proofs for asymptotic normality of estimators are provided on a case-by-case basis, which does not include our case with transportation. Therefore, in the following we will briefly describe their approach, then show how we construct our GATE estimators and provide the required assumptions for their asymptotic normality.

Semenova and Chernozhukov (2021) focuses on the setting where there exists some pseudo-outcome / signal, $Y(\eta)$, and where one is interested in summarizing the function, $\tau(x) = \mathbb{E}[Y(\eta) \mid X = x]$, with a linear regression function (in the simplest case, a set of group indicators). When $Y(\eta)$ is the doubly-robust score (Robins et al., 1994; Robins and Rotnitzky, 1995) (see Equation (B.11)), $\tau(x)$ is equal to the CATE function, and the best approximation by group indicators gives the GATE.

Our general procedure is as follows: for estimation of $\hat{\tau}(k)$ and the respective variances, we construct a score function / pseudo-outcome, $\tilde{Y}$, whose empirical conditional expectation (in each group) provides an estimate of the GATE, and whose empirical variance we use as an estimate of the variance. We describe this procedure in more detail below. Throughout, X should be taken to refer to the covariates that are observed in a given observational study.

Comparing Validation Effect Estimates In our simulation setup, all of the observational datasets are drawn from a common distribution, which differs from the RCT distribution, requiring the use of the techniques and assumptions discussed in Section B.2 to estimate the GATE, $\tau_i = \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0]$, using data from the observational distributions.

To generate the observational estimates $\hat{\tau}_i(k), \hat{\sigma}_i^2(k)$ in this setting, we cannot simply take empirical conditional expectation / variance of the score function given in Equation B.6. Rather, the GATE is identified under Assumptions B.2.1 and B.2.2 as a conditional expectation of the score times a correction factor, as discussed in the following proposition.

Proposition B.3.1. *In the setting of Section B.2, under Assumptions B.2.1 and B.2.2, the conditional mean potential outcome in the target distribution is identified as*

$$\mathbb{E}[Y_a \mid S = 0, G = i] = \frac{\mathbb{P}(S = 0)}{\mathbb{P}(S = 0 \mid G = i)} \mathbb{E}[Y^a(\eta_0, \pi_0) \mid G = i], \tag{B.7}$$

where $Y^a(\eta, \pi)$ is defined as in Equation B.8.

$$Y^a(\eta, \pi) \coloneqq \frac{1}{\pi}\left(\mathbf{1}\left\{S = 1, A = a\right\} \cdot \frac{1 - p(X)}{p(X)e_a(X)} \cdot \{Y - g_a(X)\} + (1 - S)g_a(X)\right) \tag{B.8}$$

where $\eta \coloneqq (g_a, e_a, p)$ with true underlying parameters $\eta_0 = (g_{a0}, e_{a0}, p_0)$, $\pi \coloneqq \mathbb{P}(S = 0)$ with true value π_0, $g_a(X) \coloneqq \mathbb{E}[Y \mid A = a, S = 1, X]$, $p(X) \coloneqq \mathbb{P}(S = 1 \mid X)$, and $e_a(X) \coloneqq \mathbb{P}(A = a \mid S = 1, X]$.

A proof is provided in Appendix B.4. Note that this is equivalent to replacing the estimate of $1/\mathbb{P}(S = 0)$ in the score with an estimate of $1/\mathbb{P}(S = 0 \mid G = i)$, before computing the empirical conditional expectations of the score.

Now, for each observational dataset, we construct estimates $\hat{\tau}_i(k), \hat{\sigma}_i^2(k)$ for $i \in \mathcal{I}_R$ as follows:

1. We collect observational samples from the two validation groups {lbw, married} and {hbw, married}, which we denote as $G = 0, G = 1$ respectively. We combine these observational samples with the samples from the RCT, using $S = 0$ to denote RCT samples (the target distribution) and $S = 1$ to denote observational samples.

2. We define our signal for each sample as

$$\tilde{Y}_i(\hat{\eta}, \hat{\pi}_g) := \tilde{Y}_i^1(\hat{\eta}, \hat{\pi}_g) - \tilde{Y}_i^0(\hat{\eta}, \hat{\pi}_g) \tag{B.9}$$

   where we define the modified score $\tilde{Y}_i^a$, in light of Proposition B.3.1, as

$$\tilde{Y}_i^a(\hat{\eta}, \hat{\pi}_g) := \frac{1}{\hat{\pi}_g(G_i)} \left(\mathbf{1}\left\{S_i = 1, A_i = a\right\} \cdot \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i)\hat{e}_a(X_i)} \cdot \{Y_i - \hat{g}_a(X_i)\} + (1 - S_i)\hat{g}_a(X_i) \right), \tag{B.10}$$

   where $\hat{\pi}_g(G_i)$ is defined as an estimate of $\pi_g(G_i) := \mathbb{P}(S = 0 \mid G_i)$, computed using empirical averages.

3. We use 3-fold cross-fitting as described in Semenova and Chernozhukov (2021) to generate the signals for each sample, such that for the i-th datapoint, the score $\tilde{Y}_i(\hat{\eta}, \hat{\pi})$ uses plug-in estimates $\hat{\eta} = (\hat{g}_1, \hat{g}_0, \hat{e}_1, \hat{p})$ that are learned on the folds that do not include the i-th datapoint, and $\hat{\pi}$ is estimated using empirical averages. In practice, we use a multi-layer perceptron (MLP) regressor for estimating $\hat{g}_a$, and ℓ_2-regularized logistic regression for estimating $\hat{e}_1, \hat{p}$, with hyperparameters described in Section B.6. For each model, we reserve 20% of the current fold in the cross fitting procedure as a validation set to do hyperparameter selection.

4. Finally, we estimate $\hat{\tau}_i(k)$ as the empirical average $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g) \mid G = i]$, and we use the empirical conditional variance of this score to estimate the variance $\hat{\sigma}_i^2(k)$.

We construct the RCT estimate $\hat{\tau}_i(0)$ (using the RCT sample alone) as the difference of the empirical conditional means $\mathbb{E}_{N_0}[Y \left(\frac{\mathbf{1}\{A=1\}}{\hat{P}(A=1)} - \frac{\mathbf{1}\{A=0\}}{1-\hat{P}(A=1)} \right) \mid G = i]$, where $\hat{P}(A = 1)$ is an empirical average. We compute $\hat{\sigma}_i^2(0)$ as the empirical conditional variance of this quantity. We then conduct testing, as described in Algorithm 1.

Asymptotic normality of transported estimators We herein provide sufficient assumptions that guarantee the asymptotic normality of our transported GATE estimators, i.e. the empirical average $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$:

Assumption B.3.1 (Observational dataset covers the whole support of covariates)**.**

$$\inf_{x \in \mathcal{X}} p_0(x) = \varepsilon_p > 0$$

Note that Assumption B.3.1 is implied by Assumption 4.1.

Assumption B.3.2 (Bounded within-subgroup variance of conditional treatment effects in the RCT)**.**

$$\sup_{x \in \mathcal{X}} var[g_{10}(x) - g_{00}(x)|G = i, S = 0] = \sigma^2_{\tau i} < \infty$$

Assumption B.3.3 (Overlap between treatments in the observational dataset)**.**

$$\inf_{x \in \mathcal{X}} \min(e_{00}(x), e_{10}(x)) = \varepsilon_e > 0$$

Assumption B.3.4 (Finite outcome conditional variance in the observational dataset)**.**

$$\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} \mathbb{E}[(Y - g_{a0}(x))^2 | X = x, S = 1, A = a] = \bar{\sigma}^2 < \infty$$

Assumption B.3.5 (Properties of the nuisance function estimators)**.** Let $\hat{\eta}_{(n)}$ be a sequence of estimators for η indexed by the size of the cross-fitting training fold n. We assume that there exists

- $\epsilon_n = o_P(1)$, a sequence of positive numbers
- $\mathcal{T}_n$, a sequence of nuisance function vector sets in the neighborhood of $\eta_0 = (g_{10}, g_{00}, e_{10}, p_0)$ satisfying $\mathbb{P}(\hat{\eta}_{(n)} \in \mathcal{T}_n) \geq 1 - \epsilon_n$
- $\mathbf{g}_n, \mathbf{e}_n, \mathbf{p}_n$, sequences of worst root mean square errors for the nuisance functions

g_1, g_0, e_1, p, defined as follows:

$$\mathbf{g}_n := \max_{a \in \{0,1\}} \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[g_a(X) - g_{a0}(X)]^2}$$
$$\mathbf{e}_n := \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[e_1(X) - e_{10}(X)]^2}$$
$$\mathbf{p}_n := \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[p(X) - p_0(X)]^2}$$

so that the following assumptions hold:

Assumption A: (Rate of nuisance error)

$$\mathbf{g}_n \vee \mathbf{e}_n \vee \mathbf{p}_n = o_P(1)$$

Assumption B: (Rate of nuisance error product)

$$\sqrt{n}\mathbf{g}_n(\mathbf{e}_n \vee \mathbf{p}_n) = o_P(1)$$

Assumption C: (Bounded nuisance estimates)

$$\sup_{\eta \in \cup_{n=1}^{\infty} \mathcal{T}_n} \left(\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} |g_a(x)| \vee \sup_{x \in \mathcal{X}} \left| \frac{1}{p(x)} \right| \vee \sup_{x \in \mathcal{X}} \left| \frac{1}{e_1(x)} \right| \vee \sup_{x \in \mathcal{X}} \left| \frac{1}{e_0(x)} \right| \right) = \bar{\mathcal{C}} < \infty$$

Theorem B.3.1. *Suppose Assumptions B.3.1 to B.3.5 hold. Then, the empirical average, $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$, where $\tilde{Y}$ is defined in Equation B.9 and $\hat{\eta}$ is estimated with cross-fitting, is asymptotically normal.*

Remark: As we will prove later in section B.4.2, Assumptions B.3.1 to B.3.4 guarantee that when the nuisance function vector $(g_{10}, g_{00}, e_{10}, p_0)^\top$ is known (i.e. need not be estimated), the transported GATE estimator is asymptotically normal. In practice, $(g_{10}, g_{00}, e_{10}, p_0)^\top$ is not known and has to be estimated, so Assumption B.3.5 lays out sufficient properties the nuisance function vector estimator needs to satisfy. In particular, Assumptions B.3.5.A and B.3.5.B permit that the convergence rate of estimators can be slower than $o_P(n^{-1/2})$, which is useful when X is high-dimensional and machine learning models are required to estimate the nuisance functions. To date, a variety of commonly-used machine learning models have been shown to enjoy a convergence rate of at least $o_P(n^{-1/4})$, e.g. Bühlmann and Van De Geer (2011); Belloni et al. (2011b,a) for certain ℓ_1 penalized models, Wager and Walther (2015) for a class of regression trees and random forests, and Chen and White (1999) for a class of neural nets. This implies when these models are applied to the estimation of $(g_{10}, g_{00}, e_{10}, p_0)^\top$, Assumptions B.3.5.A and B.3.5.B hold, so our transported GATE estimator is asymptotically normal and Assumption 4.4 is satisfied.

Constructing Confidence Intervals for the Extrapolated Effects In our experimental setup, the data generating distribution for all observational studies is identical, so no transportation of effects is required, which enables the application of existing results. We use the doubly-robust score (Robins et al., 1994; Robins and Rotnitzky, 1995) as the signal for the conditional average treatment effect,

$$Y(\eta) = \mu(1, X) - \mu(0, X) + \frac{A(Y - \mu(1, X))}{s(X)} - \frac{(1 - A)(Y - \mu(0, X))}{1 - s(X)}, \qquad \text{(B.11)}$$

where $\eta \coloneqq (\mu, s)$, and $\mu(A, X) \coloneqq \mathbb{E}[Y|A, X]$, and $s(X) \coloneqq \mathbb{P}(A = 1 \mid X)$. We use a multi-layer perceptron (MLP) regressor as a plug-in estimate $\hat{\mu}$ of μ, and ℓ_2-regularized logistic regression as a plug-in estimate $\hat{s}$ of s, with hyperparameters described in Section B.6.

Following example 2.2 from Semenova and Chernozhukov (2021), we approximate the conditional treatment effect with a linear combination of subgroup dummy variables $G = (G_0, G_1, G_2, G_3)^\top$, so the combination weights correspond to the GATEs $\tau(k) = (\tau(k)_0, \tau(k)_1, \tau(k)_2, \tau(k)_3)$. This amounts to regressing the estimated signal $\hat{Y}_i(\hat{\eta})$ with G. As long as the propensity score is bounded above and below away from 0 and 1 (Assumption 4.10(a) of Semenova and Chernozhukov (2021)), and the convergence rates of the response surface and propensity score estimates are sufficiently fast (Assumption 4.11), Corollary 4.1 and a set of mild technical conditions justify Theorem 3.1 in

Semenova and Chernozhukov (2021), which gives a result on pointwise asymptotic normality for the regression coeffcients $\hat{\tau}(k) = (\hat{\tau}(k)_0, \hat{\tau}(k)_1, \hat{\tau}(k)_2, \hat{\tau}(k)_3) \in \mathbb{R}^4$, so that for any unit vector $\gamma \in \mathbb{R}^4$ where $\|\gamma\| = 1$,

$$\lim_{N_k \to \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{N_k}\gamma^\top(\hat{\tau}(k) - \tau(k))}{\sqrt{\gamma^\top \Omega \gamma}} < t \right) - \Phi(t) \right| = 0$$

where Ω can be consistently estimated with Equation 2.5 in Semenova and Chernozhukov (2021)

$$\hat{\Omega} = \left(\frac{1}{N_k} \sum_j G_j G_j^\top \right)^{-1} \left(\frac{1}{N_k} \sum_j G_j G_j^\top (\hat{Y}_j(\hat{\eta}) - G_j^\top \hat{\tau}(k))^2 \right) \left(\frac{1}{N_k} \sum_j G_j G_j^\top \right)^{-1}$$

Setting γ as 1 in the $(i+1)$th element and 0 elsewhere thus yields

$$\lim_{N_k \to \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))}{\sqrt{\Omega_{ii}}} < t \right) - \Phi(t) \right| = 0$$

We therefore estimate $\hat{\sigma}_i^2(k)$, the variance of $\hat{\tau}_i(k)$, with $\hat{\Omega}_{ii}$, and as this converges in probability to Ω_{ii}, the asymptotic normality of the above follows via Slutsky's theorem.

B.4 Proofs

B.4.1 Proofs for propositions and theorems

Proposition 4.1. *For an observational estimator $\hat{\tau}(k)$, assume Assumptions 4.2 and 4.4 hold. Furthermore, let $N = N_k + N_0$ with fixed proportions, where $N_k = \rho N, N_0 = (1-\rho)N$ for $\rho \in (0,1)$. Define the test statistic*

$$\hat{T}_N(k,i) \coloneqq \frac{\hat{\tau}_i(k) - \hat{\tau}_i(0) - \mu_i(k)}{\hat{s}} \tag{4.3}$$

where $\hat{s}^2 \coloneqq \frac{\hat{\sigma}_i^2(k)}{N_k} + \frac{\hat{\sigma}_i^2(0)}{N_0}$ is the estimated variance, and $\mu_i(k) \coloneqq \tau_i(k) - \tau_i$. This test statistic converges in distribution to a normal distribution as $N \to \infty$, $\hat{T}_N(k,i) \xrightarrow{d}$

$\mathcal{N}(0,1)$.

Proof. As $N \to \infty$, we have it that

$$\sqrt{\rho N}(\hat{\tau}_i(k) - \tau_i(k)) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_i^2(k))$$
$$\sqrt{(1-\rho)N}(\hat{\tau}_i(0) - \tau_i) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_i^2(0))$$

where we have written ρN in place of N_k, and similarly for N_0. By Slutsky's theorem, we can multiply by the constants $\rho^{-1/2}$ and $(1-\rho)^{-1/2}$ to get both results in terms of $\sqrt{N}$. We can then use independence of $\hat{\tau}(k), \hat{\tau}(0)$ to write that

$$\sqrt{N} \underbrace{\begin{pmatrix} \hat{\tau}_i(k) - \tau_i(k) \\ \hat{\tau}_i(0) - \tau_i \end{pmatrix}}_{Z-\theta} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_i^2(k)/\rho & 0 \\ 0 & \sigma_i^2(0)/(1-\rho) \end{bmatrix} \right).$$

We now apply the Delta method. Let $Z = (\hat{\tau}_i(k), \hat{\tau}_i(0))$ denote the (column) vector of estimates, and similarly let $\theta = (\tau_i(k), \tau_i)$. Letting $f(X) = X_1 - X_2$, we can argue that

$$\sqrt{N}(Z-\theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma) \implies \sqrt{N}(f(Z) - f(\theta)) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \nabla f(\theta)^\top \Sigma \; \nabla f(\theta)\right),$$

where the resulting variance is given by

$$\nabla f(\theta)^\top \Sigma \; \nabla f(\theta) = \frac{\sigma_i^2(k)}{\rho} + \frac{\sigma_i^2(0)}{1-\rho},$$

and $f(Z) - f(\theta) = \tau_i(k) - \tau_i - \mu_i(k)$.

$$\sqrt{N}(\tau_i(k) - \tau_i - \mu_i(k)) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\sigma_i^2(k)}{\rho} + \frac{\sigma_i^2(0)}{1-\rho}\right),$$

and accordingly that

$$\frac{\tau(k)_i - \tau_i - \mu_i(k)}{\sqrt{\frac{\sigma_i^2(k)}{N_k} + \frac{\sigma_i^2(0)}{N_0}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

where this also holds (by Slutsky's theorem) with $\sigma^2(k)_i$ and $\sigma_i^2(0)$ replaced by their empirical estimates, which converge in probability. □

Theorem 4.1 (Properties of Algorithm 1)**.** *Under Assumptions 4.1 and 4.2, the output of Algorithm 1 has the following asymptotic properties as $N \to \infty$, where N denotes the total sample size, and the samples used for all estimators are of the same order $N_k = \rho_k N_0, \forall k \geq 1$, for some $\rho_k > 0$.*

1. *Under Assumptions 4.3 and 4.4, for each $i \in \mathcal{I}_O$,*

$$\lim_{N\to\infty} \mathbb{P}(\tau_i \in [\hat{L}_i, \hat{U}_i]) \geq 1 - \alpha \tag{4.6}$$

2. *Under Assumption 4.4, for each estimator where $\tau_i(k) \neq \tau_i$ for some $i \in \mathcal{I}_R$,*

$$\lim_{N\to\infty} \mathbb{P}(k \in \hat{\mathcal{C}}) = 0 \tag{4.7}$$

Proof. **(1)** By asymptotic normality and consistency of each dimension of $\tau(k)$, the test statistic $\hat{T}_N(k,i)$ converges in distribution to $\mathcal{N}(0,1)$. As a result, for each $i \in \mathcal{I}_R$, the probability that $\left|\hat{T}_N(k,i)\right| > z_{\alpha/(4|\mathcal{I}_R|)}$ converges to $\alpha/(2\,|\mathcal{I}_R|)$. By an application of the union bound, the probability that this occurs for any $i \in \mathcal{I}_R$ is bounded by $\alpha/2$. Similarly, by the assumed properties of $\tau(k)$, the probability that the confidence interval $[\hat{L}_i(k)(\alpha/2), \hat{U}_i(k)(\alpha/2)]$ fails to capture the true value of τ_i converges to $\alpha/2$. By another application of the union bound, for each $i \in \mathcal{I}_O$, the probability that either $\tau(k)$ is not selected or τ_i is not contained in the interval is upper bounded by α. The result follows.

(2) By asymptotic normality of each $\tau(k)$, the power calculation in Equation (4.4) holds, and as $N \to \infty$, the probability of rejecting the null hypothesis converges to zero as $\sigma^2_{k,0}$ becomes arbitrarily large, which occurs as both $N_k, N_0 \to \infty$. □

Proposition B.3.1. *In the setting of Section B.2, under Assumptions B.2.1 and B.2.2,* *the conditional mean potential outcome in the target distribution is identified as*

$$\mathbb{E}[Y_a \mid S = 0, G = i] = \frac{\mathbb{P}(S = 0)}{\mathbb{P}(S = 0 \mid G = i)}\mathbb{E}[Y^a(\eta_0, \pi_0) \mid G = i], \tag{B.7}$$

where $Y^a(\eta, \pi)$ is defined as in Equation B.8.

$$Y^a(\eta, \pi) \coloneqq \frac{1}{\pi}\left(\mathbf{1}\left\{S = 1, A = a\right\} \cdot \frac{1 - p(X)}{p(X)e_a(X)} \cdot \{Y - g_a(X)\} + (1 - S)g_a(X)\right) \tag{B.8}$$

where $\eta \coloneqq (g_a, e_a, p)$ with true underlying parameters $\eta_0 = (g_{a0}, e_{a0}, p_0)$, $\pi \coloneqq \mathbb{P}(S = 0)$ with true value π_0, $g_a(X) \coloneqq \mathbb{E}[Y \mid A = a, S = 1, X]$, $p(X) \coloneqq \mathbb{P}(S = 1 \mid X)$, and $e_a(X) \coloneqq \mathbb{P}(A = a \mid S = 1, X]$.

Proof. First, we can observe by standard arguments that the conditional expectation of $Y^a(\eta_0, \pi_0)$ given X is given by the following

$$\mathbb{E}[Y^a(\eta_0, \pi_0) \mid X = x] = \mathbb{E}\left[\frac{1 - S}{\mathbb{P}(S = 0)}g_{a0}(X)\middle| X = x\right],$$

because the first term in Equation (B.8) is mean-zero conditioned on $X = x$. This follows by the law of total expectation: for any event where $S = 1, A = a$ does not hold, the first term is zero due to the indicator, and for any other event $S = 1, A = a, X = x$, the first term is mean-zero, since the first term becomes a constant (determined by $S = 1, A = a, X = x$) times a mean-zero random variable $Y - \mathbb{E}[Y \mid A = a, S = 1, X = x]$.

As a result, we can write that

$$\begin{aligned}
&\mathbb{E}[Y^a(\eta_0, \pi_0) \mid G = i] \\
&= \mathbb{E}\left[\frac{1 - S}{\mathbb{P}(S = 0)}\mathbb{E}[Y \mid A = a, S = 1, X]\middle| G = i\right] \\
&= \frac{1}{\mathbb{P}(S = 0)}\int_x \sum_s \mathbf{1}\left\{s = 0\right\}\mathbb{E}[Y \mid A = a, S = 1, x]p(s, x \mid G = i)dx \\
&= \frac{1}{\mathbb{P}(S = 0)}\int_x \sum_s \mathbf{1}\left\{s = 0\right\}\mathbb{E}[Y_a \mid S = 1, x]p(s, x \mid G = i)dx
\end{aligned}$$

(By Assumption B.2.1, $Y_a \perp\!\!\!\perp A \mid X, S = 1$)

$$= \frac{1}{\mathbb{P}(S=0)} \int_x \sum_s \mathbf{1}\{s=0\} \mathbb{E}[Y_a \mid S=0, x] p(s, x \mid G=i) dx$$

(By Assumption B.2.2, $Y_a \perp\!\!\!\perp S \mid X$)

$$= \frac{1}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(S=0, x \mid G=i) dx$$

$$= \frac{1}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(x \mid S=0, G=i) \mathbb{P}(S=0 \mid G=i) dx$$

$$= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(x \mid S=0, G=i) dx$$

$$= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0, G=i] p(x \mid S=0, G=i) dx$$

(Since $X = x \Rightarrow G = i, \forall x : p(x \mid G=i) > 0$)

$$= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \mathbb{E}[Y_a \mid S=0, G=i]$$

and the result follows from dividing both sides by the first term on the right-hand side, which we can observe is equivalent to multiplying both sides by

$$\frac{\mathbb{P}(S=0)}{\mathbb{P}(S=0 \mid G=i)} = \frac{\mathbb{P}(S=0)\mathbb{P}(G=i)}{\mathbb{P}(S=0, G=i)} = \frac{\mathbb{P}(G=i)}{\mathbb{P}(G=i \mid S=0)} \tag{B.12}$$

□

B.4.2 Asymptotic normality of cross-fitted transported GATE estimators

Theorem B.3.1. *Suppose Assumptions B.3.1 to B.3.5 hold. Then, the empirical average, $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g) | G = i]$, where $\tilde{Y}$ is defined in Equation B.9 and $\hat{\eta}$ is estimated with cross-fitting, is asymptotically normal.*

Proof sketch: Our strategy for the proof consists of two stages. First, we show that if the nuisance function is known to be η_0 and plugged into the estimator as $\mathbb{E}[\tilde{Y}(\eta_0, \hat{\pi}_g)|G = i]$, the resulting estimator, which we later refer to as the oracle estimator, is asymptotically normal. Second, we show that even if the true nuisance function is not known, as long as we have an estimator, $\hat{\eta}$, of the nuisance function that follows certain properties, the resulting estimator $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$ converges to the oracle estimator in probability. Then, by Slutsky's Theorem, the resulting estimator is also asymptotically normal.

Before diving into the first stage of the proof, we introduce additional notation to reflect the cross-fitting nature of our GATE estimator. Let the combined sample size of the observational study and RCT be N with sample indices $[N] \coloneqq \{1, 2, ..., N\}$. We denote $(I_m)_{m=1}^{M}$ as a M-fold random partition of $[N]$, so that each fold has size $N_M = N/M$. The plug-in nuisance function estimate for the m^{th} fold, $\hat{\eta}_m$, is then estimated from the rest of the folds $I_m^c \coloneqq [N] \backslash I_m$. For brevity, we denote the size of the rest of the folds as $N_M^c = N - N/M$.

We now restate the definition of the treatment effect signal $\tilde{Y}_j(\eta, \pi_g) = \tilde{Y}_j((g_1, g_0, e_1, p)^\top, \pi_g)$:

$$
\begin{aligned}
&\tilde{Y}_j(\eta, \pi_g) \coloneqq \tilde{Y}_j^1(\eta, \pi_g) - \tilde{Y}_j^0(\eta, \pi_g) \\
&\tilde{Y}_j^a(\eta, \pi_g) \coloneqq \frac{1}{\pi_g(G_i)} \left(\mathbf{1}\{S_j = 1, A_j = a\} \cdot \frac{1 - p(X_j)}{p(X_j) e_a(X_j)} \cdot \{Y_j - g_a(X_j)\} + (1 - S_j) g_a(X_j) \right)
\end{aligned}
$$

In the remainder of the development, we will drop the subscript j, which represents one of the N samples, for conciseness.

Stage 1 — *Proving the asymptotic normality of the oracle estimator*

For brevity, we define the following unweighted signal:

Definition B.4.1 (Unweighted signal functional)**.**

$$
\begin{aligned}
\mathcal{Y}(\eta) &= \pi_g(G) \tilde{Y}(\eta, \pi_g) \\
&= \pi_g(G)(\tilde{Y}^1(\eta, \pi_g) - \tilde{Y}^0(\eta, \pi_g)) \\
&= (1 - S)(g_1(X) - g_0(X)) + S \frac{1 - p(X)}{p(X)} \frac{(A - e_1(X))(Y - g_A(X))}{e_1(X) e_0(X)}
\end{aligned}
$$

From the proof of Proposition B.3.1, we have the following identities for the unweighted signals:

Lemma B.4.1 (Conditional mean of unweighted (oracle) signal)**.** *The conditional mean of the unweighted (oracle) signal is equivalent to the following:*

$$\begin{aligned}
&\mathbb{E}[\mathcal{Y}(\eta_0)|G = i] = \tau_i \pi_g(i) \\
&\mathbb{E}[\mathcal{Y}(\eta_0)|G = i, S = 0] = \tau_i
\end{aligned}$$

.

Proof. First, we have,

$$\begin{aligned}
\mathbb{E}[\mathcal{Y}(\eta_0)|G = i] &= \mathbb{E}[\pi_g(G)\tilde{Y}(\eta_0, \pi_g)|G = i] \\
&= \mathbb{E}[\pi_g(i)\tilde{Y}(\eta_0, \pi_g)|G = i] \\
&= \pi_g(i)\mathbb{E}[\tilde{Y}(\eta_0, \pi_g)|G = i] \\
&= \tau_i \pi_g(i)
\end{aligned}$$

Next, using Definition D.1 of the unweighted signal functional and the fact that we condition on $S = 0$, we have,

$$\mathbb{E}[\mathcal{Y}(\eta_0)|G = i, S = 0] = \mathbb{E}\left[g_{10}(X) - g_{00}(X)|G = i, S = 0\right],$$

which is τ_i as desired. □

In addition, we can rewrite our estimator $\hat{\tau}_i := \mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$ with the unweighted signals:

$$
\begin{aligned}
\hat{\tau}_i = \mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i] &= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i) \tilde{Y}(\hat{\eta}_m, \hat{\pi}_g)}{\sum_j \mathbf{1}(G_j = i)} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i) \frac{1}{\hat{\pi}_g(G_j)} \mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = i)}, \quad \text{from } \textbf{Def. D.1} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i) \frac{1}{\hat{\pi}_g(i)} \mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = i)} \\
&= \frac{1}{\hat{\pi}_g(i)} \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i) \mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = i)} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i) \mathcal{Y}(\hat{\eta}_m)}{\frac{\sum_j \mathbf{1}(G_j = 1, S_j = 0)}{\sum_j \mathbf{1}(G_j = i)} \sum_j \mathbf{1}(G_j = i)} \\
&= \frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i) \mathcal{Y}(\hat{\eta}_m)}{\sum_j \mathbf{1}(G_j = 1, S_j = 0)}
\end{aligned}
$$

Now, using the above expression, we can define the oracle estimator, where we *know* the true value of η, which is η_0:

Definition B.4.2 (Oracle GATE Estimator)**.**

$$
\hat{\tau}_{i0} \coloneqq \frac{\sum_j \mathbf{1}(G_j = i) \mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j = i, S_j = 0)}
$$

To show the asymptotic distribution of the oracle GATE estimator, we restate several assumptions:

Assumption B.3.1 (Observational dataset covers the whole support of covariates)**.**

$$
\inf_{x \in \mathcal{X}} p_0(x) = \varepsilon_p > 0
$$

Assumption B.3.2 (Bounded within-subgroup variance of conditional treatment effects in the RCT)**.**

$$
\sup_{x \in \mathcal{X}} var[g_{10}(x) - g_{00}(x)|G = i, S = 0] = \sigma_{\tau i}^2 < \infty
$$

Assumption B.3.3 (Overlap between treatments in the observational dataset)**.**

$$\inf_{x \in \mathcal{X}} \min(e_{00}(x), e_{10}(x)) = \varepsilon_e > 0$$

Assumption B.3.4 (Finite outcome conditional variance in the observational dataset)**.**

$$\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} \mathbb{E}[(Y - g_{a0}(x))^2 | X = x, S = 1, A = a] = \bar{\sigma}^2 < \infty$$

These assumptions ensure that the oracle signals have finite conditional variance, which we prove in the following lemma.

Lemma B.4.2 (Finite conditional variance of unweighted oracle signal)**.** *Under Assumptions B.3.1 - B.3.4, we have that,*

$$var[\mathcal{Y}(\eta_0)|G = i] \coloneqq \sigma_i^2 < \infty, \forall i \in [d]$$

Proof.

$$
\begin{aligned}
& var[\mathcal{Y}(\eta_0)|G=i] \\
=& \mathbb{E}[\mathcal{Y}^2(\eta_0)|G=i] - [\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]]^2 \\
=& \mathbb{E}\left[\left((1-S)(g_{10}(X)-g_{00}(X)) + \right.\right. \\
& \qquad \left.\left. S\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2 \middle| G=i\right] - \pi_g(i)^2\tau_i^2 \\
=& \left\{\mathbb{E}\left[(1-S)(g_{10}(X)-g_{00}(X))^2 \middle| G=i\right] + \right. \\
& \qquad \left. \mathbb{E}\left[S\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2 \middle| G=i\right]\right\} - \pi_g(i)^2\tau_i^2 \\
=& \left\{\mathbb{E}\left[(g_{10}(X)-g_{00}(X))^2 \middle| G=i, S=0\right]\pi_g(i) + \right. \\
& \qquad \left. \mathbb{E}\left[\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2 \middle| G=i, S=1\right](1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 \\
=& \left\{\left[var\left[g_{10}(X)-g_{00}(X)|G=i, S=0\right] + \tau_i^2\right]\pi_g(i) + \right. \\
& \qquad \left. \mathbb{E}\left[\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2 \middle| G=i, S=1\right](1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 \\
<& \left\{\left[\sigma_{\tau i}^2+\tau_i^2\right]\pi_g(i) + \frac{\mathbb{E}[(Y-g_{A0}(X))^2|G=i,S=1]}{\varepsilon_\pi^2\varepsilon_e^2(1-\varepsilon_e)^2}(1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 \\
\leq& \left\{\left[\sigma_{\tau i}^2+\tau_i^2\right]\pi_g(i) + \frac{\bar{\sigma}^2}{\varepsilon_\pi^2\varepsilon_e^2(1-\varepsilon_e)^2}(1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 < \infty
\end{aligned}
$$

Where the first line follows from Lemma B.4.1, the penultimate line follows from Assumptions B.3.1-B.3.3, and the final line follows from Assumption B.3.4. □

Now, using the above lemmas, we are ready to prove the main result of stage 1 of the proof, stated below.

Proposition B.4.1 (Asymptotic normality of oracle GATE estimator)**.** *Under Assumptions B.3.1 - B.3.4,*

$$\sqrt{N}(\hat{\tau}_{i0}-\tau_i) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2 - \pi_g(i)(1-\pi_g(i))\tau_i^2}{\pi_g(i)^2\mathbb{P}(G=i)}\right)$$

Proof. We have that,

$$\begin{aligned}
&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) \\
=&\sqrt{N}\left(\frac{\sum_j \mathbf{1}(G_j = i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j = i, S_j = 0)} - \tau_i\right) \\
=&\sqrt{N}\left(\frac{\sum_j \mathbf{1}(G_j = i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j = i, S_j = 0)} - \frac{\sum_j \mathbf{1}(G_j = i, S_j = 0)\tau_i}{\sum_j \mathbf{1}(G_j = i, S_j = 0)}\right) \\
=&\sqrt{N}\frac{\sum_j \mathbf{1}(G_j = i)(\mathcal{Y}_j(\eta_0) - \tau_i \mathbf{1}(S_j = 0))}{\sum_j \mathbf{1}(G_j = i, S_j = 0)} \\
=&\frac{\sqrt{N}\frac{1}{N}\sum_j \mathbf{1}(G_j = i)(\mathcal{Y}_j(\eta_0) - \tau_i \mathbf{1}(S_j = 0))}{\frac{1}{N}\sum_j \mathbf{1}(G_j = i, S_j = 0)} \quad \begin{aligned}&\xrightarrow{d} \mathcal{N}(0, (\sigma_i^2 - \tau_i^2 \pi_g(i)(1-\pi_g(i)))\mathbb{P}(G = i)) \\ &\xrightarrow{p} \mathbb{P}(G = i, S = 0) = \mathbb{P}(G = i)\pi_g(i)\end{aligned} \\
\xrightarrow{d}&\mathcal{N}\left(0, \frac{\sigma_i^2 - \tau_i^2\pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2\mathbb{P}(G = i)}\right)
\end{aligned}$$

Where in the last line, we use Slutsky's lemma, and in the penultimate line, we use the following fact, proven below, that

$$\sqrt{N}\left[\frac{1}{N}\sum_j \mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))\right] \xrightarrow{d} \mathcal{N}(0, (\sigma_i^2 - \tau_i^2 \pi_g(i)(1-\pi_g(i)))\mathbb{P}(G = i))$$

To show this, we observe that

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))] \\
=&\mathbb{E}[\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0)|G = i]\mathbb{P}(G = i) \\
=&(\mathbb{E}[\mathcal{Y}(\eta_0)|G = i] - \tau_i \mathbb{E}[\mathbf{1}(S = 0)|G = i])\mathbb{P}(G = i) \\
=&(\mathbb{E}[\mathcal{Y}(\eta_0)|G = i] - \tau_i \pi_g(i))\mathbb{P}(G = i) = 0 && \text{Lem. } B.4.1
\end{aligned}$$

$$\begin{aligned}
&var[\mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))] \\
=&\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))]^2 \\
&\qquad - (\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))])^2 \\
=&\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))]^2 \\
=&\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S = 0))^2] && \mathbf{1}(G = i) \in \{0, 1\}
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}[(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S=0))^2 | G=i]\mathbb{P}(G=i) \\
=&\left\{\mathbb{E}[\mathcal{Y}^2(\eta_0)|G=i] + \tau_i^2 \mathbb{E}[\mathbf{1}(S=0)|G=i]\right. \\
&\qquad \left. -2\tau_i \mathbb{E}[\mathcal{Y}(\eta_0)\mathbf{1}(S=0)|G=i]\right\}\mathbb{P}(G=i) \qquad\qquad \mathbf{1}(S=0) \in \{0,1\} \\
=&\left\{var[\mathcal{Y}(\eta_0)|G=i] + (\mathbb{E}[\mathcal{Y}(\eta_0)|G=i])^2 + \tau_i^2 \pi_g(i)\right. \\
&\qquad \left. -2\tau_i \pi_g(i)\mathbb{E}[\mathcal{Y}(\eta_0)|G=i, S=0]\right\}\mathbb{P}(G=i) \\
=&\left\{\sigma_i^2 + \tau_i^2 \pi_g(i)^2 + \tau_i^2 \pi_g(i) - 2\tau_i^2 \pi_g(i)\right\}\mathbb{P}(G=i) \qquad \text{Asmp. } B.4.2\text{, Lem. } B.4.1 \\
=&(\sigma_i^2 - \tau_i^2 \pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i) < \infty
\end{aligned}
$$

Therefore, from central limit theorem,

$$
\sqrt{N}\left[\frac{1}{N}\sum_j \mathbf{1}(G=i)(\mathcal{Y}(\eta_0) - \tau_i \mathbf{1}(S=0))\right] \xrightarrow{d} \mathcal{N}(0, (\sigma_i^2 - \tau_i^2 \pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i))
$$

□

Stage 2 — *Proving the asymptotic normality of the cross-fitted estimator,* $\tilde{Y}(\hat{\eta}, \hat{\pi}_g)$

With asymptotic normality of the oracle estimator shown above in Stage 1, we can show the asymptotic normality of the cross-fitted estimator (i.e. our estimator) by decomposing its error into the error of the oracle estimator and the difference between our estimator and the oracle estimator:

$$
\begin{aligned}
\sqrt{N}(\hat{\tau}_i - \tau_i) &= \sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}(\hat{\tau}_i - \hat{\tau}_{i0}) \\
&= \sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}\frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))}{\sum_j \mathbf{1}(G_j = i, S_j = 0)} \\
&= \sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \frac{\sum_m \frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))}{\frac{1}{N}\sum_j \mathbf{1}(G_j = i, S_j = 0)}
\end{aligned}
$$

The asymptotic distribution of the cross-fitted estimator therefore hinges on the asymptotic property of $\frac{1}{\sqrt{N}}\sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))$, which in turn depends on the convergence property of the nuisance function estimate $\hat{\eta}_m$ and its influence on the signal $\mathcal{Y}$. We therefore restate the last required assumption governing the convergence properties of $\hat{\eta}_m$:

Assumption B.3.5 (Properties of the nuisance function estimators)**.** Let $\hat{\eta}_{(n)}$ be a sequence of estimators for η indexed by the size of the cross-fitting training fold n. We assume that there exists

- $\epsilon_n = o_P(1)$, a sequence of positive numbers
- $\mathcal{T}_n$, a sequence of nuisance function vector sets in the neighborhood of $\eta_0 = (g_{10}, g_{00}, e_{10}, p_0)$ satisfying $\mathbb{P}(\hat{\eta}_{(n)} \in \mathcal{T}_n) \geq 1 - \epsilon_n$
- $\mathbf{g}_n, \mathbf{e}_n, \mathbf{p}_n$, sequences of worst root mean square errors for the nuisance functions g_1, g_0, e_1, p, defined as follows:

$$\begin{aligned}
\mathbf{g}_n &:= \max_{a\in\{0,1\}} \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[g_a(X) - g_{a0}(X)]^2} \\
\mathbf{e}_n &:= \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[e_1(X) - e_{10}(X)]^2} \\
\mathbf{p}_n &:= \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[p(X) - p_0(X)]^2}
\end{aligned}$$

so that the following assumptions hold:

Assumption A: (Rate of nuisance error)

$$\mathbf{g}_n \vee \mathbf{e}_n \vee \mathbf{p}_n = o_P(1)$$

Assumption B: (Rate of nuisance error product)

$$\sqrt{n}\mathbf{g}_n(\mathbf{e}_n \vee \mathbf{p}_n) = o_P(1)$$

Assumption C: (Bounded nuisance estimates)

$$\sup_{\eta\in\cup_{n=1}^{\infty}\mathcal{T}_n} \left(\max_{a\in\{0,1\}} \sup_{x\in\mathcal{X}} |g_a(x)| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{p(x)}\right| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{e_1(x)}\right| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{e_0(x)}\right| \right) = \bar{\mathcal{C}} < \infty$$

Based on the assumptions above, we have the following bounds on the convergence rate of the signals when the nuisance function estimates are in the high-probability neighborhood, $\mathcal{T}_n$:

Lemma B.4.3 (Bounds on bias of signal)**.** *Under Assumptions B.3.5.B and B.3.5.C*

$$\sqrt{n} \sup_{\eta \in \mathcal{T}_n} |\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| = o_P(1)$$

Lemma B.4.4 (Bounds on MSE of signal)**.** *Under Assumptions B.3.1, B.3.3, B.3.4, B.3.5.A and B.3.5.C*

$$\sup_{\eta \in \mathcal{T}_n} \mathbb{E} \left| \mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0)) \right|^2 = o_P(1)$$

which in turn implies that $\frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))$ converges to zero in probability:

Lemma B.4.5 (Numerator of difference is $o_P(1)$)**.** *Under Assumptions B.3.1, B.3.3, B.3.4 and B.3.5,*

$$\frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0)) = o_P(1), \quad \forall m \in \{1, 2, ..., M\}$$

The proofs for Lemmas B.4.3 to B.4.5 are more labor-intensive and we defer these proofs to later subsections. Based on these lemmas, we arrive at the main result of Stage 2.

Theorem B.4.1 (Asymptotic normality of the cross-fitted transported GATE estimator)**.** *Under Assumptions B.3.1 - B.3.5,*

$$\sqrt{N}(\hat{\tau}_i - \tau_i) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2 - \tau_i^2 \pi_g(i)(1 - \pi_g(i))}{\pi_g(i)^2 \mathbb{P}(G = i)}\right)$$

Proof.

$$
\begin{aligned}
&\sqrt{N}(\hat{\tau}_i - \tau_i)\\
=&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}(\hat{\tau}_i - \hat{\tau}_{i0})\\
=&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}\frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))}{\sum_j \mathbf{1}(G_j = i, S_j = 0)}\\
=&\sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \frac{\sum_m \frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0)) \xrightarrow{p} 0}{\frac{1}{N}\sum_j \mathbf{1}(G_j = i, S_j = 0) \xrightarrow{p} \mathbb{P}(G = i, S = 0)} && \begin{matrix}\text{Lem. } B.4.5\\ \text{WLLN}\end{matrix}\\
\xrightarrow{d}&\mathcal{N}\left(0, \frac{\sigma_i^2 - \tau_i^2 \pi_g(i)(1 - \pi_g(i))}{\pi_g(i)^2 \mathbb{P}(G = i)}\right) && \text{Prop. } B.4.1
\end{aligned}
$$

□

Note that Theorem B.4.1 is simply Theorem B.3.1, which is the primary result of this section, with the variance explicitly stated. Thus, Theorem B.3.1 is proven.

B.4.3 Proof for Lemmas B.4.3 and B.4.4

First, we prove Lemmas B.4.3 and B.4.4, which will be necessary for Lemma B.4.5. Recall that Lemma B.4.5 was essential for the proof of the asymptotic normality result in Theorem B.4.1.

Lemma B.4.3 (Bounds on bias of signal)**.** *Under Assumptions B.3.5.B and B.3.5.C*

$$\sqrt{n} \sup_{\eta \in \mathcal{T}_n} |\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| = o_P(1)$$

Lemma B.4.4 (Bounds on MSE of signal)**.** *Under Assumptions B.3.1, B.3.3, B.3.4, B.3.5.A and B.3.5.C*

$$\sup_{\eta \in \mathcal{T}_n} \mathbb{E}\left|\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))\right|^2 = o_P(1)$$

Proof. We first define partial unweighted signal functionals for the two counterfactual outcomes

Definition B.4.3 (Partial unweighted signal functionals)**.**

$$\begin{aligned}
\mathcal{Y}^1(\eta) &\coloneqq \left[(1-S)g_1(X) + S\frac{1-p(X)}{p(X)}\frac{A(Y-g_1(X))}{e_1(X)}\right] \\
\mathcal{Y}^0(\eta) &\coloneqq \left[(1-S)g_0(X) + S\frac{1-p(X)}{p(X)}\frac{(1-A)(Y-g_0(X))}{e_0(X)}\right] \\
\Rightarrow \mathcal{Y}(\eta) &= \mathcal{Y}^1(\eta) - \mathcal{Y}^0(\eta)
\end{aligned}$$

At a high level, we will prove the above lemmas by decomposing the errors of signal functionals into simpler terms that can be bounded by standard concentration inequalities. This idea will be repeated for both the bias and MSE of the signals. To simplify the analysis, we can split up the unweighted signal into "partial signals" (for the treatment and control groups). Therefore, we set out to show the following lemmas:

Lemma B.4.6 (Bounds on bias of partial signal)**.** *Under Assumptions B.3.5.B and B.3.5.C*

$$\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\right| = o_P(1)$$
$$\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right| = o_P(1)$$

Lemma B.4.7 (Bounds on MSE of partial signal)**.** *Under Assumptions B.3.1, B.3.3, B.3.4, B.3.5.A and B.3.5.C*

$$\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2 = o_P(1)$$
$$\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2 = o_P(1)$$

In the following subsections, we prove the $\mathcal{Y}_1$ part of Lemmas B.4.6 and B.4.7. The $\mathcal{Y}_0$ part will follow by symmetry. First, we further define $\eta(X) = \eta_0(X) + \delta_\eta(X)$, in detail:

$$\begin{aligned} g_1(X) &= g_{10}(X) + \delta_{g_1}(X) \\ p(X) &= p_0(X) + \delta_p(X) \\ e(Z) &= e_0(X) + \delta_e(X) \end{aligned}$$

so that (omitting the parameter X for brevity),

$$\begin{aligned} &\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0) \\ =& \left[(1-S)(g_{10}+\delta_{g_1}) + \frac{1-p_0-\delta_p}{p_0+\delta_p}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0+\delta_e}\right] - \left[(1-S)g_{10} + \frac{1-p_0}{p_0}\frac{SA(Y-g_{10})}{e_0}\right] \\ =&(1-S)\delta_{g_1} + \frac{1-p_0-\delta_p}{p_0+\delta_p}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0+\delta_e} - \frac{1-p_0}{p_0}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0} - \frac{1-p_0}{p_0}\frac{SA}{e_0}\delta_{g_1} \\ =& \left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1} + \left(\frac{1-p_0-\delta_p}{(p_0+\delta_p)(e_0+\delta_e)} - \frac{1-p_0}{p_0e_0}\right)SA(Y-g_{10}-\delta_{g_1}) \\ =& \left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1} - \frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10}-\delta_{g_1}) \\ =& \underbrace{\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}}_{S_1} - \underbrace{\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})}_{S_2} \\ &\qquad + \underbrace{\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA\delta_{g_1}}_{S_3} \\ :=& S_1 - S_2 + S_3 \end{aligned}$$

Proof for Lemma B.4.6

For Lemma B.4.6 we want to bound

$$\begin{aligned} &\left|\mathbb{E}\left[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))\right]\right| \\ &= \left|\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0)\right)\middle|X\right]\right]\right| \\ &= \left|\mathbb{E}\left[\mathbf{1}(G=i)\mathbb{E}\left[S_1 - S_2 + S_3|X\right]\right]\right| && G \text{ is a function of } X \\ &= \left|\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathbb{E}\left[S_1|X\right] - \mathbb{E}\left[S_2|X\right] + \mathbb{E}\left[S_3|X\right]\right)\right]\right| \end{aligned}$$

For the term $\mathbb{E}[S_1|X]$,

$$\begin{aligned}
&\mathbb{E}[S_1|X] \\
=&\mathbb{E}\left[\left((1-S)-\frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}\middle|X\right] \\
=&\left((1-\mathbb{E}[S|X])-\frac{1-p_0}{p_0}\frac{\mathbb{E}[SA|X]}{e_0}\right)\delta_{g_1} && \eta_0,\delta_\eta \text{ are functions of } X \\
=&\left((1-p_0)-\frac{1-p_0}{p_0}\frac{p_0e_0}{e_0}\right)\delta_{g_1}=0 && p_0(X)=\mathbb{P}[S=1|X] \\
& && e_0(X)=\mathbb{P}[A=1|S=1,X]
\end{aligned}$$

For the term $\mathbb{E}[S_2|X]$

$$\begin{aligned}
&\mathbb{E}[S_2|X] \\
=&\mathbb{E}\left[\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})\middle|X\right] \\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\mathbb{E}[SA(Y-g_{10})\mid X] && \eta_0,\delta_\eta \text{ are functions of } X \\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\cdot 0=0 && g_{10}(X)=\mathbb{E}[Y|S=1,A=1,X]
\end{aligned}$$

For the term $\mathbb{E}[S_3|X]$,

$$\begin{aligned}
&\mathbb{E}[S_3|X] \\
=&\mathbb{E}\left[\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA\delta_{g_1}\middle|X\right] \\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\mathbb{E}\left[SA|X\right]\delta_{g_1} && \eta_0,\delta_\eta \text{ are functions of } X \\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}p_0e_0\delta_{g_1} && p_0(X)=\mathbb{P}[S=1|X] \\
& && e_0(X)=\mathbb{P}[A=1|S=1,X] \\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\left|\mathbb{E}\left[\mathbf{1}(G=i)[\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0)]\right]\right|^2 \\
=&\left|\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathbb{E}[S_1|Z]-\mathbb{E}[S_2|Z]+\mathbb{E}[S_3|Z]\right)\right]\right|^2 \\
=&\left|\mathbb{E}\left[\mathbf{1}(G=i)\mathbb{E}[S_3|Z]\right]\right|^2 \\
\leq&\left(\mathbb{E}\left|\mathbf{1}(G=i)\mathbb{E}[S_3|Z]\right|\right)^2 && |EA|\leq E|A| \\
\leq&\left(\mathbb{E}\left|\mathbb{E}[S_3|Z]\right|\right)^2 \\
=&\left(\mathbb{E}\left|\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}\right|\right)^2 \\
\leq&\left(\mathbb{E}\left|\frac{e_0\delta_p}{pe}\delta_{g_1}\right|+\mathbb{E}\left|\frac{(1-p_0)p_0\delta_e}{pe}\delta_{g_1}\right|+\mathbb{E}\left|\frac{(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}\right|\right)^2 && \text{Triangular ineq.} \\
\leq&\bar{\mathcal{C}}^4\left(\mathbb{E}\left|\delta_p\delta_{g_1}\right|+\mathbb{E}\left|\delta_e\delta_{g_1}\right|+\mathbb{E}\left|\delta_p\delta_e\delta_{g_1}\right|\right)^2 && \text{Assmp. } B.3.5.C \\
=&\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}|+\mathbb{E}|\delta_e\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_e||\delta_p\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_p||\delta_e\delta_{g_1}|\right)^2 \\
\leq&\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}|+\mathbb{E}|\delta_e\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_p\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_e\delta_{g_1}|\right)^2 && |\delta_p|,|\delta_e|\leq 1 \\
=&\frac{9}{4}\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}|+\mathbb{E}|\delta_e\delta_{g_1}|\right)^2 \\
\leq&\frac{9}{4}\bar{\mathcal{C}}^4\left(\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_e^2}\sqrt{\mathbb{E}\delta_{g_1}^2}\right)^2 && \text{Hölder's ineq.}
\end{aligned}
$$

So we have,

$$
\begin{aligned}
&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}\mathbf{1}(G=i)[Y^1(\eta)-Y^1(\eta_0)]\right| \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\frac{3}{2}\bar{\mathcal{C}}^2\left(\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_e^2}\sqrt{\mathbb{E}\delta_{g_1}^2}\right) \\
\leq&\frac{3}{2}\bar{\mathcal{C}}^2\sqrt{n}\mathbf{g}_N\left(\mathbf{p}_N+\mathbf{e}_N\right) && \text{Assump. } B.3.5 \\
=&o_P(1) && \text{Assump. } B.3.5.B
\end{aligned}
$$

Proof for Lemma B.4.7

Here we first place bounds on $\mathbb{E}S_1^2, \mathbb{E}S_2^2$ and $\mathbb{E}S_3^2$ for future use. For the term $\mathbb{E}S_1^2$, we have,

$$
\begin{aligned}
&\mathbb{E}S_1^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(\left((1-S)-\frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}\right)^2\middle|X\right]\right] \\
=&\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\left(\frac{1-S}{1-p_0}-\frac{SA}{p_0e_0}\right)^2\middle|X\right]\delta_{g_1}^2\right] && \eta_0, \delta_\eta \text{ are functions of } X \\
=&\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\frac{(1-S)^2}{(1-p_0)^2}+\frac{(SA)^2}{(p_0e_0)^2}\middle|X\right]\delta_{g_1}^2\right] && S \in \{0,1\} \\
=&\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\frac{1-S}{(1-p_0)^2}+\frac{SA}{(p_0e_0)^2}\middle|X\right]\delta_{g_1}^2\right] && 1-S, SA \in \{0,1\} \\
=&\mathbb{E}\left[(1-p_0)^2\left(\frac{1}{1-p_0}+\frac{1}{p_0e_0}\right)\delta_{g_1}^2\right] && p_0(X) = \mathbb{P}[S=1|X] \\
& && e_0(X) = \mathbb{P}[A=1|S=1,X] \\
\le&\frac{2}{\varepsilon_p\varepsilon_e}\mathbb{E}\delta_{g_1}^2 && \text{Assmp. } B.3.1, B.3.3
\end{aligned}
$$

We can similarly bound $\mathbb{E}S_2^2$,

$$
\begin{aligned}
&\mathbb{E}S_2^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})\right)^2\middle|X\right]\right] \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\mathbb{E}\left[(SA(Y-g_{10}))^2\middle|X\right]\right] && \eta_0, \delta_\eta \text{ are functions of } X \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\right. && S, A \in \{0,1\} \\
&\qquad\left.\cdot\,\mathbb{E}\left[(Y-g_{10})^2\middle|X,S=1,A=1\right]\mathbb{P}(S=1,A=1|X)\right] \\
\le&E\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\bar{\sigma}^2\mathbb{P}(S=1,A=1|X)\right] && \text{Assmp. } B.3.4 \\
=&E\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pep_0e_0}\right)^2\bar{\sigma}^2p_0e_0\right] && p_0(X) = \mathbb{P}[S=1|X] \\
& && e_0(X) = \mathbb{P}[A=1|S=1,X]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left| e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e \right|^2 && \text{Assmp. } B.3.1, B.3.3, B.3.5.C \\
&\leq \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left[e_0 |\delta_p| + (1-p_0) p_0 |\delta_e| + (1-p_0) |\delta_p \delta_e| \right]^2 && \text{Triangular ineq.} \\
&\leq \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left[|\delta_p| + |\delta_e| + |\delta_p \delta_e| \right]^2 \\
&= \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left[|\delta_p| + |\delta_e| + \frac{1}{2} |\delta_p| |\delta_e| + \frac{1}{2} |\delta_p| |\delta_e| \right]^2 \\
&\leq \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left[|\delta_p| + |\delta_e| + \frac{1}{2} |\delta_p| + \frac{1}{2} |\delta_e| \right]^2 && |\delta_p|, |\delta_e| \leq 1 \\
&= \frac{9}{4} \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left[|\delta_p| + |\delta_e| \right]^2 \\
&= \frac{9}{4} \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \left[\mathbb{E} \delta_p^2 + \mathbb{E} \delta_e^2 + 2 \mathbb{E} |\delta_p| |\delta_e| \right] \\
&\leq \frac{9}{4} \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \left[\mathbb{E} \delta_p^2 + \mathbb{E} \delta_e^2 + 2 \sqrt{\mathbb{E} \delta_p^2} \sqrt{\mathbb{E} \delta_e^2} \right] && \text{Cauchy-Schwartz} \\
&= \frac{9}{4} \frac{\bar{\sigma}^2 \bar{C}^4}{\varepsilon_p \varepsilon_e} \left(\sqrt{\mathbb{E} \delta_p^2} + \sqrt{\mathbb{E} \delta_e^2} \right)^2
\end{aligned}
$$

Finally, we bound $\mathbb{E} S_3^2$,

$$
\begin{aligned}
&\mathbb{E} S_3^2 \\
&= \mathbb{E} \left[\mathbb{E} \left[\left(\frac{e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e}{(p_0 + \delta_p)(e_0 + \delta_e) p_0 e_0} S A \delta_{g_1} \right)^2 \middle| X \right] \right] \\
&= \mathbb{E} \left[\left(\frac{e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e}{(p_0 + \delta_p)(e_0 + \delta_e) p_0 e_0} \delta_{g_1} \right)^2 \mathbb{E} \left[(SA)^2 | X \right] \right] && \eta_0, \delta_\eta \text{ are functions of } X \\
&= \mathbb{E} \left[\left(\frac{e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e}{(p_0 + \delta_p)(e_0 + \delta_e) p_0 e_0} \delta_{g_1} \right)^2 \mathbb{E} \left[SA | X \right] \right] && SA \in \{0, 1\} \\
&= \mathbb{E} \left[\left(\frac{e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e}{p e p_0 e_0} \delta_{g_1} \right)^2 p_0 e_0 \right] && \begin{aligned} p_0(X) &= \mathbb{P}[S=1|X] \\ e_0(X) &= \mathbb{P}[A=1|S=1,X] \end{aligned} \\
&\leq \frac{\bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left| \left(e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e \right) \delta_{g_1} \right|^2 && \text{Assmp. } B.3.1, B.3.3, B.3.5.C \\
&= \frac{\bar{C}^4}{\varepsilon_p \varepsilon_e} \mathbb{E} \left[|e_0 \delta_p + (1-p_0) p_0 \delta_e + (1-p_0) \delta_p \delta_e|^2 |\delta_{g_1}|^2 \right]
\end{aligned}
$$

$$\le \frac{\bar{C}^4}{\varepsilon_p \varepsilon_e}\mathbb{E}\left[\left(|e_0\delta_p| + |(1-p_0)p_0\delta_e| + |(1-p_0)\delta_p\delta_e|\right)^2 |\delta_{g_1}|^2\right] \qquad \text{Triangular ineq.}$$

$$\le \frac{9\bar{C}^4}{\varepsilon_p \varepsilon_e}\mathbb{E}\delta_{g_1}^2 \qquad 0 \le p_0, e_0, |\delta_p|, |\delta_e| \le 1$$

From the above, we have

$$\begin{aligned}
&\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))\right|^2 \\
=&\mathbb{E}|\mathbf{1}(G=i)||S_1 - S_2 + S_3|^2 \\
\le&\mathbb{E}|S_1 - S_2 + S_3|^2 \\
\le&\mathbb{E}(|S_1| + |S_2| + |S_3|)^2 && \text{Triangular ineq.} \\
=& \left[\mathbb{E}S_1^2 + \mathbb{E}S_2^2 + \mathbb{E}S_3^2 + 2\mathbb{E}|S_1S_2| + 2\mathbb{E}|S_1S_3| + 2\mathbb{E}|S_2S_3|\right] \\
\le& \Big[\mathbb{E}S_1^2 + \mathbb{E}S_2^2 + \mathbb{E}S_3^2 + \\
&\qquad 2\sqrt{\mathbb{E}|S_1|^2}\sqrt{\mathbb{E}|S_2|^2} + 2\sqrt{\mathbb{E}|S_1|^2}\sqrt{\mathbb{E}|S_3|^2} + 2\sqrt{\mathbb{E}|S_2|^2}\sqrt{\mathbb{E}|S_3|^2}\Big] && \text{Cauchy-Schwartz} \\
=& \left[\sqrt{\mathbb{E}S_1^2} + \sqrt{\mathbb{E}S_1^2} + \sqrt{\mathbb{E}S_1^2}\right]^2 \\
\le& \left[\sqrt{\frac{2}{\varepsilon_p\varepsilon_e}}\sqrt{\mathbb{E}\delta_{g_1}^2} + \sqrt{\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}}\left(\sqrt{\mathbb{E}\delta_p^2} + \sqrt{\mathbb{E}\delta_e^2}\right) + \sqrt{\frac{9\bar{C}^4}{\varepsilon_p\varepsilon_e}}\sqrt{\mathbb{E}\delta_{g_1}^2}\right]^2 \\
\le& C\left[\sqrt{\mathbb{E}\delta_{g_1}^2} + \sqrt{\mathbb{E}\delta_p^2} + \sqrt{\mathbb{E}\delta_e^2}\right]^2 && C := \frac{\left(3\bar{C}^2 + \sqrt{2}\right)^2}{\varepsilon_p\varepsilon_e} \vee \frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}
\end{aligned}$$

So we have,

$$\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))\right|^2 \\
\le& C\sup_{\eta\in\mathcal{T}_n}\left[\sqrt{\mathbb{E}\delta_{g_1}^2} + \sqrt{\mathbb{E}\delta_p^2} + \sqrt{\mathbb{E}\delta_e^2}\right]^2 \\
\le& C\left[\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_{g_1}^2} + \sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_p^2} + \sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_e^2}\right]^2 \\
\le& C\left[\mathbf{g}_n + \mathbf{p}_n + \mathbf{e}_n\right]^2 && \text{Assmp. } B.3.5 \\
\le& 9C\left[\mathbf{g}_n \vee \mathbf{p}_n \vee \mathbf{e}_n\right]^2 = o_P(1) && \text{Assmp. } B.3.5.A
\end{aligned}$$

Assembling the proofs for Lemmas B.4.3 and B.4.4

For Lemma B.4.3:

$$\begin{aligned}
&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))]| \\
=&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\big|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))] \\
&\qquad -\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\big| \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\big\{\big|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\big| \\
&\qquad +\big|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\big|\big\} && \text{Triangular ineq.} \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\big|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\big| \\
&\qquad +\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\big|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\big| \\
=&o_P(1) && \text{Lem. } B.4.6
\end{aligned}$$

For Lemma B.4.4:

$$\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))\right|^2 \\
=&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))-\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2 \\
\leq&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left\{\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|+\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|\right\}^2 && \text{Triangular ineq.} \\
=&\sup_{\eta\in\mathcal{T}_n}\Big\{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+ \\
&\qquad 2\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|\Big\} \\
\leq&\sup_{\eta\in\mathcal{T}_n}\Big\{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+ && \text{Cauchy-Schwartz} \\
&\qquad 2\sqrt{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2}\sqrt{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2}\Big\} \\
\leq&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+ \\
&\qquad 2\sqrt{\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2}\sqrt{\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2}
\end{aligned}$$

$$=o_P(1) \qquad \text{Lem. } B.4.7$$

□

B.4.4 Proof for Lemma B.4.5

Now that we have shown Lemmas B.4.3 and B.4.4, it remains to show Lemma B.4.5.

Lemma B.4.5 (Numerator of difference is $o_P(1)$)**.** *Under Assumptions B.3.1, B.3.3, B.3.4 and B.3.5,*

$$\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))=o_P(1), \quad \forall m\in\{1,2,...,M\}$$

Proof. First we observe

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\\
=&\frac{1}{\sqrt{N}}\left[\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right]+\\
&\qquad\qquad\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\\
=&\underbrace{\frac{N_M}{\sqrt{N}}\left[\left(\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\right)-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right]}_{R_1(m)}+\\
&\qquad\qquad\underbrace{\frac{N_M}{\sqrt{N}}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]}_{R_2(m)}\\
:=&R_1(m)+R_2(m)
\end{aligned}$$

We define the event $\mathcal{E}_N$ as $\cap_k\left(\hat{\eta}_m\in\mathcal{T}_{N_M^c}\right)$, i.e. all M nuisance function estimates falling into the high-probability neighborhood where Lemmas B.4.3 and B.4.4 apply.

From union bound,

$$1 - \mathbb{P}(\mathcal{E}_N) \leq \sum_k \mathbb{P}(\hat{\eta}_m \notin \mathcal{T}_{N_M^c}) \ \leq K\epsilon_{N_M^c} = o_P(1) \qquad \because \epsilon_n = o_P(1)$$

Conditional on $\mathcal{E}_N$ and the data complementary to fold m, which we denote as D_m, we have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}(|R_1(m)| \geq \epsilon|\mathcal{E}_N, D_m) \\
=&\mathbb{P}\left(\left|\left(\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))\right) - \right.\right. \\
&\qquad\qquad \left.\left.\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0))]\right| \geq \frac{\sqrt{N}}{N_M}\epsilon \right| \mathcal{E}_N, D_m\right) \\
\leq&\frac{N_M^2}{N\epsilon^2}var\left[\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))\middle|\mathcal{E}_N, D_m\right] && \text{Chebyshev ineq.} \\
=&\frac{N_M}{N\epsilon^2}var\left[\mathbf{1}(G = i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0))|\mathcal{E}_N, D_m\right] \\
\leq&\frac{1}{M\epsilon^2}\mathbb{E}\left[(\mathbf{1}(G = i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0)))^2\middle|\mathcal{E}_N, D_m\right] \\
\leq&\frac{1}{M\epsilon^2}\sup_{\eta\in\mathcal{T}_{N_M}}\mathbb{E}\left[(\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0)))^2\middle|D_m\right] && \hat{\eta}_m \in \mathcal{T}_{N_M^c} \text{ under } \mathcal{E}_N \\
\leq&\frac{1}{M\epsilon^2}\sup_{\eta\in\mathcal{T}_{N_M}}\mathbb{E}\left(\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))\right)^2 && \text{Fold } m \text{ is independent of } D_m \\
=&\frac{1}{M\epsilon^2}o_P(1) = o_P(1) && \text{Lem. } B.4.4
\end{aligned}$$

Also, conditional on $\mathcal{E}_N$ and D_m

$$\begin{aligned}
|R_2(m)| =&\left|\frac{N_M}{\sqrt{N}\sqrt{N_M^c}}\sqrt{N_M^c}\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0))]\middle|\mathcal{E}_N, D_m\right| \\
=&\frac{1}{\sqrt{M-1}}\sqrt{N_M^c}\,|\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0))]|\mathcal{E}_N, D_m| \\
\leq&\frac{1}{\sqrt{M-1}}\sqrt{N_M^c}\sup_{\eta\in\mathcal{T}_{N_M^c}}|\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))|D_m]| && \hat{\eta}_m \in \mathcal{T}_{N_M^c} \text{ under } \mathcal{E}_N \\
=&\frac{1}{\sqrt{M-1}}\sqrt{N_M^c}\sup_{\eta\in\mathcal{T}_{N_M^c}}|\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| && \text{Fold } m \text{ is independent of } D_m
\end{aligned}$$

$$= \frac{1}{\sqrt{M-1}} o_P(1) = o_P(1) \qquad \text{Lem. } B.4.3$$

Which implies that, conditional on $\mathcal{E}_N$ and D_m, $R_1(m) + R_2(m) = o_P(1)$. So for any $\epsilon > 0$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))\right| > \epsilon\right) \\
=&\mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon) \\
=&\mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon|\mathcal{E}_N)\mathbb{P}(\mathcal{E}_N) + \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon|\bar{\mathcal{E}}_N)(1 - \mathbb{P}(\mathcal{E}_N)) \\
\leq&\mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon|\mathcal{E}_N) + (1 - \mathbb{P}(\mathcal{E}_N)) \\
=&\int \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon|\mathcal{E}_N, D_m) d\mathbb{P}(D_m|\mathcal{E}_N) + (1 - \mathbb{P}(\mathcal{E}_N)) \\
=&o_P(1) + o_P(1) = o_P(1)
\end{aligned}$$

and the lemma is proven. □

B.5 Details on WHI Experiments

We assess our algorithm on clinical trial data and observational data available from the Women's Health Initiative (WHI). The RCTs were run by the WHI via 40 US clinical centers from 1993-2005 (1993-1998: enrollment + randomization; 2005: end of follow-up) on postmenopausal women aged 50-79 years, and the observational dataset was designed and run in parallel on a similar population. Note that this data is publicly available to researchers and requires only an application on BIOLINCC (https://biolincc.nhlbi.nih.gov/studies/whi_ctos/).

B.5.1 Data

WHI RCT – There are three clinical trials associated with the WHI. The RCT that we will be leveraging in this set of experiments is the Postmenopausal Hormone Therapy

(PHT) trial, which was run on postmenopausal women aged 50-79 years who had an intact uterus. This trial included a total of $N_{HT} = 16608$ patients. The intervention of interest was a hormone combination therapy of estrogen and progesterone. Specifically, post-randomization, the treatment group was given 2.5 mg of medroxyprogesterone as well as 0.625 mg of estrogen a day. The control group was given a placebo. Finally, there are several outcomes that were tracked and studied in the principal analysis done on this trial (Rossouw et al., 2002). These outcomes are of three broad categories: a) cardiovascular events, including coronary heart disease, which served as a primary endpoint b) cancer (e.g. endometrial, breast, colorectal, etc.), and c) fractures.

WHI OS – The observational study component of the WHI tracked the medical events and health habits of $N = 93676$ women. Recruitment for the study began in 1994 and participants were followed until 2005, i.e. a similar follow-up to the RCT. Follow-up was done in a similar fashion as in the RCT (i.e. patients would have annual visits, in addition to a "screening" visit, where they would be given survey forms to fill out to track any events/outcomes). Thus, the same outcomes, including cancers, fractures, and cardiovascular events, are tracked in the observational study.

B.5.2 Outcome

The outcome of interest in our analysis is a "global index", which is a summary statistic of several outcomes, including coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. Events or outcomes are tracked for each patient, and are recorded as "day of event/outcome" in the data, where the initial time-point for follow-up is the same for both the RCT and OS. At a high level, the "global index" is essentially the minimum "event day" when considering all the previously mentioned events.

We binarize the "global index," by choosing a time point, t, before the end of follow-up and letting $Y = 1$ if the observed event day is before t and $Y = 0$ otherwise. Thus, we are looking at whether the patient will experience the event within some particular period of time or not. We set $t = 7$ years. Note that we sidestep censorship of a patient before the threshold by defining the outcomes in the following way: $Y = 1$ indicates that a patient is observed to have the event before the threshold, and $Y = 0$ indicates that a patient is not observed to have the event before the threshold. We apply this binarization in the same way for both the RCT and OS. Extending our method to a survival analysis framing is beyond the scope of this chapter, but an interesting direction for future work.

B.5.3 Intervention

Recall from above that the intervention studied in the RCT was 2.5mg of medroxyprogesterone + 0.625 mg of estrogen and the control was a placebo pill. The RCT was run as an “intention-to-treat” trial. To establish “treatment” and “control” groups in the OS, we leverage the annual survey data collected from patients and assign a patient to the treatment group if they confirm usage of both estrogen and progesterone in the first three years. A patient is assigned to the control group if they deny usage of both estrogen and progesterone in the first three years. We exclude a patient from the analysis if she confirms usage of one and not the other OR if the field in the survey is missing OR if they take some other hormone therapy. We end up with a total of $N_{obs} = 33511$ patients.

B.5.4 Data Processing + Covariates

We use only covariates that are measured both in the RCT and OS to simplify the analysis. Because this information is gathered via the same set of questionnaires, they each indicate the same type of covariate. In other words, there is consistency of meaning across the same covariates across the RCT and the OS. We end up with a total of 1576 covariates.

B.5.5 Details of Experimental Setup

We give a more detailed exposition of the steps in our experimental workflow, which were described in brief in the corresponding chapter of this thesis.

- **Step 0**: *Replicate the principal results from the PHT trial, given in Table 2 of (Rossouw et al., 2002), using the WHI OS data.* In this step, we fit a doubly robust estimator of the style given in Appendix B.3.

- **Step 1**: *While treating the WHI OS dataset as the "unbiased" observational dataset (hence the need for Step 0), simulate additional "biased" observational datasets by inducing bias into the WHI OS.* We construct four additional "biased" datasets (for a total of five observational datasets, including the WHI OS dataset), where we use the following procedure to induce selection bias – of the people who were not exposed to the treatment and did not end up getting the event, we drop each person with some probability, p. We set $p = [0.1, 0.3, 0.5, 0.7]$ to get the four additional observational datasets.

  This type of selection bias may reflect the following clinical scenario: consider a patient who is relatively healthy who does not end up taking any hormone therapy. This patient might enroll initially in the OS, but may drop out or stop responding to the surveys. If the committee running the study does not explicitly account for this drop-out rate, then the resultant study will suffer from selection bias. (Banack et al., 2019) detail additional examples of selection bias that can occur in observational studies. Importantly, this part is the only part of our setup that involves any simulation. However, in order to properly evaluate our method, we need to know which datasets are biased and unbiased in our set. Thus, we opt to simulate the bias.

- **Step 2**: *Run our procedure over "multiple tasks," generating confidence intervals on the treatment effect for different subgroups.* To do so, we compile a list of covariates, taking both from (Schnatz et al., 2017) as well as covariates with high feature importance in both the propensity score model and response surface model from the estimator in Step 0. We generate all pairs from this list and use each pair to generate four subgroups. We treat two of the subgroups as validation subgroups and two of them as extrapolated subgroups in that we "hide" the RCT data in those subgroups when fitting our doubly robust transported estimator. (This gives us the benefit of knowing the RCT result for the extrapolated subgroups, which is useful in evaluation). Pairs that do not have enough support (threshold of 400 observations) in each group are removed. The total number of "tasks" (or covariate pairs) that we have is 592 (and therefore 2368 subgroups).

- **Step 3**: *Evaluate ExPCS (our method), ExOCS, Simple, and Meta-Analysis for each of the covariate pairs.* Additionally, we evaluate an "oracle" method, which always selects only the original observational study (i.e. the base WHI OS to which we have not added any selection bias) and reports the interval estimate computed on this study. To evaluate these methods, we will treat the RCT point estimates as "correct." For each, we compute the following metrics: **Length** – length of the confidence interval for the subgroup; **Coverage** – percentage of tasks for which the method's interval covers the RCT point estimate; **Unbiased OS Percentage** – across all tasks, the percentage at which our approach retains the unbiased study after the falsification step.

Note that we utilize sample splitting when running the above procedure. Namely, we use 50% of the data as a "training" set, where we experiment with different classes of covariates and different types of bias, and then reserve 50% of the data as a "testing" set, on which we do the final run of the analysis and report results. All nuisance functions in the doubly robust estimator are fit with a Gradient Boosting Classifier with significant regularization. In practice, we found that any highly-regularized tree-based model works well.

B.5.6 Covariate List for Task Generation

Below is the list of covariates used to generate the tasks in **Step 2** of our experiment:

- ALCNOW (current alcohol user)
- BMI ≤ 30
- BLACK
- SMOKING (current smoker)
- DIAB (diabetes ever)
- HYPT (hypertension ever)
- BRSTFEED (breastfeeding)
- MSMINWK ≤ 106 (minutes of moderate to strenuous activity per week)
- BRSTBIOP (breast biopsy done)
- RETIRED
- EMPLOYED
- OC (oral contraceptive use ever)
- LIVPRT (live with husband or partner)
- MOMALIVE (natural mother still alive)
- LATREGION-Northern > 40 degrees north
- BKBONE (broke bone ever)
- NUMFALLS
- GRAVID (gravidity)
- AGE ≤ 64

- ANYMENSA $\leq$ 51 (age at last bleeding)
- MENOPSEA $\leq$ 50 (age at last regular period)
- MENO $\leq$51 (age at menopause)
- LSTPAPDY (days from randomization to last pap smear)
- BMI $\leq$ 27.7
- TMINWK $\leq$ 191 (minutes of recreational exercise per week)
- HEIGHT $\leq$ 161
- WEIGHT $\leq$ 72
- WAIST $\leq$ 86
- HIP $\leq$ 105
- WHR $\leq$ 0.81 (waist to hip ratio)
- TOTHCAT (HRT duration by category)
- MEDICARE (on medicare)
- HEMOGLBN $\leq$ 13
- PLATELET $\leq$ 244
- WBC $\leq$ 6
- HEMATOCR $\leq$ 40

B.6 Details on Semi-Synthetic Experiment (Data Generation and Model Hyperparameters)

B.6.1 Data Generation

For each simulated dataset, we generate 1 RCT and K observational studies. The RCT is assumed to have covariate values identical to the IHDP dataset but is restricted to infants with married mothers. For the observational studies, we resample the rows of the IHDP dataset to the desired sample size $n = rn_0$. The covariate distribution of the observational studies are made different from the RCT by weighted sampling, with the relative weights set as

$$w = 0.8^{\mathbf{1}(\text{male infant})+\mathbf{1}(\text{mother smoked})+\mathbf{1}(\text{mother worked during pregnancy})}$$

Then, to introduce confounding (in the observational data), we generate m_c continuous confounders and m_b binary confounders. Each continuous confounder is drawn from a mixture of $0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(3,1)$ in the RCT and $(0.25 + 0.5A)\mathcal{N}(3,1) + (0.75 - 0.5A)\mathcal{N}(0,1)$ in the observational studies, where A is the treatment indicator. Similarly, each binary confounder is drawn from $\text{Bern}(0.5)$ in the RCT and $\text{Bern}(0.25 + 0.5A)$ in the observational studies. In the following, we denote the covariate vector as $X \in \mathbb{R}^{m_x}$ where $m_x = 28$ is the number of covariates in the IHDP dataset, and the generated confounder vector as $Z \in \mathbb{R}^{(m_c+m_b)}$. For brevity, we also denote the vector $(A, X^\top)^\top$ as $\tilde{X}$.

For outcome simulation in the datasets, we modify *response surface B* from Hill (2011) to account for additional confounding variables. We set the following counterfactual outcome distributions:

$$\begin{aligned}
&Y_0 \sim \mathcal{N}\left(\exp\left[\left(\tilde{X} + \frac{1}{2}\mathbf{1}\right)^\top \beta\right] + Z^\top\gamma, 1\right)\\
&Y_1 \sim \mathcal{N}(\tilde{X}^\top\beta + Z^\top\gamma + \omega, 1),
\end{aligned}$$

where $\mathbf{1} \in \mathbb{R}^{(m_x+1)}$ is vector of ones, $\beta \in \mathbb{R}^{(m_x+1)}$ is a vector where each element is randomly sampled from $(0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, $\gamma \in \mathbb{R}^{(m_c+m_b)}$ is a vector where each element is randomly sampled from $(0.1, 0.2, 0.5, 0.75, 1)$ with uniform probability, and $\omega = 23$ is a constant chosen to limit the size of the GATEs. The observed outcome is then $Y \coloneqq AY_1 + (1 - A)Y_0$. We then conceal a number of confounders, chosen in order from the highest to lowest weighted, from each observational study to mimic the scenario of unobserved confounding. The number of concealed confounders in each observational study is denoted as $\mathbf{c_z} = (c_{z1}, c_{z2}, ..., c_{zK})$.

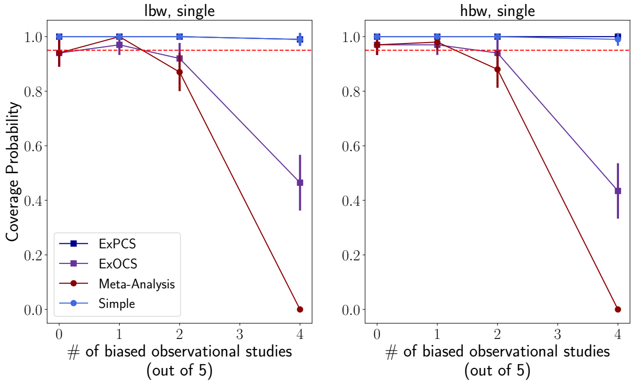
B.6.2 Hyperparameters

Logistic regression

Hyperparametes	Value set
Penalty type	ℓ_2
Penalty coefficient	$\{1, 0.1, 0.01, 0.001\}$

Multilayer perceptron regression

Hyperparametes	Value set
# of hidden layers and # of perceptrons	$[1, (100)], [2, (50, 50)], [2, (25, 25)]$
Activation function	ReLU, tanh
Solver	Adam
Alpha	$(1, 0.1, 0.01, 0.001, 0.0001)$
Learning rate	0.001
# of epochs	$(250, 500)$



Mean width of 95% confidence intervals (4 biased studies)				
	ExPCS	ExOCS	Meta	Simple
LS	5.29	3.55	2.34	5.51
HS	5.84	3.76	2.48	6.15

Figure B-2: *Coverage probabilities of confidence intervals shown as a function of the number of biased observational datasets (out of five). In the 4/5 biased studies case, the average interval widths for each approach is shown for two subgroups. We observe that ExPCS achieves the best balance of interval width and coverage.*

Upsampling ratio	1.0	3.0	5.0	10.0
P(selecting biased study)	0.98	0.80	0.68	0.60

Table B.1: *P(selecting biased study) as a function of upsampling ratio*

B.7 Additional Semi-Synthetic Experimental Results

An analysis of including biased observational studies: In Figure B-2, we study coverage probability and width of confidence intervals in the presence of biased studies. *Meta-Analysis* intervals approach zero coverage probability as the number of biased studies increases. Indeed, a fundamental assumption of this approach is that differences between estimates are only due to random variation, leading to poorer coverage probability when there are more biased studies. *ExOCS* allows for elimination of biased studies in principle through falsification, resulting in improved coverage. However, it does not maintain the desired threshold of coverage (95%), since biased estimators may still be included after falsification either due to chance or by being underpowered. Finally, *ExPCS* and *Simple Union* intervals have good coverage across the board, but as before, *ExPCS* results in narrower intervals.

Overall, we find that our method is robust to biased studies, yielding a good balance between coverage and width. In the case where one has adequate power, *ExOCS* could be a reasonable alternative to get narrower intervals for a sacrifice in coverage (even in the presence of biased studies). However, this implicitly assumes that an estimator consistent for the validation effects will be consistent for the extrapolated effects. If this assumption does not hold, then *ExOCS* will have poor coverage.

Biased estimator selection: In Table B.1, we see that the probability of selecting the biased estimator goes down with increasing sample size of the observational studies, reflected by the increasing sample size ratio, r. This result validates our intuition that our method is more useful and results in more precise estimates of bias as we obtain more observational samples.

B.8 Additional related work

Combining observational and experimental data Prior work has sought to combine RCTs and observational studies for the purpose of more precise estimation of treatment effects (Rosenman et al., 2020, 2021), or for the purpose of generalizing or transporting estimates from RCTs to observational populations when overlap holds between the two (see Degtiar and Rose (2021) for a recent review). In contrast, our work is motivated by settings where there are populations in the observational studies who are not at all represented in trials, e.g., due to a lack of eligibility. Kallus et al. (2018) also seek to combine observational and experimental data to extrapolate beyond tfhe support of an RCT. They propose to learn a CATE function on observational data, and then learn a parametric additive correction term on the sample that overlaps between the RCT and observational data. In contrast to this approach, we do not assume that confounding can be corrected for, and instead seek to choose an observational estimate (if one exists) that is already consistent for each sub-population.

Calibrating observational confidence intervals An alternative method for calibration of confidence intervals for observational studies makes use of negative controls (Lipsitch et al., 2010), such as drug-outcome pairs known to have no causal relationship. Methods range using uses these negative controls to form an empirical null distribution for callibrated p-values(Schuemie et al., 2014), and Schuemie et al. (2018) extend this approach to calibration of confidence intervals. These techniques have been used in several large-scale observational studies such as the LEGEND-HTN study for comparing antihypertensive drugs (Suchard et al., 2019). By contrast, our method does not assume the existence of negative controls, but instead uses a form of positive control (i.e., validation effects), one that is based on the same underlying treatment as the extrapolated effect.

Appendix C

Appendix for Chapter 5

This appendix is organized as follows

- (Section C.1): First, we give a simple 1D example to build intuition for the theoretical results.
- (Section C.2): In the context of Section 5.3.1, we give a concrete example to demonstrate the non-identifiability of Ω_W, defined in (5.12). We focus on the simple case when W is one dimensional, and the matrix Ω_W reduces to a single number $\rho_W \coloneqq \beta_W^2/(\beta_W^2 + \sigma_W^2)$, indicating the signal-to-variance ratio of W. We give an example of an observed distribution for which ρ_W is not identified, and moreover, the optimal predictor with respect to the robustness set $C_A(\lambda)$ is not identified (see Figure C-2).
- (Section C.3): Proofs for results stated in the corresponding chapter of this thesis.
- (Section C.4): Additional results (and proofs) for Proxy Targeted Anchor Regression (PTAR) and Cross-Proxy TAR, deferred from the corresponding chapter of this thesis.
- (Section C.5): Details for implementation of all experiments
- (Section C.6): Additional synthetic experimental results

C.1 An example for building intuition

To illustrate the problem, consider the following setup, where we observe A, X, Y at training time, and wish to learn a predictor $\hat{y} = \alpha + \gamma x$ that will generalize to a new environment where $\mathbb{P}_{te}(A) \neq \mathbb{P}_{tr}(A)$.

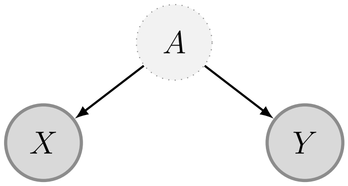


Figure C-1: *Simple example where* $X, Y, A \in \mathbb{R}$.

Suppose that our data is generated under $\mathbb{P}_{tr}$ as follows

$$\begin{aligned} A &= \epsilon_A, & \epsilon_A &\sim \mathcal{N}(0, 1) \\ X &= A + \epsilon_X, & \epsilon_X &\sim \mathcal{N}(0, \sigma_X^2) \\ Y &= A + \epsilon_Y, & \epsilon_Y &\sim \mathcal{N}(0, \sigma_Y^2), \end{aligned}$$

where $\epsilon_A, \epsilon_X, \epsilon_Y$ are jointly independent. This simple example demonstrates a few concepts:

- Assuming $\sigma_X^2 > 0$, the conditional expectation $\mathbb{E}[Y \mid X]$ changes as the distribution of A changes.
- We can write the residuals $Y - \hat{Y}$ as a linear function in A and the noise variables. This holds, even if the errors are non-Gaussian.
- The test population MSE is a convex function of α, γ.

In particular, we will see that the parameters α, γ trade off between the variance of A and ϵ_X: There exists an invariant solution, where $\alpha = 0, \gamma^* = 1$, such that the MSE is completely independent of A, but this is only optimal in the setting where $\mathrm{Var}(A) \to \infty$.

Conditional Expectation depends on A Starting with the assumption that A, X, Y are multivariate Gaussian, we can write down the optimal predictor in the target environment, supposing that at test time $\mathbb{P}_{te}(A) \overset{(d)}{=} \mathcal{N}(\mu_A, \sigma_A^2)$.

$$\begin{aligned}\mathbb{E}_{te}[Y \mid X = x] &= \mathbb{E}_{te}[Y] + \frac{\mathrm{Cov}_{te}(X, Y)}{\mathrm{Var}_{te}(X)} \cdot (x - \mathbb{E}_{te}[X]) \\ &= \mu_A + \underbrace{\frac{\sigma_A^2}{\sigma_A^2 + \sigma_X^2}}_{\gamma} \cdot (x - \mu_A) \\ &= \mu_A(1 - \gamma) + \gamma x,\end{aligned}$$

where if $\epsilon_X = 0$, then $\gamma = 1$ and the optimal solution does not depend on the parameters of A, and is given by

$$\mathbb{E}_{te}[Y|X = x] = x. \tag{C.1}$$

However, for any $\sigma_x^2 > 0$, the optimal solution under $\mathbb{P}_{te}(A)$ depends on μ_A, σ_A^2.

Rewriting residuals Regardless of whether the Gaussian assumption holds, for a given predictor $\hat{Y} = \alpha + \gamma x$, we can write the error $Y - \hat{Y}$ as a function that is linear in A and the noise variables

$$\begin{aligned}Y - \hat{Y} &= (A + \epsilon_Y) - \gamma(A + \epsilon_X) - \alpha \\ &= A(1 - \gamma) + (\epsilon_Y - \gamma\epsilon_X - \alpha).\end{aligned}$$

Optimizing for a known target distribution The mean squared error $\mathbb{E}[(Y - \hat{Y})^2]$ can be written as a function of α, γ, and the mean and variance of A under $\mathbb{P}_{te}(A)$. Here, all expectations are taken with respect to the test distribution.

$$\begin{aligned}\mathbb{E}_{te}[(y - \hat{y})^2] &= \mathbb{E}_{te}[\mathbb{E}_{te}[(y - \hat{y})^2 \mid A]] \\ &= \alpha^2 - 2\alpha\mathbb{E}_{te}[A](1 - \gamma) \\ &\quad + (1 - \gamma)^2\mathbb{E}_{te}[A^2] + \gamma^2\sigma_x^2 + \sigma_y^2.\end{aligned} \tag{C.2}$$

By first-order conditions, this expression is minimized by

$$\alpha^* = \mu_A(1-\gamma^*) \qquad\qquad \gamma^* = \frac{\sigma_A^2}{\sigma_A^2+\sigma_X^2}. \tag{C.3}$$

When $\sigma_A^2 \to \infty$, then $\gamma^* \to 1$ from Equation (C.3). This is intuitive, because in Equation (C.2), $\gamma = 1$ renders the MSE functionally independent of the distribution of A.

Optimizing for a worst-case distribution Equation (C.3) shows the optimal solution under a known target distribution, if μ_A, σ_A^2 were known in advance. However, a similar intuition applies to the case where $\mathbb{P}_{te}(A)$ is unknown, but we expect it to lie in a particular class. Consider interventions of the form $do(A \coloneqq \nu)$, where we constrain ν to lie in the set of random variables $C(\lambda) \coloneqq \{\nu : \mathbb{E}[\nu^2] \leq \lambda\}$. In this case, our worst-case loss is given by

$$\begin{aligned}
&\sup_{\nu \in C(\lambda)} \mathbb{E}_\nu[(Y-\hat{Y})^2] \\
&= \sup_{\nu \in C(\lambda)} (1-\gamma)\left[-2\alpha\mathbb{E}[\nu] + (1-\gamma)\mathbb{E}[\nu^2]\right] \\
&\quad + \alpha^2 + \gamma^2\sigma_X^2 + \sigma_Y^2,
\end{aligned}$$

where the last line does not depend on ν. We observe that $\alpha^* = 0$, by analyzing two cases. First, if $\gamma = 1$, then the first term is eliminated, and the only term that depends on α is α^2. Second, if $\gamma \neq 1$, then $(1-\gamma)^2 > 0$, the first term is partially maximized when $\mathbb{E}[\nu^2] = \lambda$, and if $\alpha \neq 0$, then the expression can be made even larger by choosing a deterministic $\nu = \pm\sqrt{\lambda}$ (instead of e.g., a random $\nu \sim \mathcal{N}(0, \lambda^2)$), depending on the sign of $\alpha(1-\gamma)$. From this (and the presence of the α^2 term in the second line) it follows that $\alpha^* = 0$, in this case as well. When $\alpha = 0$, the supremum is obtained by any random or deterministic ν such that $\mathbb{E}[\nu^2] = \lambda$.

With $\alpha^* = 0$ and taking $\mathbb{E}[\nu^2] = \lambda$ in the supremum, this expression simplifies to

$$\begin{aligned}&\sup_{\nu \in C(\lambda)} \mathbb{E}_\nu[(Y - \hat{Y})^2] \\ &= (1-\gamma)^2 \lambda + \gamma^2 \sigma_X^2 + \sigma_Y^2.\end{aligned}$$

Differentiating with respect to γ, we obtain

$$\gamma^* = \frac{\lambda}{\sigma_X^2 + \lambda}.$$

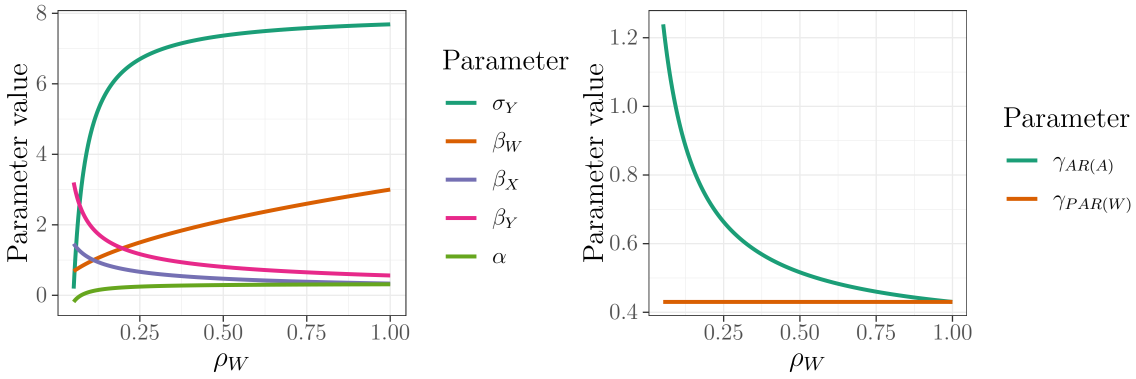
Here, λ trades off accuracy and stability; As $\lambda \to \infty$, we recover the solution where $\gamma^* = 1$, but for situations where σ_X^2 is large and λ is bounded, we are better off choosing $\gamma^* < 1$.

C.2 Example: Non-identifiability of Ω_W

Overview In the context of Section 5.3.1, we give a concrete example to demonstrate the non-identifiability of Ω_W, defined in (5.12). We focus on the simple case when W is one dimensional, and the matrix Ω_W reduces to a single number $\rho_W \coloneqq \beta_W^2/(\beta_W^2 + \sigma_W^2)$, indicating the signal-to-variance ratio of W. We give an example of an observed distribution for which ρ_W is not identified, and moreover, the optimal predictor with respect to the robustness set $C_A(\lambda)$ is not identified (see Figure C-2).

Setup If $(X, Y, W) \in \mathbb{R}^3$ is distributed multivariate normal with zero mean, then their covariance matrix fully determines the observed distribution. Let that covariance matrix be denoted by $\Sigma_{(X,Y,W)} \in \mathbb{R}^{3\times 3}$, which gives us six observed moments of the distribution

$$\Sigma_{(X,Y,W)} \coloneqq \begin{pmatrix} \mathbb{E}[X^2] & \cdot & \cdot \\ \mathbb{E}[XY] & \mathbb{E}[Y^2] & \cdot \\ \mathbb{E}[WX] & \mathbb{E}[WY] & \mathbb{E}[W^2] \end{pmatrix},$$



(a) **(b)**

Figure C-2: *(a) SCM parameters that all give rise to the same observational distribution, and observe that (b) the parameter $\gamma_{AR(A)}$ (as if A were observed) can diverge substantially from the solution $\gamma_{PAR(W)}$, when a single proxy is available. $\lambda = 5$ for this example.*

where we only show the lower triangular portion, since the matrix is symmetric. Suppose that we knew that this observed distribution was generated by the following SCM, but that we do not know the values for the parameters $(\beta_W, \beta_X, \beta_Y, \alpha, \sigma^2_W, \sigma^2_X, \sigma^2_Y)$

$$\begin{aligned} A &\coloneqq \epsilon_A & \epsilon_A &\sim \mathcal{N}(0, 1) \\ W &\coloneqq \beta_W A + \epsilon_W & \epsilon_W &\sim \mathcal{N}(0, \sigma^2_W) \\ X &\coloneqq \beta_X A + \epsilon_X & \epsilon_X &\sim \mathcal{N}(0, \sigma^2_X) \\ Y &\coloneqq \alpha X + \beta_Y A + \epsilon_Y & \epsilon_Y &\sim \mathcal{N}(0, \sigma^2_Y), \end{aligned}$$

where $\epsilon_A, \epsilon_W, \epsilon_X, \epsilon_Y$ are jointly independent. We can attempt to identify the parameters using the following relationships implied by the SCM, and matching these to the moments that we observe

$$\begin{aligned} \mathbb{E}[WX] &= \beta_W \beta_X \\ \mathbb{E}[XY] &= \beta_Y \beta_X + \alpha \mathbb{E}[X^2] \\ \mathbb{E}[WY] &= \beta_W(\beta_Y + \alpha \beta_X) \\ \mathbb{E}[W^2] &= \beta^2_W + \sigma^2_W \end{aligned}$$

$$\mathbb{E}[X^2] = \beta_X^2 + \sigma_X^2$$
$$\mathbb{E}[Y^2] = \alpha^2\mathbb{E}[X^2] + 2\alpha\beta_Y\beta_X + \beta_Y^2 + \sigma_Y^2$$

However, as we will see, this does not identify the parameters. In particular, there is a set of parameterizations which all give rise to the same observed distribution, and which imply different values of the signal-to-variance ratio $\rho_W \coloneqq \beta_W^2/(\beta_W^2 + \sigma_W^2)$.

A class of observationally equivalent SCMs Let $\theta \coloneqq (\beta_W, \beta_X, \beta_Y, \alpha, \sigma_W^2, \sigma_X^2, \sigma_Y^2) \in \mathbb{R}^7$ be the parameters of the SCM, and let $\Sigma = f(\theta)$ be the covariance matrix over (X, Y, W) implied by these parameters.

For any covariance matrix Σ, there exists a subset $C \subset [0, 1]$ such that for any $\rho_W \in C$, we can write the parameters as a function of ρ_W, such that $f(\theta(\rho_W)) = \Sigma$. The set C is constrained by the observed moments: In particular, as we show below, $\rho_W \geq \text{corr}(W, X)^2$ due to the constraint that $\sigma_X^2 \geq 0$, and the condition that $\sigma_Y^2 \geq 0$ also imposes a lower bound. In particular, for the covariance matrix below, we demonstrate numerically that $[0.06, 1] \subset C$.

$$\Sigma_{(X,Y,W)} \coloneqq \begin{pmatrix} 9 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 9 \end{pmatrix}.$$

We now give a strategy for constructing $\theta(\rho_W)$, given a desired ρ_W (including checking the constraint that this $\rho_W \in C$). Suppose that W and X are positively correlated, as in this example. Fixing some $\rho_W \in [0, 1]$, we start by writing β_W, σ_W as functions of ρ_W, where

$$\beta_W \coloneqq \sqrt{\mathbb{E}[W^2]\rho_W}$$
$$\sigma_W^2 \coloneqq \mathbb{E}[W^2](1 - \rho_W).$$

The first constraint, that $\sigma_X^2 \geq 0$, can be captured as follows. Let $\rho_X \coloneqq \beta_X^2/\mathbb{E}[X^2]$.

Observe that $\sqrt{\rho_X \rho_W} = \operatorname{corr}(W, X)$. This implies a lower bound on ρ_W, given by $\rho_W \geq \operatorname{corr}(W, X)^2$, since $\rho_X \leq 1$ due to $\sigma_X^2 \geq 0$. This also implies that ρ_X is determined uniquely by ρ_W, and is given by $\rho_X = \operatorname{corr}(W, X)^2 / \rho_W$. From this we can write

$$\begin{aligned}
\beta_X &\coloneqq \sqrt{\mathbb{E}[X^2]\rho_X} \\
\sigma_X^2 &\coloneqq \mathbb{E}[X^2](1 - \rho_X).
\end{aligned}$$

These choices for $(\beta_W, \sigma_W^2, \beta_X, \sigma_X^2)$ match the observed moments $\mathbb{E}[X^2], \mathbb{E}[W^2], \mathbb{E}[WX]$. Then the rest of the parameters can be found as follows, where β_W, β_X are fixed as above

$$\begin{aligned}
\beta_Y &\coloneqq \frac{1}{\beta_W(1 - \rho_X)} \left(\mathbb{E}[WY] - \frac{\mathbb{E}[XY]\mathbb{E}[WX]}{\mathbb{E}[X^2]} \right) \\
\alpha &\coloneqq \frac{\mathbb{E}[XY] - \beta_Y \beta_X}{\mathbb{E}[X^2]} \\
\sigma_Y^2 &\coloneqq \mathbb{E}[Y^2] - \beta_Y^2 - 2\alpha\beta_Y\beta_X - \alpha^2 \mathbb{E}[X^2]
\end{aligned}$$

where all of these are functions of ρ_W, in that β_W, β_X are functions of ρ_W. It remains to verify that for a given choice of ρ_W, we satisfy the constraint that $\sigma_Y^2 \geq 0$. For simplicity, we check this constraint computationally in the context of Example 1, for a range of values of ρ_W, and we give the set of observationally-equivalent parameters in Figure C-2a, where valid values of ρ_W range over $[0.06, 1]$.

Next we show that the Proxy Anchor Regression estimator, $\gamma_{PAR(W)}$, differs from the Anchor Regression estimator, $\gamma_{AR(A)}$, and more so when ρ_W becomes small. This is shown in Figure C-2b, for $\lambda = 5$, and we give the relevant computations here.

Solution to PAR(W) If we have a single proxy, then we can write down the optimization problem Equation (5.10) as

$$\min_{\gamma} \mathbb{E}[(Y - \gamma X)^2] + \lambda \mathbb{E}[(Y - \gamma X)W]^2 \mathbb{E}[W^2]^{-1}$$

$$
\begin{aligned}
&= \min_{\gamma} \mathbb{E}[Y^2] - 2\gamma \mathbb{E}[YX] + \gamma^2 E[X^2] \\
&\quad + \lambda(\mathbb{E}[YW] - \gamma \mathbb{E}[XW])^2 \mathbb{E}[W^2]^{-1},
\end{aligned}
$$

from which we obtain the optimal solution

$$
\gamma_{PAR(W)} = \frac{\mathbb{E}[YX]\mathbb{E}[W^2] + \lambda \mathbb{E}[YW]}{\mathbb{E}[X^2]\mathbb{E}[W^2] + \lambda \mathbb{E}[XW]}.
$$

Solution to AR(A) First, we can write the residual as

$$
\begin{aligned}
Y - \hat{Y} &= Y - \gamma X \\
&= \alpha X + \beta_Y A + \epsilon_Y - \gamma \beta_X A - \gamma \epsilon_X \\
&= \alpha(\beta_X A + \epsilon_X) + \beta_Y A + \epsilon_Y - \gamma \beta_X A - \gamma \epsilon_X \\
&= A((\alpha - \gamma)\beta_X + \beta_Y) + (\alpha - \gamma)\epsilon_X + \epsilon_Y,
\end{aligned}
$$

such that the expected squared error is given by

$$
\begin{aligned}
&\mathbb{E}_{do(A:=\nu)}(Y - \hat{Y})^2 \\
&= ((\alpha - \gamma)\beta_X + \beta_Y)^2 \mathbb{E}[\nu^2] + (\alpha - \gamma)^2 \sigma_X^2 + \sigma_Y^2, \qquad \text{(C.4)}
\end{aligned}
$$

and when $\nu \in \{\nu : \mathbb{E}[\nu^2] \leq (1 + \lambda)\}$, taking the supremum involves replacing $\mathbb{E}[\nu^2]$ with $(1 + \lambda)$. Optimizing Equation (C.4) with respect to γ, we obtain

$$
\begin{aligned}
&\frac{\partial}{\partial \gamma} \left[((\alpha - \gamma)\beta_X + \beta_Y)^2 (1 + \lambda) + (\alpha - \gamma)^2 \sigma_X^2 + \sigma_Y^2 \right] \\
&= -2\beta_X(\alpha\beta_X - \gamma\beta_X + \beta_Y)(1 + \lambda) - 2(\alpha - \gamma)\sigma_X^2,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
0 &= \beta_X(\alpha\beta_X - \gamma\beta_X + \beta_Y)(1 + \lambda) + (\alpha - \gamma)\sigma_X^2 \\
&= (\alpha\beta_X^2 + \beta_X\beta_Y)(1 + \lambda) - \gamma\beta_X^2(1 + \lambda) + \alpha\sigma_X^2 - \gamma\sigma_X^2,
\end{aligned}
$$

so that the optimal choice of γ is given by

$$\gamma_{AR(A)} = \frac{(\alpha\beta_X^2 + \beta_X\beta_Y)(1+\lambda) + \alpha\sigma_X^2}{\beta_X^2(1+\lambda) + \sigma_X^2}.$$

If $\lambda = -1$, this recovers the causal effect of X on Y, and if $\lambda \to \infty$, this recovers a set of coefficients that are invariant to variation in A, as can be seen by plugging the resulting coefficient $\gamma = \alpha + \beta_Y/\beta_X$ into Equation (C.4).

C.3 Proofs

C.3.1 Auxiliary results

First, we show that the proof of Theorem 1 of Rothenhäusler et al. (2021) can be decomposed into two parts, and use this observation to simplify the proof of our Theorem 5.1. Proposition A1 establishes that ℓ_{PLS} can be written as a quadratic form in the structural parameters $w_\gamma^\top M_A$. Proposition A2 is a straightforward generalization of the techniques used in Rothenhäusler et al. (2021), and establishes that any regularization term that can be written in this way naturally implies a robustness guarantee.

By Assumption 5.1, our SCM can be written in the following form, where $\epsilon \perp\!\!\!\perp A$, and all variables are mean-zero and have bounded covariance.

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (Id - B)^{-1}(M_A A + \epsilon). \tag{C.5}$$

In this context, we use the following notational shorthand,

$$w_\gamma := \left((Id - B)^{-1}_{d_X+1,\cdot} - \gamma^\top (Id - B)^{-1}_{1:d_X,\cdot}\right)^\top, \tag{C.6}$$

such that we can write the residual as a function of both the exogenous noise ϵ and A

as

$$R(\gamma) := Y - \gamma^\top X = w_\gamma^\top(\epsilon + M_A A), \tag{C.7}$$

under the training distribution. (This identity explains the valley in the loss landscape displayed in Figure 5-3: If $d_A \geq 2$, for any parameter γ, there exist an orthogonal intervention direction $\nu \in (w_\gamma^\top M_A)^\perp$, to which the loss is invariant.)

Proposition A1. *Under Assumption 5.1,*

$$\begin{aligned}&\ell_{PLS}(X, Y, A; \gamma)\\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top w_\gamma,\end{aligned} \tag{C.8}$$

and

$$\begin{aligned}&\ell_{PLS}(X, Y, W; \gamma)\\ &= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] M_A^\top w_\gamma,\end{aligned} \tag{C.9}$$

where w_γ is defined by Equation (C.6).

Proof. The first statement follows from Equation (5.6) and the observation that

$$\begin{aligned}\mathbb{E}[R(\gamma)A^\top] &= \mathbb{E}[w_\gamma^\top(\epsilon + M_A A)A^\top]\\ &= w_\gamma^\top \mathbb{E}[\epsilon A^\top] + w_\gamma^\top M_A \mathbb{E}[AA^\top]\\ &= w_\gamma^\top M_A \mathbb{E}[AA^\top],\end{aligned}$$

where we used $\epsilon \perp\!\!\!\perp A$. Similarly

$$\begin{aligned}&\ell_{PLS}(X, Y, W; \gamma)\\ &= \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WR(\gamma)^\top]\\ &= \mathbb{E}[w_\gamma^\top(\epsilon + MA)W^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WR(\gamma)^\top]\\ &= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] M_A^\top w_\gamma,\end{aligned}$$

where the first equality follows from Equation (5.6), and the final equality follows from the fact that $\epsilon \perp\!\!\!\perp W$. $\square$

Proposition A2. *Under Assumption 5.1, for any λ and any real, symmetric Ω such that $0 \preceq \mathbb{E}[AA^\top] + \lambda\Omega$, any loss function of the form*

$$\ell(\gamma, \lambda) := \ell_{LS}(X, Y; \gamma) + \lambda w_\gamma^\top M_A \Omega M_A^\top w_\gamma, \tag{C.10}$$

where w_γ is defined by Equation (C.6), is equal to the following worst-case loss under bounded perturbations

$$\ell(\gamma, \lambda) = \sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2],$$

where

$$C(\lambda) := \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega\}.$$

Proof. We have, making use of the fact that $\epsilon \perp\!\!\!\perp A$, and $\mathbb{E}[\epsilon] = 0$

$$\begin{aligned}
&\sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)} \left[(Y - \gamma^\top X)^2\right] \\
&= \sup_{\nu \in C(\lambda)} \mathbb{E}_{do(A:=\nu)} \left[(w_\gamma^\top(\epsilon + M_A\nu))^2\right] \\
&= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + \sup_{\nu \in C(\lambda)} \mathbb{E}[(w_\gamma^\top M_A \nu)^2] \\
&= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + \sup_{\nu \in C(\lambda)} w_\gamma^\top M_A \mathbb{E}[\nu\nu^\top] {M_A}^\top w_\gamma \\
&= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + w_\gamma^\top M_A(\mathbb{E}[AA^\top] + \lambda\Omega){M_A}^\top w_\gamma \\
&= \mathbb{E}\left[(w_\gamma^\top \epsilon)^2\right] + w_\gamma^\top M_A \mathbb{E}[AA^\top]{M_A}^\top w_\gamma \\
&\quad + \lambda w_\gamma^\top M_A \Omega {M_A}^\top w_\gamma \\
&= \mathbb{E}\left[(w_\gamma^\top(\epsilon + M_A A))^2\right] + \lambda w_\gamma^\top M_A \Omega {M_A}^\top w_\gamma \\
&= \ell_{LS}(X, Y; \gamma) + \lambda w_\gamma^\top M_A \Omega {M_A}^\top w_\gamma \\
&= \ell(\gamma, \lambda),
\end{aligned}$$

where in the fifth line we used the definition of $C(\lambda)$. The supremum is achievable even if ν is a deterministic vector, since we can take $\nu := \frac{Sb}{\sqrt{b^\top Sb}}$ where $S := \mathbb{E}[AA^\top] + \lambda\Omega$ and $b := {M_A}^\top w_\gamma$. Then the supremum value is achieved by ν, as $\nu\nu^\top = \frac{Sbb^\top S}{b^\top Sb}$ and $b^\top \nu\nu^\top b = \frac{b^\top Sbb^\top Sb}{b^\top Sb} = b^\top Sb$. To show that $\nu\nu^\top \preceq S$, such that $\nu \in C(\lambda)$, we can take any conformable vector x to see that

$$\begin{aligned} x^\top(S - \nu\nu^\top)x &= x^\top Sx - \frac{x^\top Sbb^\top Sx}{b^\top Sb} \\ &= \langle x, x\rangle - \frac{\langle x, b\rangle^2}{\langle b, b\rangle} \\ &\geq 0, \end{aligned}$$

where we use the fact that $\langle e, f\rangle := e^\top Sf$ defines an inner product, and we apply Cauchy-Schwarz: $\langle x, x\rangle\langle b, b\rangle \geq \langle x, b\rangle^2$. □

In the proofs for Section 5.3, we will occasionally make use of the following fact, which we prove here to simplify exposition later on.

Proposition A3. *In the setting of a single proxy (i.e., under Assumptions 5.1 and 5.2) let Ω_W be defined as follows*

$$\Omega_W := \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]. \tag{C.11}$$

Then $\Omega_W \preceq \mathbb{E}[AA^\top]$. Furthermore, if $\mathbb{E}[\epsilon_W \epsilon_W^\top]$ is positive definite, then this inequality is strict, that is, $\Omega_W \prec \mathbb{E}[AA^\top]$.

Proof. Recall that $\mathbb{E}[AA^\top]$ and $\mathbb{E}[WW^\top]$ are invertible (and hence positive definite) by assumption.

The inequality $\Omega_W \preceq \mathbb{E}[AA^\top]$ is equivalent to showing that

$$S := \mathbb{E}[AA^\top] - \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \succeq 0. \tag{C.12}$$

Observe that S is the Schur complement of the matrix $K := \mathbb{E}\left[\begin{pmatrix} A \\ W \end{pmatrix}\begin{pmatrix} A \\ W \end{pmatrix}^\top\right]$. The matrix K is positive semi-definite (PSD) if and only if $\mathbb{E}[AA^\top]$ is positive definite (true by assumption) and S is PSD (see Zhang (2006, Theorem 1.12b)). Since K is PSD by construction, as the covariance matrix of A, W, this implies that $S \succeq 0$.

Similarly, K is positive definite (PD) if and only if $\mathbb{E}[AA^\top]$ and S are both PD (see Zhang (2006, Theorem 1.12a)). Under the condition that $\mathbb{E}[\epsilon_W \epsilon_W^\top]$ is full-rank, then K is PD, and the second inequality follows. □

C.3.2 Proof of additional results

Proof of Equation (5.9). It follows from Proposition A1 that

$$\ell_{PLS}(X, Y, A; \gamma) = w_\gamma^\top M_A \Omega_A M_A^\top w_\gamma$$
$$\ell_{PLS}(X, Y, W; \gamma) = w_\gamma^\top M_A \Omega_W M_A^\top w_\gamma,$$

where $\Omega_W := \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]$ and $\Omega_A := \mathbb{E}[AA^\top]$ are both full rank because $\mathbb{E}[AW^\top] = \mathbb{E}[AA^\top]\beta_W$ and by assumptions that $\mathbb{E}[WW^\top], \mathbb{E}[AA^\top]$ and β_W are full rank. Hence both $\ell_{PLS}(X, Y, A; \gamma)$ and $\ell_{PLS}(X, Y, W; \gamma)$ are zero exactly when $w_\gamma^\top M_A = 0$. □

C.3.3 Proof of main results

Section 5.3

Proof of Theorem 5.1. We use the fact that ϵ is mean-zero and independent of both A and W. Recall that

$$\ell_{PAR}(W; \gamma, \lambda) = \ell_{LS}(\gamma) + \lambda \ell_{PLS}(W; \gamma),$$

where we suppress the dependence on X, Y in the notation. Letting w_γ be as defined in Equation (C.6), it follows from Equation (C.9) that

$$\begin{aligned}&\ell_{PLS}(X, Y, W; \gamma)\\ &= w_\gamma^\top M_A \underbrace{\mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top]}_{\Omega_W} M_A^\top w_\gamma.\end{aligned}$$

The statement then follows from the application of Proposition A2, and the fact that $\Omega_W \preceq \mathbb{E}[AA^\top]$ (by Proposition A3), such that $\mathbb{E}[AA^\top] + \lambda\Omega_W \succeq 0$ for all $\lambda \geq -1$. □

Proof of Proposition 5.1. Recall that the guarantee regions are given by

$$\begin{aligned}C_A(\lambda) &= \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\mathbb{E}[AA^\top]\}\\ C_W(\lambda) &= \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top] + \lambda\Omega_W\}\\ C_{OLS} &= \{\nu : \mathbb{E}[\nu\nu^\top] \preceq \mathbb{E}[AA^\top]\},\end{aligned}$$

where

$$\Omega_W = \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top].$$

The fact that $\mathbb{E}[WW^\top]^{-1} \succ 0$ implies $\Omega_W \succeq 0$, and this implies that $C_{OLS} \subseteq C_W(\lambda)$ for $\lambda \geq 0$. Showing $C_W(\lambda) \subset C_A(\lambda)$ amounts to showing that $\Omega_W \prec \mathbb{E}[AA^\top]$, which holds by Proposition A3 when $\mathbb{E}[\epsilon_W\epsilon_W^\top] \succ 0$.

Next, we prove that C_W is monotonically decreasing in the noise $\mathbb{E}[\epsilon_W\epsilon_W^\top]$, in the sense that if $\mathbb{E}[\epsilon_W\epsilon_W^\top] \preceq \mathbb{E}[\eta_W\eta_W^\top]$ then

$$\begin{aligned}&\mathbb{E}_\eta[AW^\top]\mathbb{E}_\eta[WW^\top]^{-1}\mathbb{E}_\eta[WA^\top]\\ &\preceq \mathbb{E}_\epsilon[AW^\top]\mathbb{E}_\epsilon[WW^\top]^{-1}\mathbb{E}_\epsilon[WA^\top],\end{aligned}$$

where $\mathbb{E}_\eta$ is the expectation in the SCM where $W := \beta_W^\top A + \eta_W$ (and similar for $\mathbb{E}_\epsilon$). Suppose that $\mathbb{E}[\epsilon_W\epsilon_W^\top] \preceq \mathbb{E}[\eta_W\eta_W^\top]$. Then $\mathbb{E}_\eta[WW^\top]^{-1} \preceq \mathbb{E}_\epsilon[WW^\top]^{-1}$, and since

$\mathbb{E}_\eta[AW^\top] = \mathbb{E}_\epsilon[AW^\top]$, for any vector $x \in \mathbb{R}^{d_A}$ it holds that,

$$\begin{aligned}
&(\mathbb{E}_\eta[WA^\top]x)^\top \mathbb{E}_\eta[WW^\top]^{-1}(\mathbb{E}_\eta[WA^\top]x) \\
&\leq (\mathbb{E}_\epsilon[WA^\top]x)^\top \mathbb{E}_\epsilon[WW^\top]^{-1}(\mathbb{E}_\epsilon[WA^\top]x).
\end{aligned}$$

This establishes the matrix inequality.

To conclude the proof, suppose that $\mathbb{E}[\epsilon_W \epsilon_W^\top] = 0$, $d_A = d_W$ and that β_W has full rank. It then follows that

$$\begin{aligned}
\Omega_W &= \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \\
&= \mathbb{E}[AA^\top]\beta_W(\beta_W^\top \mathbb{E}[AA^\top]\beta_W)^{-1}\beta_W^\top \mathbb{E}[AA^\top] \\
&= \mathbb{E}[AA^\top]\beta_W \beta_W^{-1}\mathbb{E}[AA^\top]^{-1}{\beta_W^\top}^{-1}\beta_W^\top \mathbb{E}[AA^\top] \\
&= \mathbb{E}[AA^\top],
\end{aligned}$$

such that $C_W(\lambda) = \mathbb{E}[AA^\top] + \lambda\Omega_W = (1+\lambda)\mathbb{E}[AA^\top] = C_A(\lambda)$. □

Proof of Theorem 5.2. Let w_γ be defined as in Equation (C.6). We can write the population quantity as follows, making use of the fact that ϵ, ϵ_Z, and ϵ_W are jointly independent, and that all errors have zero mean.

$$\begin{aligned}
&\ell_\times(W, Z; \gamma) \\
&= \mathbb{E}[(Y - \gamma^\top X)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[Z(Y - \gamma^\top X)^\top] \\
&= \mathbb{E}[w_\gamma^\top (M_A A + \epsilon)W^\top]\mathbb{E}[ZW^\top]^{-1} \\
&\qquad \cdot \mathbb{E}[Z(A^\top M_A^\top + \epsilon^\top)w_\gamma] \\
&= w_\gamma^\top M_A \mathbb{E}[AW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZA^\top]M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[A(A^\top \beta_W + \epsilon_W^\top)] \\
&\qquad \mathbb{E}[(\beta_Z^\top A + \epsilon_Z)(A^\top \beta_W + \epsilon_W^\top)]^{-1} \\
&\qquad \mathbb{E}[(\beta_Z^\top A + \epsilon_Z)A^\top]M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[AA^\top]\beta_W \big(\beta_Z^\top \mathbb{E}[AA^\top]\beta_W\big)^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\beta_Z^\top \mathbb{E}[AA^\top] M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[AA^\top]\beta_W \beta_W^{-1} \mathbb{E}[AA^\top]^{-1} (\beta_Z^\top)^{-1} \\
&\quad \beta_Z^\top \mathbb{E}[AA^\top] M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[AA^\top]\mathbb{E}[AA^\top]^{-1}\mathbb{E}[AA^\top] M_A^\top w_\gamma \\
&= w_\gamma^\top M_A \mathbb{E}[AA^\top] M_A^\top w_\gamma
\end{aligned}
$$

The result follows from Proposition A1. □

In the main text, we state that the xPAR(W, Z) objective is convex in γ and has a closed form solution. We give the proof here:

Proposition A4. *Under Assumptions 5.1, 5.3 and 5.4, the loss in Equation (5.14) is convex in γ, and its minimizer is given by*

$$
\begin{aligned}
\gamma^*_{\times PAR} := &\left(2\mathbb{E}[XX^\top] + \lambda(L + L^\top)\right)^{-1} \\
&\left(2\mathbb{E}[XY^\top] + \lambda(K_1 + K_2)\right),
\end{aligned}
$$

where we define

$$
\begin{aligned}
L &:= \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZX^\top], \\
K_1 &:= \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZY^\top] \\
K_2 &:= \mathbb{E}[XZ^\top]\mathbb{E}[WZ^\top]^{-1}\mathbb{E}[WY^\top].
\end{aligned}
$$

Proof. By Theorem 5.2 and Equation (5.7), $\ell_{\times PAR}(W, Z; \gamma, \lambda) = \ell_{AR}(X, Y, A; \gamma, \lambda)$, and the latter is convex in γ, since it is the sum ℓ_{LS}, which is convex, and $\lambda \ell_{PLS}(X, Y, A; \gamma)$, which is a quadratic form by Proposition A1 and hence convex.

Consequently optimal solution can be found by taking the gradient of $\ell_{\times PAR}(W, Z; \gamma, \lambda) = \ell_{LS} + \lambda \ell_\times$ with respect to γ and equating it to 0. Letting $D := \mathbb{E}[ZW^\top]^{-1}$, we can differentiate $\ell_{\times PAR}$ term wise, using Equation (5.13) to rewrite $\ell_\times$:

$$
0 = 2\gamma^\top \mathbb{E}[XX^\top] - 2\mathbb{E}[YX^\top]
$$

$$
\begin{aligned}
&- \lambda \mathbb{E}[YW^\top] D \mathbb{E}[ZX^\top] \\
&- \lambda \mathbb{E}[YZ^\top] D^\top \mathbb{E}[WX^\top] \\
&+ \lambda \gamma^\top (L + L^\top),
\end{aligned}
$$

where $L := \mathbb{E}[XW^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZX^\top]$. Defining $K_1 := \mathbb{E}[XW^\top]D\mathbb{E}[ZY^\top]$ and $K_2 := \mathbb{E}[XZ^\top]D^\top\mathbb{E}[WY^\top]$, and rearranging, we obtain:

$$
\begin{aligned}
&\gamma^\top (2\mathbb{E}[XX^\top] + \lambda(L + L^\top)) \\
&= 2\mathbb{E}[YX^\top] + \lambda(K_1^\top + K_2^\top),
\end{aligned}
$$

so by transposing and solving for γ, we get the expression from the statement. □

Section 5.4

Proof of Proposition 5.2. Let w_γ be defined by (C.6) and for any γ let $b_\gamma^\top := w_\gamma^\top M_A$. We can write the loss as follows

$$
\begin{aligned}
&\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top(\epsilon + M_A\nu) - \alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top \epsilon + w_\gamma^\top M_A \nu - \alpha)^2] \\
&\overset{\epsilon \perp\!\!\!\perp \nu}{=} \mathbb{E}[(w_\gamma^\top \epsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A \nu - \alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top \epsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&\qquad - \mathbb{E}[(w_\gamma^\top M_A A)^2] + \mathbb{E}[(w_\gamma^\top M_A \nu - \alpha)^2] \\
&= \ell_{LS}(\gamma) - \mathbb{E}[(b_\gamma^\top A)^2] + \mathbb{E}[(b_\gamma^\top \nu - \alpha)^2] \\
&= \ell_{LS}(\gamma) - b_\gamma^\top \mathbb{E}[AA^\top] b_\gamma^\top \\
&\qquad + b_\gamma^\top \mathbb{E}[\nu\nu^\top] b_\gamma - 2\mathbb{E}[b_\gamma^\top \nu]\alpha + \alpha^2 \\
&= \ell_{LS}(\gamma) + b_\gamma^\top \left(\mathbb{E}[\nu\nu^\top] - \mathbb{E}[AA^\top]\right) b_\gamma \\
&\qquad - 2\mathbb{E}[b_\gamma^\top \nu]\alpha + \alpha^2 \\
&= \ell_{LS}(\gamma)
\end{aligned}
$$

$$\begin{aligned} &+ b_\gamma^\top \left(\mathbb{E}[\nu\nu^\top] - \mathbb{E}[AA^\top]\right) b_\gamma - \left(b_\gamma^\top \mathbb{E}[\nu]\right)^2 \\ &+ \left(b_\gamma^\top \mathbb{E}[\nu]\right)^2 - 2\mathbb{E}[b_\gamma^\top \nu]\alpha + \alpha^2 \\ = {}& \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma + \left(b_\gamma^\top \mathbb{E}[\nu] - \alpha\right)^2, \end{aligned}$$

where for any value of γ, that minimizing with respect to α yields $\alpha^* = b_\gamma^\top \mathbb{E}[\nu]$, where $b_\gamma^\top = w_\gamma^\top M_A$. Given that we can write the structural relationship $Y - \gamma^\top X = b_\gamma^\top A + w_\gamma^\top \epsilon$, and knowing that $\mathbb{E}[\epsilon] = 0$ and that $\epsilon \perp\!\!\!\perp A$, we know that $b_\gamma^\top A$ is the conditional expectation of $R(\gamma)$ given A. □

In the main text, we note that Equation (5.16) (the objective function ℓ_{TAR}) is convex in γ, α, and has a closed form solution. We prove that result here.

Proposition A5. *Under Assumption 5.1, the minimizer $\gamma^*_{TAR}, \alpha^*_{TAR}$ of Equation (5.16) is given by*

$$\begin{aligned} \gamma^* &= \left(\mathbb{E}[XX^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AX^\top]\right)^{-1} \\ &\quad \left(\mathbb{E}[XY^\top] + \mathbb{E}[XA^\top]\Omega\mathbb{E}[AY^\top]\right) \\ \alpha^* &= b_{\gamma^*}^\top \mu_\nu, \end{aligned}$$

where $\Omega = \mathbb{E}[AA^\top]^{-1}(\Sigma_\nu - \Sigma_A)\mathbb{E}[AA^\top]^{-1}$, and $b_\gamma^\top$ is defined in Equation (5.15).

Proof of Proposition A5. Let w_γ be as defined in Equation (C.6) and let $b_\gamma^\top := w_\gamma^\top M_A$. Since $\mathbb{E}[(Y - \gamma^\top X) \mid A] = \mathbb{E}[w_\gamma^\top [M_A A + \epsilon] \mid A] = b_\gamma^\top A$, for any γ, $b_\gamma^\top$ is the linear regression coefficient of $(Y - \gamma^\top X)$ onto A, so we may write $b_\gamma^\top = \mathbb{E}[(Y - \gamma^\top X)A^\top]\mathbb{E}[AA^\top]^{-1}$. Plugging in the optimal value $\alpha(\gamma) := b_\gamma^\top \mu_\nu$, we obtain

$$\begin{aligned} &\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha(\gamma)) \\ &= \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma \\ &= \ell_{LS}(\gamma) + \mathbb{E}[(Y - \gamma^\top X)A^\top]\Omega\mathbb{E}[A(Y - \gamma^\top X)^\top] \end{aligned}$$

This objective is convex in γ. The derivative of the loss with respect to γ is

$$-2(\mathbb{E}[(Y-\gamma^\top X)X^\top]+\mathbb{E}[(Y-\gamma^\top X)A^\top]\Omega\mathbb{E}[AX^\top]),$$

and equating to 0 and solving for γ yields

$$\begin{aligned}\gamma^* = &\left(\mathbb{E}[XX^\top]+\mathbb{E}[XA^\top]\Omega\mathbb{E}[AX^\top]\right)^{-1}\\ &\left(\mathbb{E}[XY^\top]+\mathbb{E}[XA^\top]\Omega\mathbb{E}[AY^\top]\right).\end{aligned}$$

□

We also claim in the main text that if ν is a constant, then the minimizer of Equation (5.16) can be found by performing OLS using both X, A as predictors, and then plugging in the known value ν for A in prediction. We prove that result here.

Proof. If ν is a constant, then we can write the first two terms as follows, where w_γ is defined in Equation (C.6).

$$\begin{aligned}&\ell_{LS}-b_\gamma^\top\Sigma_A b_\gamma\\ &=\mathbb{E}[(w_\gamma^\top(M_A A+\epsilon))^2]-w_\gamma^\top M_A\mathbb{E}[AA^\top]M_A^\top b_\gamma\\ &=\mathbb{E}[(w_\gamma^\top(M_A A+\epsilon))^2]-\mathbb{E}[(w_\gamma^\top M_A A)^2]\\ &=\mathbb{E}[(w_\gamma^\top\epsilon)^2]\end{aligned}$$

which is equivalent to the objective for the loss when Y, X are residualized with respect to A (see Section 8.6 of Rothenhäusler et al. (2021)). By the Frish-Waugh-Lovell theorem (Lovell, 1963, 2008), this yields the same coefficients γ for X as if we had performed regression on X, A together. For this value of γ, $b_\gamma^\top$ is the coefficient that we would obtain for A in the joint regression, because it equals the regression coefficients for $Y-\gamma^\top X$ on A. □

Proof of Proposition 5.3. We use ν to denote the random shift. Let $\nu \in T(\mu_\nu, \Sigma_\nu)$, or equivalently, let $\nu := \mu_\nu + \delta$, where μ_ν is fixed and δ satisfies the constraint that

$\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$, where Σ_ν is a symmetric positive definite matrix. Let w_γ be defined by (C.6) and for any γ let $b_\gamma^\top := w_\gamma^\top M_A$. We can write the loss as follows

$$\begin{aligned}
&\sup_{\nu\in T} \mathbb{E}_{do(A:=\nu)}[(Y-\gamma^\top X-\alpha)^2] \\
&= \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top(\epsilon+M_A\nu)-\alpha)^2] \\
&= \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top\epsilon+w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top\epsilon)^2] + \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \mathbb{E}[(w_\gamma^\top\epsilon)^2] + \mathbb{E}[(w_\gamma^\top M_A A)^2] \\
&\qquad - \mathbb{E}[(w_\gamma^\top M_A A)^2] + \sup_{\nu\in T} \mathbb{E}[(w_\gamma^\top M_A\nu-\alpha)^2] \\
&= \ell_{LS}(\gamma) - \mathbb{E}[(b_\gamma^\top A)^2] + \sup_{\nu\in T} \mathbb{E}[(b_\gamma^\top\nu-\alpha)^2],
\end{aligned}$$

where on the fourth line we used the fact that $\mathbb{E}[\epsilon\nu]=0$ by the fact that $\nu=\mu_v+\delta$, and δ is independent of ϵ. In the last line we replaced $w_\gamma^\top M_A$ by $b_\gamma^\top$. We can re-write the last term as follows, where the supremum with respect to δ is constrained in the set $\mathbb{E}[\delta\delta^\top] \preceq \Sigma_\nu$

$$\begin{aligned}
&\sup_{\nu\in T} \mathbb{E}[(b_\gamma^\top\nu-\alpha)^2] \\
&= \sup_{\delta:\mathbb{E}[\delta\delta^\top]\preceq\Sigma_\nu} \mathbb{E}[(b_\gamma^\top(\delta+\mu_\nu)-\alpha)^2] \\
&= \sup_{\delta} \mathbb{E}[(b_\gamma^\top\delta+b_\gamma^\top\mu_\nu-\alpha)^2] \\
&= \sup_{\delta} \mathbb{E}[(b_\gamma^\top\delta)^2] + 2\mathbb{E}[(b_\gamma^\top\delta)](b_\gamma^\top\mu_\nu-\alpha) + \mathbb{E}[(b_\gamma^\top\mu_\nu-\alpha)^2] \\
&= b_\gamma^\top\Sigma_\nu b_\gamma + 2\left\|b_\gamma\right\|_{\Sigma_\nu} \cdot \left|b_\gamma^\top\mu_\nu-\alpha\right| + (b_\gamma^\top\mu_\nu-\alpha)^2,
\end{aligned}$$

where $\left\|b_\gamma\right\|_{\Sigma_\nu} := \sqrt{b_\gamma^\top\Sigma_\nu b_\gamma}$ is the norm induced by the inner product defined with respect to Σ_ν. In the last line, we have used the fact that the expression is maximized (subject to the constraint) by the deterministic distribution $\delta_* = \pm\frac{\Sigma_\nu b_\gamma}{\sqrt{b_\gamma^\top\Sigma_\nu b_\gamma}}$ where the sign depends on the sign of $(b_\gamma^\top\mu_\nu-\alpha)$: δ_* satisfies $b_\gamma^\top\delta_*\delta_*^\top b_\gamma = b_\gamma^\top\Sigma_\nu b_\gamma$, maximizing the first term. Further, the second term is also maximized by δ_*, because if any other random or deterministic δ satisfies $|\mathbb{E}b_\gamma^\top \delta| > |b_\gamma^\top \delta_*|$, it follows by Jensens inequality that $\mathbb{E}[(b_\gamma^\top \delta)^2] \geq (\mathbb{E}[(b_\gamma^\top \delta)])^2 > (b_\gamma^\top \delta_*)^2 = b_\gamma^\top \Sigma_\nu b_\gamma$, such that $\mathbb{E}[\delta\delta^\top] \succ \Sigma_\nu$, so δ is not in the set over which the supremum is taken. Consequently, the supremum is attained at δ_*, because δ_* maximizes both terms.

Using this expression for the supremum, we can write the objective as

$$\begin{aligned}
&\sup_{\nu \in T} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2] \\
&= \ell_{LS}(\gamma) + b_\gamma^\top (\Sigma_\nu - \Sigma_A) b_\gamma \\
&\qquad + 2 \left\|b_\gamma\right\|_\Sigma \cdot \left|b_\gamma^\top \mu_\nu - \alpha\right| + (b_\gamma^\top \mu_\nu - \alpha)^2,
\end{aligned}$$

for which the optimal choice of α^* is given by $b_\gamma^\top \mu_\nu$, for any γ, and for this choice of α, we can see that $\gamma^* = \arg\min_\gamma \ell_{LS}(\gamma) + b_\gamma^\top \left(\Sigma_\nu - \Sigma_A\right) b_\gamma$. □

C.4 Targeting with proxies

Definition C.1 (Proxy Targeted Anchor Regression)**.** Let $\tilde{\mu} := \mathbb{E}_{do(A:=\nu)}[W]$ denote the mean of W under intervention, and let $\tilde{\Sigma}_W := \mathrm{Cov}_{do(A:=\nu)}(W)$ denote the covariance. We define

$$\begin{aligned}
&\ell_{PTAR}(W; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) \\
&= \ell_{LS}(\gamma) + c_\gamma^\top \left(\tilde{\Sigma}_W - \Sigma_W\right) c_\gamma + (c_\gamma^\top \tilde{\mu} - \alpha)^2,
\end{aligned} \tag{C.13}$$

where $c_\gamma^\top := \mathbb{E}[R(\gamma) W^\top] {\Sigma_W}^{-1}$.

As mentioned in the main text, Equation (C.13) is not generally equal to Equation (5.16), and does not generally yield the optimal predictor under the targeted loss. A simple example is given in Proposition A6.

Proposition A6. *Assume Assumptions 5.1, 5.2, and that* $\mathbb{E}[\epsilon_W \epsilon_W^\top]$ *is full rank. Let* $\nu \overset{(d)}{=} A + \eta$ *for the deterministic vector* $\eta^T = \mathbb{E}[R(\gamma_{OLS}^*) A^\top]$*, where* $\overset{(d)}{=}$ *indicates* *equality of distribution, and assume $\eta \neq 0$. Then, the minimizers of Equations (5.16) and (C.13) differ, in that*

$$\alpha^*_{PTAR} < \alpha^*_{TAR}$$

and if $d_W = d_A = 1$, and A has unit variance, then $\frac{\alpha^*_{PTAR}}{\alpha^*_{TAR}} = \rho_W$, where $\rho_W := \beta_W^2/(\beta_W^2 + \mathbb{E}[\epsilon_W^2])$.

Proof. The assumption that $\nu = A + \eta$ implies that $\Sigma_\nu - \Sigma_A = 0$, and $\mathbb{E}[\nu] = \eta$. That is, we have changed the mean of the distribution, but not the covariance. This implies

$$\begin{aligned} \mathbb{E}[\tilde{W}] &= \beta_W^\top \mathbb{E}[\nu] = \beta_W^\top \eta \\ \Sigma_{\tilde{W}} - \Sigma_W &= \beta_W^\top (\Sigma_\nu - \Sigma_A)\beta_W = 0, \end{aligned}$$

where in the second equation we use the fact that $\Sigma_W = \beta_W^\top \mathbb{E}[AA^\top]\beta_W + \mathbb{E}[\epsilon_W \epsilon_W^\top]$ (and similarly for $\Sigma_{\tilde{W}}$), and the ϵ_W terms cancel in the subtraction. We can then write both objectives as follows

$$\begin{aligned} &\ell_{PTAR}(W, \tilde{W}; \gamma, \alpha) \\ &\quad = \ell_{LS}(\gamma) + \left(c_\gamma^\top \beta_W^\top \eta - \alpha\right)^2 \\ &\quad = \ell_{LS}(\gamma) + \left(\mathbb{E}[R(\gamma)A^T]\beta_W \Sigma_W^{-1} \beta_W^\top \eta - \alpha\right)^2 \\ &\ell_{TAR}(A, \nu; \gamma, \alpha) \\ &\quad = \ell_{LS}(\gamma) + \left(b_\gamma^\top \eta - \alpha\right)^2 \\ &\quad = \ell_{LS}(\gamma) + \left(\mathbb{E}[R(\gamma)A^T]\Sigma_A^{-1}\eta - \alpha\right)^2 \end{aligned}$$

This gives the optimal value of α in both cases as the value that minimizes the second term

$$\begin{aligned} \alpha^*_{PTAR} &= \mathbb{E}[R(\gamma^*_{PTAR})A^T](\beta_W \Sigma_W^{-1} \beta_W^\top)\eta \\ \alpha^*_{TAR} &= \mathbb{E}[R(\gamma^*_{TAR})A^T]\Sigma_A^{-1}\eta, \end{aligned}$$

and since the second term can be made equal to zero by these choices of α, the optimal

γ in both cases is identically $\gamma^*_{PTAR} = \gamma^*_{TAR} = \gamma^*_{OLS}$, the value of γ that minimizes the first term $\ell_{LS}(\gamma)$. Hence, we can write the difference between these terms as

$$\begin{aligned}&\alpha^*_{TAR} - \alpha^*_{PTAR} \\ &\quad = \mathbb{E}[R(\gamma^*_{OLS})A^T](\Sigma_A^{-1} - \beta_W \Sigma_W^{-1} \beta_W^\top)\mathbb{E}[AR(\gamma^*_{OLS})],\end{aligned}$$

where we have replaced η with the assumed value of $\mathbb{E}[AR(\gamma^*_{OLS})]$. By assumption, Σ_A is full-rank, so that matrix $\Omega := (\Sigma_A^{-1} - \beta_W \Sigma_W^{-1} \beta_W^\top)$ is positive definite if and only if $\Sigma_A \Omega \Sigma_A$ is positive definite. Working with this representation, we can see that

$$\begin{aligned}\Sigma_A \Omega \Sigma_A &= \Sigma_A - \Sigma_A \beta_W \Sigma_W^{-1} \beta_W^T \Sigma_A \\ &= \mathbb{E}[AA^\top] - \mathbb{E}[AW^\top]\mathbb{E}[WW^\top]^{-1}\mathbb{E}[WA^\top] \\ &\succ 0,\end{aligned}$$

where the last line follows from Proposition A3. In the case where $d_W = d_A = 1$, and A has unit variance, then let $\rho_W = \beta_W^2/(\beta_W^2 + \mathbb{E}[\epsilon_W^2])$, and observe that

$$\alpha^*_{PTAR} = \eta^2 \rho_W \qquad\qquad \alpha^*_{TAR} = \eta^2.$$

□

Proposition A6 describes a worst-case mean-shift in A, where η is taken in the direction that maximizes the loss of the OLS solution γ^*_{OLS}. This is also a particularly simple case to analyze for building intuition, because the optimal solution to both Equations (5.16) and (C.13) is to take $\gamma = \gamma^*_{OLS}$ and to estimate an intercept term α equal to the bias incurred by the shift in the mean of A. However, the noise in W results in under-estimating the impact of the shift, and the gap to the optimal solution depends on the signal-to-variance relationship in W, which (as discussed in Section 5.3) is not generally identified.

We also prove that the Cross-Proxy Targeted Anchor Regression objective is equal to that of Targeted Anchor Regression.

Theorem C.1. *Under Assumptions 5.1, 5.3, and 5.4, for all $\gamma \in \mathbb{R}^{d_X}, \alpha \in \mathbb{R}$,*

$$\ell_{\times TAR}(W, Z; \tilde{\mu}, \tilde{\Sigma}_W, \gamma, \alpha) = \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X - \alpha)^2]$$

where $\tilde{\mu} := \mathbb{E}_{do(A:=\nu)}[W]$ is the mean of W under intervention, and $\tilde{\Sigma}_W$ is the covariance $\tilde{\Sigma}_W := Cov_{do(A:=\nu)}(W)$.

Proof of Theorem C.1. We have

$$\begin{aligned} a_\gamma^\top &= \mathbb{E}[R(\gamma)Z^\top](\mathbb{E}[WZ^\top])^{-1} \\ &= \mathbb{E}[R(\gamma)(A^\top \beta_Z + \epsilon_Z^\top)] \\ &\qquad \mathbb{E}[(\beta_W^\top A + \epsilon_W)(\beta_Z^\top A + \epsilon_Z)^\top]^{-1} \\ &= \mathbb{E}[R(\gamma)A^\top]\beta_Z(\beta_W^\top \mathbb{E}[AA^\top]\beta_Z)^{-1} \\ &= \mathbb{E}[R(\gamma)A^\top](\mathbb{E}[AA^\top])^{-1}(\beta_W^\top)^{-1}, \end{aligned}$$

while

$$\begin{aligned} \tilde{\mu} &= \beta_W^\top \mathbb{E}[\nu] \\ \tilde{\Sigma}_W - \Sigma_W &= \beta_W^\top(\Sigma_\nu - \Sigma_A)\beta_W. \end{aligned}$$

With $b_\gamma^\top := w_\gamma^\top M_A$ and w_γ defined by (C.6), we have that

$$\begin{aligned} a_\gamma^\top \tilde{\mu} &= b_\gamma^\top \mathbb{E}[\nu] \\ a_\gamma^\top(\tilde{\Sigma}_W - \Sigma_W)a_\gamma &= b_\gamma^\top(\Sigma_\nu - \Sigma_A)b_\gamma, \end{aligned}$$

which is equivalent to $\ell_{TAR}(A; \mu_\nu, \Sigma_\nu, \gamma, \alpha)$ (Definition 5.4, Equation (5.16)). The proof is complete by Proposition 5.2. □

Note that the argument is symmetric for using an observed shift in either Z or W, so it suffices to know the anticipated shift with respect to one proxy.

C.5 Details for experiments

C.5.1 Details of Section 5.5.1

We outline the details of the simulation experiment in Section 5.5.1.

Summary We simulate a training data set $\mathcal{D}_{\text{train}}$ from a SCM that induces the structure in Figure 5-2, fix $\lambda := 5$ and fit estimators PAR(W) and xPAR(W, Z). We consider the intervention $\mathbb{P}_{do(A:=\nu)}$ with $\nu = (-2.83, 0.35, 0.71)^\top$, and simulate a test data set $\mathcal{D}_{\text{test}}$ from that distribution. We then compute the intervention mean squared prediction error (MSPE) $\hat{\mathbb{E}}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$ both for PAR(W) and xPAR(W, Z). We repeat this procedure $m = 10^5$ times for several signal-to-variance ratios x (not including 0), and display the quantiles of the losses in Figure 5-5. We also plot the population losses $\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$ for PAR(W) and xPAR(W, Z), as well as AR(A) and OLS.

Technical details We let $\mathbb{E}[AA^\top] = \beta = \mathrm{Id}$ and $\mathbb{E}[\epsilon_W \epsilon_W^\top] = s^2\,\mathrm{Id}$, such that $W = \beta^\top A + s \cdot \epsilon_W$. Then Ω_W as defined in Equation (5.11) simplifies to

$$\begin{aligned}\Omega_W &= \mathbb{E}[AA^\top]\beta(\beta^\top\mathbb{E}[AA^\top]\beta + \mathbb{E}[\epsilon_W\epsilon_W^\top])^{-1}\beta^\top\mathbb{E}[AA^\top] \\ &= \quad \frac{1}{1+s^2}\,\mathrm{Id}\,.\end{aligned}$$

We call $x = (1+s^2)^{-1}$ the signal-to-variance ratio, and we can obtain a given signal-to-variance ratio x, by setting $s = \sqrt{(1-x)/x}$.

For each $n \in \{150, 500\}$ and signal-to-variance ratio $x \in \{1/20, 2/20, \dots, 20/20\}$, we set $s = \sqrt{(1-x)/x}$ and sample a data set $\mathcal{D}_{n,s}^i$ for $i = 1, \dots, 5000$, each with sample size n, from the structural equations:

$$\begin{aligned}A &:= \epsilon_A \qquad \text{(C.14)} \\ W &:= A + s \cdot \epsilon_W\end{aligned}$$

$$Z := A + s \cdot \epsilon_Z$$
$$(Y, X, H) := (\mathrm{Id} - B)^{-1}(MA + \epsilon),$$

where $d_A = d_W = d_Z = d_X = 3$, $d_Y = d_H = 1$. M and B are given by

$$M = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 2 & 1 \\ -1 & 3 & 0 \\ 2 & 2 & -3 \\ 0 & -2 & 2 \end{pmatrix}, B = \begin{pmatrix} 0 & -2 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and all noise variables are i.i.d., $\epsilon_A, \epsilon_W, \epsilon_Z, \epsilon \sim \mathcal{N}(0, \mathrm{Id})$. For every combination (n, s) we have 5000 data sets $\mathcal{D}^i_{n,s}$, $i = 1, \dots, 5000$. For each data set, we compute the proxy estimators $\gamma^i_{n,s,W}$ and $\gamma^i_{n,s,W;Z}$, using one or two proxies respectively, and we simulate 5000 corresponding test data sets of size n from $\mathbb{P}_{do(A:=\nu)}$ (using the structural equations above, except for changing the assignment for A to $A := \nu$). The prediction MSE for the i'th test data set is then $\frac{1}{n}\sum_{j=1}^{n}(Y_j - \gamma^\top X_j)^2$, resulting in 5000 values of the MSE for each combination of (n, s).

At each combination of (n, s) we plot the median by a line of the estimated worst case losses, and by a shaded region indicate the interval between the 25% and 75% quantiles of the observed distribution. We plot the median instead of the mean since for small x, $s^2 = \frac{1-x}{x}$ is large, and especially for $\mathrm{WCL}^i_{n,s}(W, Z)$ and $n = 150$, the mean will be driven very much by outliers for small x.

The population versions of losses for any s is computed first by computing the population estimators γ from the parameter matrices M, B, and then computing the loss at ν by $\mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2] = w_\gamma^\top M\nu\nu^\top M^\top w_\gamma + w_\gamma^\top \mathbb{E}[\epsilon\epsilon^\top]w_\gamma$.

C.5.2 Details of Section 5.5.2

We outline the details of the simulation experiment in Section 5.5.2.

Summary We analyze the effect of applying anchor regression with one proxy, PAR(W), when the signal-to-variance ratio is potentially misspecified. To do so, we simulate data from the same SCM as in Section 5.5.1 ($n = 10^4$), and in particular from a range of true (unknown) signal-to-variance ratios $x \in (0, 1]$. To each data set, we apply anchor regression with one proxy, PAR(W), and with $\lambda := 5$. We further assume the signal-to-variance ratio to be 40% – independently of its true value. This means, by Proposition 5.1, that we assume that PAR(W) minimizes the worst case mean squared prediction error (MSPE) over the region $C := \{\nu\nu^\top \preceq (1 + 0.4 \cdot \lambda)\mathbb{E}[AA^\top]\}$, with the worst case MSPE for being equal to the optimal value of the PAR(W) objective function. If $x = 0.4$, then PAR(W) indeed minimizes the worst case MSPE over C and the estimated worst case MSPE over C is close to the actual worst case MSPE over C. But if $x \neq 0.4$, the estimator minimizes the worst case MSPE over a different set, and then expect that the true worst case MSPE over C differs from its estimate. Figure 5-6 shows that this is indeed the case: We observe that if the true signal-to-variance ratio is larger than the assumed 40%, our estimate of the MSPE is too conservative. On the contrary, if the true signal-to-variance ratio is smaller than assumed, our estimates of the MSPE over C are too small, meaning that we underestimate the worst case MSPE in the region C.

Technical details For a fixed signal-to-variance ratio x, we simulate a training data set $\mathcal{D}_{\text{train}}$ ($n = 10^4$) from the same procedure as in Section C.5.1, i.e. using the structural equations in Equation (C.14), and with the same parameters M and B. We fit the PAR(W) estimator to the data using $\lambda := 5$, and the estimated worst case mean squared prediction error (MSPE) over C is then the value of the objective function in the estimated parameter (by Theorem 5.1).

To find the actual worst case MSPE over C for a given estimator λ, we use the fact from Equation (C.7) that

$$\mathbb{E}_{do(A:=v)}[(R - \gamma^\top X)^2] = (b_\gamma^\top v)^2 + w_\gamma^\top w_\gamma, \tag{C.15}$$

where we use that $\mathbb{E}[\epsilon\epsilon^\top] = \mathrm{Id}$, w_γ is given by Equation (C.6) and $b_\gamma^\top = w_\gamma^\top M_A$. The second term doesn't depend on v, and since C is spherical, the worst case MSPE over C is attained in the direction $v \propto b_\gamma$, with v normalized such that $\|v\|^2 = (1 + 0.4 \cdot \lambda)$ (that is v lies on the boundary of C). Using the known M and B, we compute w_γ, b_γ, and the actual worst case MSPE over C is given by Equation (C.15) plugging in $v = b_\gamma \cdot \sqrt{(1 + 0.4 \cdot \lambda)}/\|b_\gamma\|$.

We compute also the worst case MSPE over C when using an OLS estimator for the prediction. We fit $\hat{\gamma}_{OLS}$ from $\mathcal{D}_{train}$, and, as for the actual MSPE of PAR(W), the worst case MSPE over C using OLS can be computed, by computing vectors $b_{\hat{\gamma}_{OLS}}, w_{\hat{\gamma}_{OLS}}$. Again the worst case MSPE over C using $\hat{\gamma}_{OLS}$ is attained by setting $v = b_{\hat{\gamma}_{OLS}} \cdot \sqrt{(1 + 0.4 \cdot \lambda)}/\|b_{\hat{\gamma}_{OLS}}\|$ and plugging v, $b_{\hat{\gamma}_{OLS}}$ and $w_{\hat{\gamma}_{OLS}}$ into Equation (C.15).

For every signal-to-variance ratio $x \in \{1/20, \ldots, 20/20\}$, we repeat the procedure $m = 1000$ times, for each computing the estimated and actual MSPEs. In Figure 5-6 we plot the median MSPE as well as the interval from the 25% quantile to the 75% quantile.

C.5.3 Details of Section 5.5.3

We outline the details of the simulation experiment in Section 5.5.3.

Summary We demonstrate the ability of Proxy Anchor Regression to select invariant predictors, in a synthetic setting where predictors X may contain both causal and anti-causal predictors. We simulate data sets ($n = 10^5$) from a SCM with the structure shown in Figure 5-7 (top), where one anchor, A_1, is a parent of the causal predictors, while the other A_2 is a parent of the anti-causal predictors.

We consider two identically distributed noisy proxies W, Z of $A := (A_1, A_2)$. The challenge, in this scenario, is that A_2 is measured with significantly more noise than A_1, across both proxies. As a consequence, proxy anchor regression with one proxy, PAR(W), puts more weight on anti-causal features: the noise in W is mistaken for fluctuations in A_2, resulting in $X_{\text{anti-causal}}$ mistakenly appearing invariant to shifts in A_2. In contrast, when two proxies W, Z are available, the estimator xPAR(W, Z) asymptotically equals that of anchor regression with observed anchors, and its regression coefficients puts more weight on the causal predictors; see Figure 5-7 (bottom).

Technical details With $d_{A_1} = d_{A_2} = d_W = d_W = 6$, $d_{X_{\text{causal}}} = d_{X_{\text{anti-causal}}} = 3$ and $d_Y = 1$, we simulate data from the SCM in Figure 5-7 (top) which amounts to simulating from the following structural equations:

$$\begin{aligned}
A_1 &:= \epsilon_{A_1} \\
A_2 &:= \epsilon_{A_2} \\
W &:= (A_1, A_2)^\top + (\epsilon_{W,1}, \epsilon_{W,2})^\top \\
Z &:= (A_1, A_2)^\top + (\epsilon_{Z,1}, \epsilon_{Z,2})^\top \\
X_{\text{causal}} &:= M_1 A_1 + \epsilon_{X_{\text{causal}}} \\
Y &:= \gamma_{\text{causal}}^\top X_{\text{causal}} + \epsilon_Y \\
X_2 &:= M_2 A_2 + \gamma_{\text{anti-causal}} Y + \epsilon_{X_{\text{anti-causal}}}.
\end{aligned}$$

Here $M_1 \in \mathbb{R}^{d_{X_{\text{causal}}} \times d_{A_1}}$ and $M_2 \in \mathbb{R}^{d_{X_{\text{anti-causal}}} \times d_{A_2}}$ are matrices with 1 in every entry, $\gamma_{\text{causal}} = (1/4, 1/4, 1/4)^\top$ and $\gamma_{\text{anti-causal}} = (4, 4, 4)^\top$ (such that the regression coefficients of Y onto $X_{\text{causal}}, X_{\text{anti-causal}}$ are of similar magnitudes). All noise terms are independent and $\epsilon_{A_1}, \epsilon_{A_2}, \epsilon_{X_{\text{causal}}}, \epsilon_{X_{\text{anti-causal}}}, \epsilon_Y \sim \mathcal{N}(0, \text{Id})$, and $\epsilon_{W,1}, \epsilon_{Z,1} \sim \mathcal{N}(0, \text{Id})$, $\epsilon_{W,2}, \epsilon_{Z,2} \sim \mathcal{N}(0, 3^2 \cdot \text{Id})$.

We simulate a data set $\mathcal{D}$ $(n = 10^5)$ from these structural equations, and fit the proxy anchor regression estimators $\gamma(W)$ and $\gamma(W, Z)$ from Section 5.3. We repeat this $m = 10^4$ times, and display the mean absolute value of the regression coefficients (that is the entries of the vectors $\gamma(W)$ and $\gamma(W, Z)$) in Figure 5-7 (bottom), as well as the standard deviation of the absolute value of the regression coefficients as error bars.

C.5.4 Details of Section 5.5.4

Summary We demonstrate the trade-off made by Targeted Anchor Regression (TAR) versus Anchor Regression (AR), considering the case when A is observed for simplicity. We simulate training data and fit estimators γ_{OLS}, γ_{AR} and γ_{TAR}, where γ_{TAR} is targeted to a particular mean and covariance of a random intervention $do(A := \nu)$, and we select λ for γ_{AR} such that this intervention is contained within $C_A(\lambda)$. We then simulate test data from two distributions: $\mathbb{P}_{do(A:=\nu)}$ (i.e., the shift occurs), and $\mathbb{P}$ (where it does not), and evaluate the mean squared prediction error (MSPE). The results are shown in Figure 5-8, and demonstrated that TAR performs better than AR and OLS in the first scenario, but this comes at the cost of worse performance on the training distribution.

Technical details The entire procedure below produces a prediction MSE for each of three methods and two settings, and we repeat this $m = 10^5$ times, to produce the histograms of MSEs shown in Figure 5-8.

We simulate a training data set $\mathcal{D}_{\text{train}}$ ($n_{\text{train}} = 10^5$) from the structural equations

$$\begin{aligned} A &:= \epsilon_A \\ (Y, X, H) &:= (\text{Id} - B)^{-1}(MA + \epsilon), \end{aligned}$$

where $d_A = d_X = 2$ and $d_Y = d_H = 1$, $\epsilon_A, \epsilon \sim \mathcal{N}(0, \text{Id})$ and M and B were selected by a simulation resulting in:

$$M = \begin{pmatrix} 2 & 1 \\ 0 & 1 \\ 2 & 2 \\ 0 & 3 \end{pmatrix}, B = \begin{pmatrix} 0 & -0.06 & 0.07 & 0.04 \\ 0.05 & 0 & 0.19 & 0.03 \\ 0.11 & -0.11 & 0 & 0.1 \\ -0.02 & 0.02 & 0.09 & 0 \end{pmatrix}.$$

We consider the target distribution $do(A := \kappa^\top A + \eta)$ where

$$\kappa = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{pmatrix}, \eta = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and so we fit the targeted AR estimator $(\gamma_{\text{targeted-AR}}, \alpha_{\text{targeted-AR}})$ from Equation (5.16), where the covariance of the anticipated shift is given by $\Sigma_\nu := \kappa^\top \mathbb{E}[AA^\top]\kappa$, and the mean shift is simply η. We also fit OLS estimates $\gamma_{\text{OLS}}(X, Y)$ and $\gamma_{\text{AR}}(X, Y, A)$ where for AR we select λ such that $(1+\lambda)$ equals the largest eigenvalue of $\kappa^\top \mathbb{E}[AA^\top]\kappa + \eta\eta^\top$, such that $\mathbb{E}[(\kappa^\top A + \eta)(\kappa^\top A + \eta)^\top] \preceq (1+\lambda)\mathbb{E}[AA^\top]$.

We then simulate a test data set ($n_{\text{test}} = 10^5$) both from 1) the training distribution (i.e. same simulation procedure as for the training set) or 2) by changing the structural equation for A to $A := \kappa^\top \epsilon_A + \eta$, and keeping all other quantities as for the simulation of training data (i.e. the test distribution is the anticipated distribution). We evaluate the prediction MSE on each of the data sets by $\frac{1}{n_{\text{test}}} \sum_j (Y_j - \gamma^\top X_j)^2$ (including the term $\alpha_{\text{targeted-AR}}$ for the targeted AR).

C.5.5 Details of Section 5.6

Features The dataset contains time-stamps as well as season indicators, which we do not use anywhere as features. The remaining features are Dew Point (Celsius Degree), Temperature (Celsius Degree), Humidity (%), Pressure (hPa), Combined wind direction (NE, NW, SE, SW, or CV, indicating calm and variable), Cumulated wind speed (m/s), Hourly precipitation (mm), and Cumulated precipitation (mm).

Data Processing Each city has PM2.5 readings from multiple sites, which we average to get a single reading, and we take a log transformation. For Precipitation (Cumulative) we subtract off the (current hour) precipitation to avoid co-linearity. We take a log transformation of the variable for Wind Speed, Precipitation (Hourly) and Precipitation (Cumulative), due to skewness. We drop all rows that contain any missing data.

Proxies (Temperature) We use temperature as our proxy variable, and treat it as unavailable at test time. We construct two synthetic proxies of temperature to serve as W, Z, adding independent Gaussian noise while controlling the signal-to-variance ratio (in the training distribution) at $\mathrm{Var}(A)/\mathrm{Var}(W) = 0.9$. This results in different standard deviations of the Gaussian noise across different environments, because of differences in the training distributions across training seasons and cities. The standard error of the noise varies between 2 and 5 degrees, to maintain the same signal-to-variance ratio.

Training Details (PAR, xPAR) For the distributional robustness approaches described in Section 5.3, we choose $\lambda \in [0, 40]$ by leave-one-group-out cross-validation on the three training seasons, using the first year (2013) of data. For Proxy Anchor Regression using Temperature directly, there is heterogeneity in the cross-validated choice of λ: In 9 out of 20 scenarios, $\lambda = 40$ is chosen, but in the remaining 11, $\lambda = 0$ is chosen, which is equivalent to OLS. We saw a similar result when the maximum value of λ was 20, and increased the maximum limit to 40 without seeing much difference, so we did not increase it further. Concretely, with λ in [0, 20], there are some scenarios where PAR (TempC) has slightly worse or slightly better MSE (vs. λ in [0, 40]), but the differences are all less than 0.001. The only observable difference in Table 5.1 when running with λ in [0, 20] is that the "best" performance is -0.040 ($\lambda = 20$), as opposed to -0.041 ($\lambda = 40$) [where lower is better, rounded to nearest 0.001]. For Proxy Anchor Regression using W and for Cross-Proxy Anchor Regression (xPAR) using W, Z together, we use the same values of λ as above, for comparability.

Training Details (PTAR, xPTAR) For the targeted approaches described in Section 5.4, we use the mean and variance of the temperature in the test distribution to target our predictors, and similarly use the distribution of the proxies when using Proxy Targeted Anchor Regression (PTAR) with W and Cross-Proxy TAR (xPTAR)

Table C.1: *MSE (lower is better) over 20 scenarios consisting of five cities and four held-out seasons. Average difference to OLS estimator (lower is better) given in the second column, and minimum / maximum difference in remaining columns.*

Estimator	Mean	Diff	Min	Max
OLS	0.457			
OLS (TempC)	0.455	-0.002	-0.028	0.026
OLS + Est. Bias	0.474	0.018	-0.072	0.150
PAR (TempC)	0.454	-0.003	-0.041	0.006
PAR (W)	0.454	-0.002	-0.037	0.006
xPAR (W, Z)	0.454	-0.003	-0.039	0.007
PTAR	0.450	-0.007	-0.061	0.002
PTAR (W)	0.452	-0.005	-0.038	0.001
xPTAR (W, Z)	0.450	-0.007	-0.059	0.003

with W, Z. Note that xPTAR (unlike xPAR) is asymmetric in the proxies, but in this case the proxies are distributed identically.

Benchmarks As described in the main text, our primary benchmark is OLS, trained on the three training seasons, evaluated on the held-out season. We also include two other baselines: First, OLS that has access to temperature during both train and test, which we denote OLS (TempC), and OLS that includes temperature during training, and attempts to estimate a bias term by plugging in the mean (test) value for temperature during prediction.

In Table C.1 we give the full results over all 20 scenarios, which includes the 11 scenarios where $\lambda = 0$ is chosen by cross-validation, rendering the PAR and xPAR solutions equivalent to OLS.

Regularization paths In Figure C-4 we have shown how the solution in the "best" scenario differs for Proxy Anchor Regression (PAR) with $\lambda = 40$ versus OLS (i.e., $\lambda = 0$). In Figure C-5, we show how the coefficients change in-between these two extremes: for every integer value of λ in [0, 40] we show the difference in the PAR vs. OLS coefficients for each feature. Increasing λ further does not make a significant difference for this particular example.

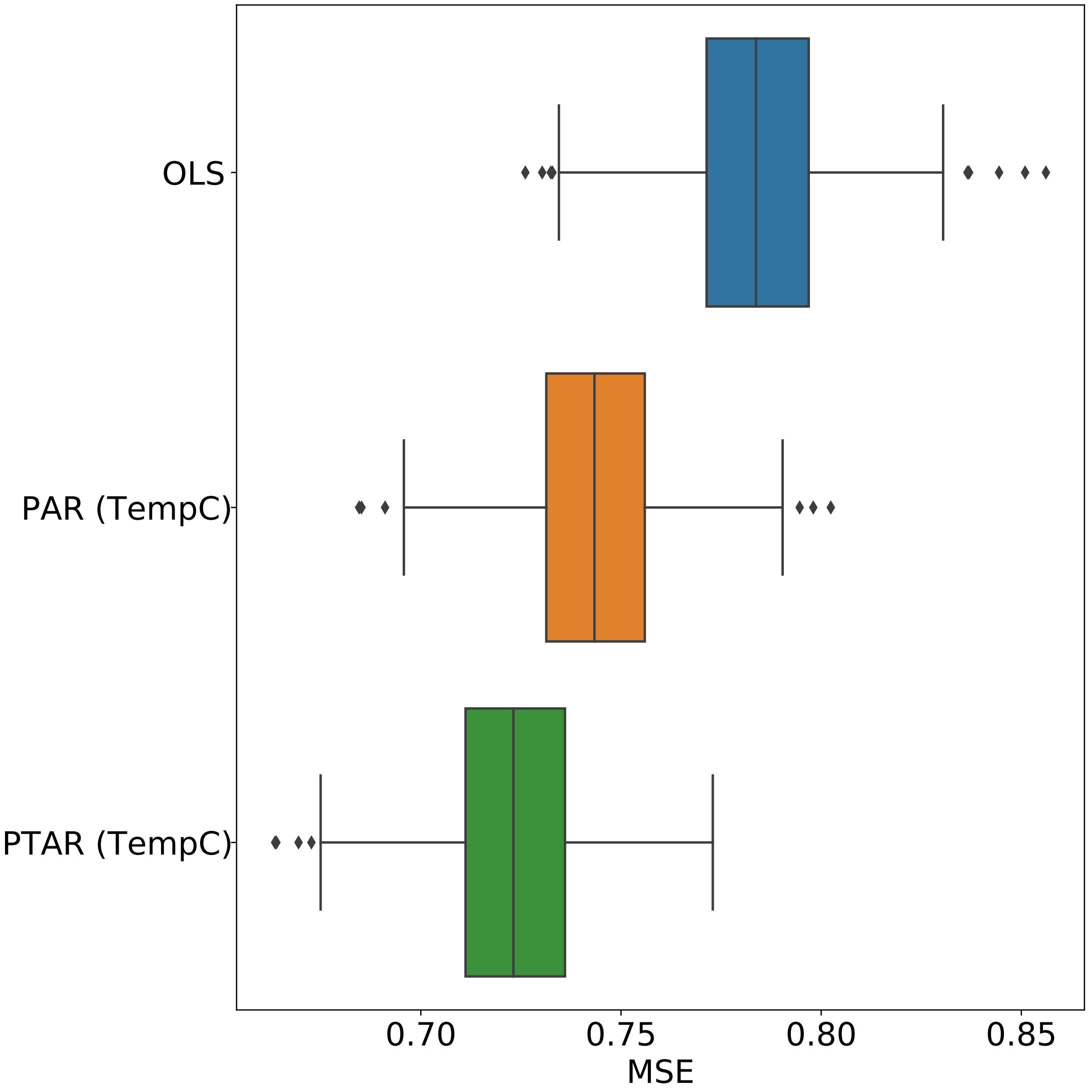


Figure C-3: *Best performance for Proxy Anchor Regression (PAR) and Proxy Targeted AR (PTAR), corresponding to Summer in Beijing. Variance estimates generated by bootstrapping the test residuals of the fitted models.*

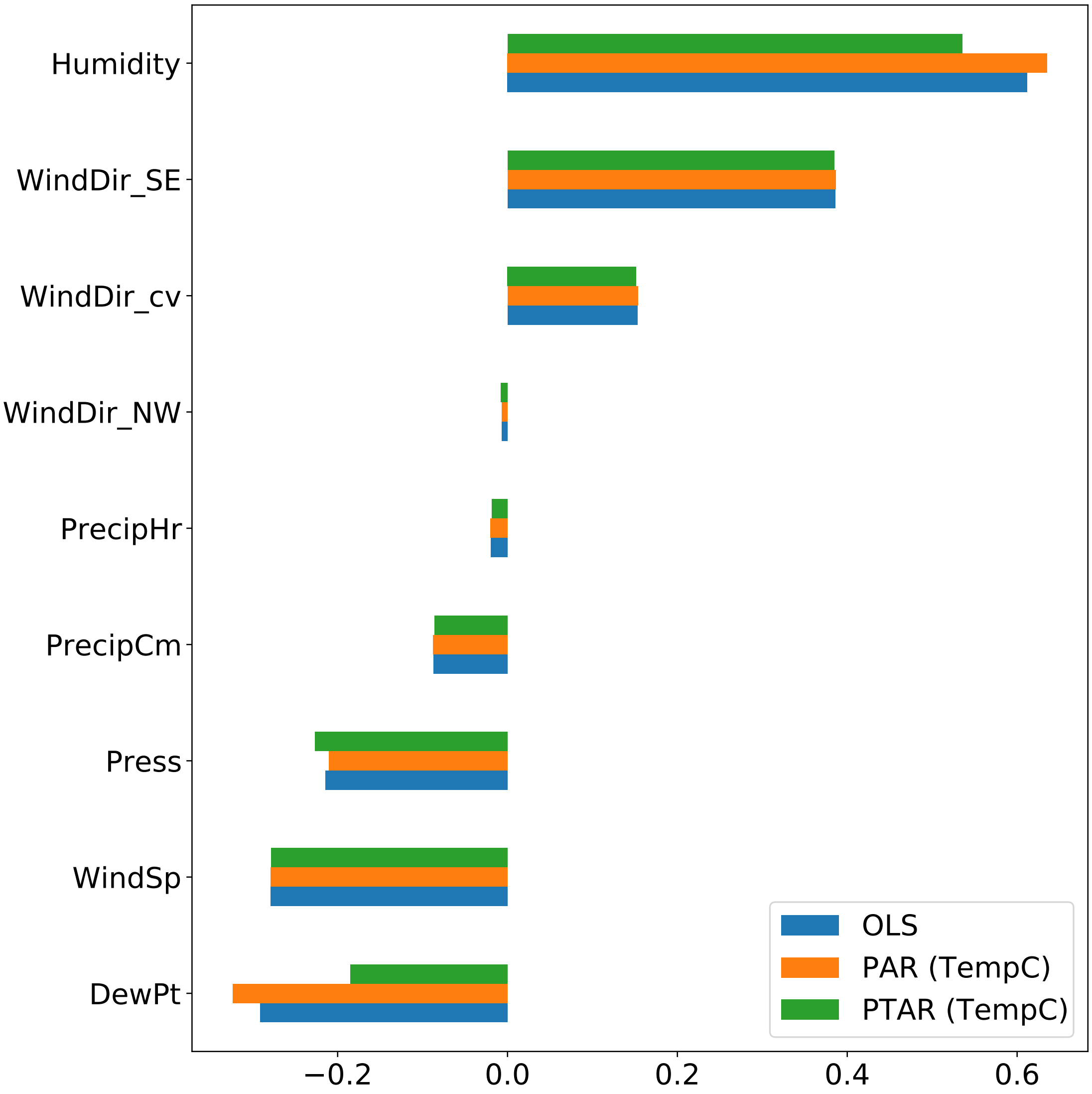


Figure C-4: *Comparison of learned coefficients. All variables were standardized to unit variance. The intercept for OLS and AR is the same (by construction) at* $\alpha = 4.087$ *while the intercept for TAR is lower at* $\alpha = 3.885$.

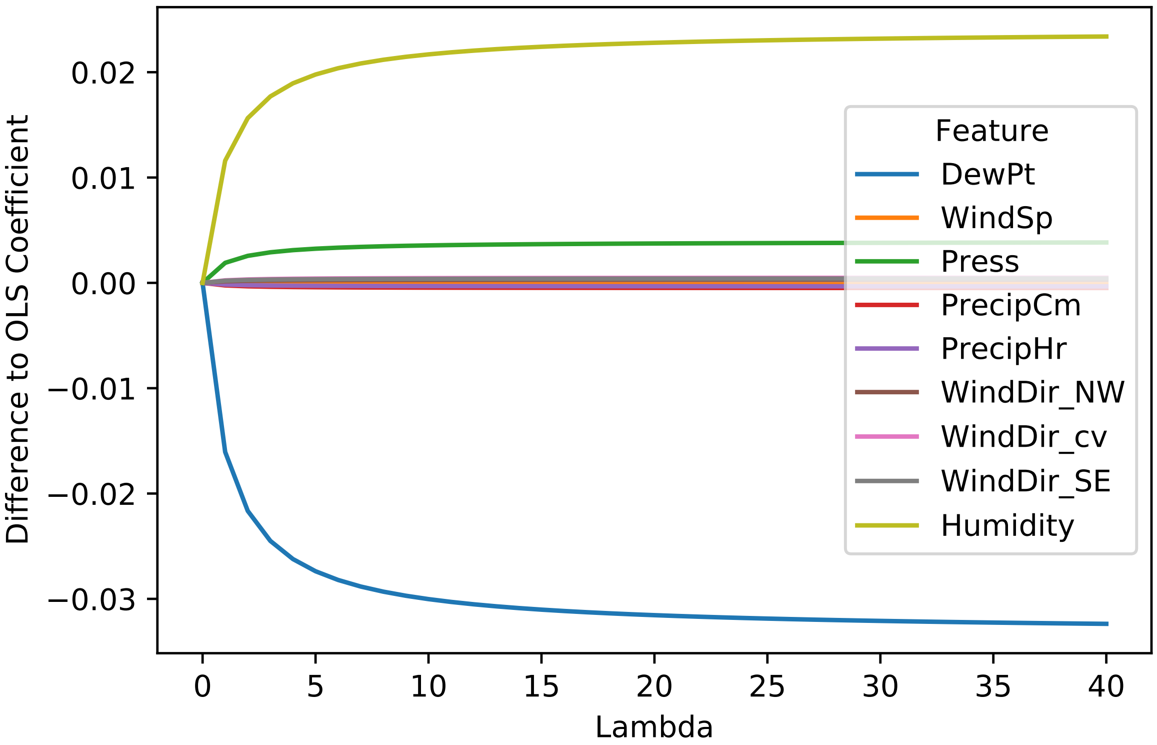


Figure C-5: *Coefficient path, showing the difference between the PAR and OLS coefficients in Figure C-4 for different values of* λ*.*

C.6 Additional experiment: Signal-to-variance ratio

To examine the effect of the signal strengths β_W and β_Z, we scale the signals $\beta_{W,s} = \beta_{Z,s} = s\,\mathrm{Id}$ for $s \in \{0, \sqrt{2/3}, 0.8\}$, which for the single proxy estimator $\hat{\gamma}_{\mathrm{PAR}}$ amounts to optimizing over worst case loss in the robustness regions $C(\lambda) = \{vv\top \preceq (1+\lambda\frac{s^2}{1+s^2})\,\mathrm{Id}\}$.

For $s \in \{1, 3\}$, such that the signal-to-variance ratio $\frac{s^2}{1+s^2}$ equals either 10% or 50%, we simulate a training data set $\mathcal{D}_{\mathrm{train}}$ with two proxies W and Z from the structural equations $A := \epsilon_A, (X^\top, Y^\top, H^\top)^\top := (1-B)^{-1}(M_a A + \epsilon), W := \beta_{W,s}^\top A + \epsilon_W$ and $Z := \beta_{Z,s}^\top A + \epsilon_Z$ where all noise terms are i.i.d with unit covariance and M_A, B are given by:

$$M := \begin{pmatrix} 2 & 1 \\ 0 & 1 \\ 2 & 2 \\ 0 & 3 \end{pmatrix}, B := \begin{pmatrix} 0 & -0.57 & 0.73 & 0.37 \\ 0.53 & 0 & 1.91 & 0.33 \\ 1.14 & -1.13 & 0 & 0.96 \\ -0.22 & 0.16 & 0.87 & 0 \end{pmatrix}.$$

Since for this experiment we are not interested in finite sample properties of the estimators, we use sample size $n = 10^7$.

For each data set we fit estimators $\hat{\gamma}_{\mathrm{PAR}(W)}$ (using only one proxy), $\hat{\gamma}_{\mathrm{xPAR(W,\ Z)}}$ (using both proxies), $\hat{\gamma}_{\mathrm{AR}(A)}$, and $\hat{\gamma}_{\mathrm{OLS}}$, and evaluate the estimators at data sampled from interventional distributions $\mathbb{P}_{do(A:=v)}$ for several interventions v of increasing strength (i.e. increasing distance from $\mathbb{E}[A] = 0$).

As the signal to variance ratio increases, the PAR(W) loss approaches the AR(A). Further we observe that xPAR(W, Z) coincides with the $AR(A)$ estimator for both signal-to-variance levels. This is illustrated in Figure C-6.

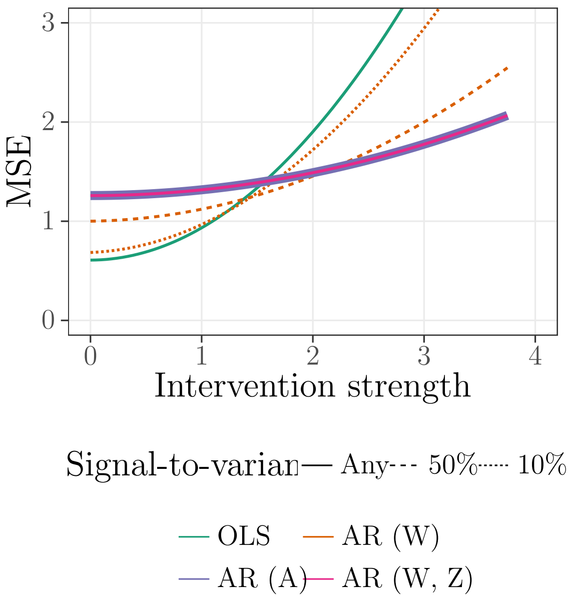


Figure C-6: *Anchor and proxy estimators for different levels of signal-to-variance ratio* $\beta(\mathbb{E}[WW^\top])^{-1}\beta^\top$*. A training data set (*$n = 10^7$*) with two proxies* W, Z *is simulated and the estimators* $\hat{\gamma}_{PAR(A)}$*,* $\hat{\gamma}_{xPAR(W,Z)}$*,* $\hat{\gamma}_{AR(A)}$*, and* $\hat{\gamma}_{OLS}$ *are fitted using a fixed* λ*. Interventions* v *of increasing strength is sampled, and for each a new data set (*$n = 10^5$*) is sampled from* $\mathbb{P}^{do(A:=v)}$*, and for each estimator* $\hat{\gamma}$*, the prediction mean squared error* $\mathbb{E}_{do(A:=(v_1,v_2))}[(Y - \hat{\gamma}^\top X)^2]$ *is computed. This procedure is repeated for signal-to-variance ratios* 10% *and* 50%*.*

Appendix D

Appendix for Chapter 6

This appendix is structured as follows:

- In Appendix D.1, we provide details on the synthetic lab testing example, including how we generate the loss landscape in Figure 6-1b.

- In Appendix D.2, we provide a "user's guide" to defining and interpreting parametric shifts, including worked examples for many common conditional distributions, as well as guidance on how to define and interpret the shift functions $s(Z; \delta)$.

- In Appendix D.3, we provide additional details on the worst-case optimization problem, as well as comparisons of the reweighting-based approach to the Taylor approximation approach. We also demonstrate that the quadratic approximation is exact, for particularly simple structural causal models.

- In Appendix D.4, we compare our approach to that of worst-case conditional subpopulation shifts, in the context of a simpler laboratory testing example where we can explicitly compute the worst-case conditional subpopulations. Here, we demonstrate that our approach can capture more realistic intuition regarding which shifts are plausible in practice.

- In Appendix D.5, we give additional experimental details, as well as illustrative samples from the generative model, for the CelebA experiment described in Section 6.4.

- In Appendix D.6, we give an extended discussion of related work.

- In Appendix D.7, we give proofs for all the results in the corresponding chapter of this thesis.

D.1 Details of Figure 6-1b

In Figure 6-1b, we consider the following, artificial, generative model, which resembles the setup in Section 6.4.1, but with the addition of age as a continuous variable.

$$\begin{aligned} \text{Age} &\sim \mathcal{N}(0, 0.5^2) \\ \mathbb{P}(\text{Disease} = 1|\text{Age}) &= \text{sigmoid}(0.5 \cdot \text{Age} - 1) \\ \mathbb{P}(\text{Order} = 1|\text{Disease, Age}) &= \text{sigmoid}(2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1) \\ \text{Test Result}|\text{Order} = 1, \text{Disease} &\sim \mathcal{N}(-0.5 + \text{Disease}, 1) \end{aligned}$$

where if Order $= 0$, the test result is a placeholder value of zero. In Figure 6-1b, we consider a simple predictive model: If lab tests are not available (Order $= 0$), this model predicts disease based on an unregularized logistic regression model, which uses age to predict disease. If a lab test is available, then it uses both age and the lab test for prediction. This model is trained on 100,000 samples from the training distribution. To construct the loss landscape shown in Figure 6-1b, we first observe that

$$\mathbb{P}(O = 1|\text{Disease}, \text{Age}) = \text{sigmoid}(\eta(\text{Disease}, \text{Age})),$$

where

$$\eta(\text{Disease}, \text{Age}) = 2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1.$$

We construct shifts using the shift function $s(\text{Disease}, \text{Age}; \delta) = \delta_0 \cdot (1 - \text{Disease}) + \delta_1 \cdot \text{Disease}$, and for a grid of values for $(\delta_0, \delta_1) \in [-5, 5]^2$ we consider perturbed distributions with a different conditional distribution of testing,

$$\mathbb{P}_\delta(O = 1|\text{Disease, Age}) = \text{sigmoid}\left(\eta(\text{Disease}, \text{Age}) + \delta_0 \cdot (1 - \text{Disease}) + \delta_1 \cdot \text{Disease}\right),$$

but where all other parts of the generative model are fixed. For each value of $(\delta_0, \delta_1) \in [-5, 5]^2$, we draw 10,000 samples from the corresponding distribution, and compute the negative log-likelihood of the original predictive model under this new distribution. The resulting surface is plotted in Figure 6-1b.

D.2 A user's guide to defining parametric shifts

In this section, we discuss practical considerations in designing parametric shift functions for different distributions.

- In Appendix D.2.1, we give examples of conditional exponential families, illustrative shift functions, and how to interpret them.

- In Appendix D.2.2, we formalize the idea that one can choose shift functions which depend on additional variables, other than the causal parents of a variable W_i.

- In Appendix D.2.3 we give guidance on how to define shift functions when the parameters $\eta(Z)$ are constrained to lie in a particular domain, which is relevant for considering shifts such as changing the variance of a conditional Gaussian.

Table D.1: *Examples of conditional exponential family distributions.*

Distribution	Parameter space	Sufficient statistic	Inverse parameter map
Binary(p)	$\eta(Z) \in \mathbb{R}$	$T(W) = W$	$p(W = 1\|Z) = \text{sigmoid}(\eta(Z))$
Categorical$(p_1, \dots, p_k)$	$\eta(Z) \in \mathbb{R}^k$	$[T(W)]_i = \mathbf{1}\{W = i\}$	$\mathbb{P}(W = i\|Z) = [\text{softmax}(\eta(Z))]_i$
Poisson(λ)	$\eta(Z) \in \mathbb{R}$	$T(W) = W$	$\lambda = \exp(\eta(Z))$
Gaussian(μ, σ^2)	$\eta(Z)_1 \in \mathbb{R}, \eta(Z)_2 < 0$	$T(W) = (W, W^2)$	$\mu(Z) = -\frac{\eta(Z)_1}{2\eta(Z)_2}, \sigma^2(Z) = -\frac{1}{2\eta(Z)_2}$
Gamma(α, β)	$\eta(Z)_1 > -1, \eta(Z)_2 < 0$	$T(W) = (\log W, W)$	$\alpha(Z) = \eta(Z)_1 + 1, \beta(Z) = -\eta(Z)_2$

D.2.1 Conditional exponential family models and interpretations of shifts

In this section, we give examples of exponential families and their sufficient statistics, and discuss design considerations in specifying the shift function $s(Z; \delta)$. Here, we restrict attention to shifts in a single variable, for ease of notation. In Table D.1 we give examples of conditional exponential families, along with their typical parameterizations. In the examples below, we review how shift functions $s(Z; \delta)$ impact these parameters, and how they can also be interpreted on the scale of more commonly considered parameters (e.g., conditional means and variances).

Example D.2.1 (Log-odds shift in a binary variable)**.** Consider the distribution of a binary variable W conditioned on variables Z. Without loss of generality, we can write that

$$\mathbb{P}(W = 1|Z) = \sigma(\eta(Z))$$

where σ is the sigmoid function, and $\eta(Z)$ is an arbitrary measurable function of Z, taking on values in the extended real line $\eta(Z) \in \mathbb{R} \cup \{-\infty, +\infty\}$. This can be written in canonical form as

$$\mathbb{P}(W|Z) = \exp\left\{\eta(Z) \cdot W - \log(1 + \exp^{\eta(Z)})\right\}$$

where $\eta(Z)$ is the canonical parameter (the log-odds ratio), $T(W) = W$ is the sufficient statistic, and $h(\theta) = \log(1 + \exp^{\eta(Z)})$ is the normalizing constant. We can consider shifts $\eta_\delta(Z) := \eta(Z) + \delta$, yielding the new conditional distribution

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta),$$

which is well-defined for any $\delta \in \mathbb{R}$.

Here, we note that these shifts occur on the "natural" parameter scale $\eta(Z)$ (e.g., the log-odds), which at first glance may seem difficult to interpret: Why should we care about changes on the log-odds scale, instead of on the original probability scale? In addition to mathematical convenience, we argue that in some settings, working with natural parameters is advantageous for retaining a common scale across across multiple variables.

For instance, consider shifts in the two independent variables W_1 and W_2, where $V_i \sim \text{Bernoulli}(p_i)$, with $p_1 = 10^{-4}$ and $p_2 = 0.6$. Suppose we wished to consider an additive shift on the probability scale, e.g., $p_1' = p_1 + 0.1, p_2' = p_2 + 0.1$. Setting aside the inconvenience that we need to ensure $p_1', p_2' \in [0, 1]$, we argue that these shifts are not truly of a comparable scale. In particular, this shift in p_1 may seem implausible in magnitude, while the same shift in p_2 seems more reasonable. On the other hand, an additive shift in the log-odds captures some aspect of this idea.

Of course, there is some flexibility to incorporate prior expectations of shifts in absolute probabilities. For instance, in binary variable with no causal parents, we can always construct a one-to-one map of δ to a change in the marginal probability. For conditional shifts, we can similarly construct a one-to-one map between the value of δ in a shift $s(Z; \delta) = \delta$ and the resulting marginal probability of W_i, as formalized below.

Proposition D.2.1. *Consider a binary random variable W with conditional distribution*

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta)$$

for an arbitrary measurable function $\eta(Z)$ whose range is the extended real numbers $\eta(Z) \in \mathbb{R} \cup \{+\infty, -\infty\}$. Let $p_+ := \mathbb{P}(\eta(Z) = +\infty)$, $p_- := \mathbb{P}(\eta(Z) = -\infty)$, and assume *that $p_+ + p_- < 1$. Then, the marginal probability*

$$p_\delta = \mathbb{P}_\delta(W = 1)$$

is a strictly monotonically increasing function of $\delta \in \mathbb{R}$ whose range is $(p_+, 1 - p_-)$,

Proposition D.2.1 states that, for any achievable marginal probability $p_\delta = \mathbb{P}_\delta(W = 1)$, there exists a unique value of δ that achieves this probability. Because this relationship is strictly monotonic, we can hope to efficiently find such a value by e.g., binary search. In the laboratory testing example of Example 6.1, this would allow us to specify a plausible strength for the conditional shift δ in terms of an impact on the overall testing rate, e.g., modelling a scenario where the testing rate decreases from 20% to 15%.

Similar to the binary case, we can (if desired) directly parameterize shifts in terms of the conditional mean of a Gaussian distribution, as illustrated in Example D.2.2, which operates on the scale of $\mu(Z)$ alone.

Example D.2.2 (Mean shift in a conditional Gaussian)**.** Consider the distribution of a multi-variate Gaussian variable W conditioned on a binary variable Z, where we write

$$p(w|z) \overset{\text{(d)}}{=} \mathcal{N}(w; \mu(z), \Sigma(z))$$

where $\mathcal{N}(w; \mu(z), \Sigma(z))$ denotes the Gaussian density with mean $\mu(z)$ and covariance $\Sigma(z)$. This can be written as an exponential family model with natural parameters $\eta(Z) = [\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1}]$ and sufficient statistic $T(W) = [W, WW^\top]$. Here, a shift in the mean can be parameterized by $s(Z; \delta) = [\Sigma(Z)^{-1}\delta, 0]$, such that

$$p_\delta(w|z) \overset{\text{(d)}}{=} \mathcal{N}(w; \mu(z) + \delta, \Sigma(z)).$$

However, shifts of the same magnitude in the conditional mean may not be comparable.

Suppose that

$$\mathbb{P}(W|Z=0) \overset{(\mathrm{d})}{=} \mathcal{N}(0,1) \qquad \text{and} \qquad \mathbb{P}(W|Z=1) \overset{(\mathrm{d})}{=} \mathcal{N}(0,0.001),$$

such that $\delta = 1$ in Example D.2.2 corresponds to

$$\mathbb{P}_{\delta=1}(W|Z=0) \overset{(\mathrm{d})}{=} \mathcal{N}(1,1) \qquad \text{and} \qquad \mathbb{P}_{\delta=1}(W|Z=1) \overset{(\mathrm{d})}{=} \mathcal{N}(1,0.001).$$

While it may seem plausible that the mean of $W|Z=0$ can increase by 1, it may seem unrealistic for $W|Z=1$. Here, it may be more reasonable to consider a different parameterization of $s(Z;\delta)$, where the impact of the shift in a direction is proportional to the variance in that direction; we discuss this in the next example.

Example D.2.3 (Variance-scaled mean shift in a conditional Gaussian)**.** Consider the distribution of a multi-variate Gaussian variable W conditioned on variables Z, where we write

$$p(w|z) \overset{(\mathrm{d})}{=} \mathcal{N}(w;\mu(z),\Sigma(z))$$

where $\mathcal{N}(w;\mu(z),\Sigma(z))$ denotes the Gaussian density with mean $\mu(z)$ and covariance $\Sigma(z)$. This can be written as an exponential family model with natural parameters $\eta(Z) = [\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1}]$ and sufficient statistic $T(W) = [W, WW^\top]$. Here, a shift in the mean can be parameterized by $s(Z;\delta) = [\delta, 0]$, such that

$$p_\delta(w|z) \overset{(\mathrm{d})}{=} \mathcal{N}(w;\mu(z)+\delta^\top\Sigma(Z),\Sigma(z)).$$

In Example D.2.3, the parameter δ has a different interpretation, as a variance-scaled mean-shift. If W is one-dimensional, we can see that this becomes

$$p_\delta(w|z) \overset{(\mathrm{d})}{=} \mathcal{N}(w;\mu(z)+\delta\sigma^2(Z),\sigma^2(z)).$$

As we demonstrate in Appendix D.3.2, this particular example of a parameterization has other benefits: For instance, for estimation of shift gradients and Hessians at $\delta = 0$

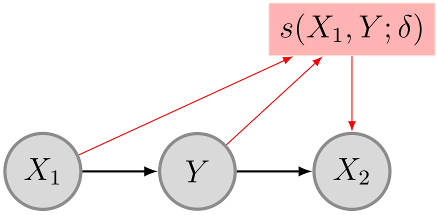


Figure D-1: *Illustrative example of an intervention $s(X_1, Y; \delta)$, and modified causal graph, which creates a dependence between X_1 and X_2 that bypasses Y.*

can be done without knowledge of $\Sigma(Z)$.

D.2.2 Adding causal edges to the graph

In Section 6.2, we consider the case where the shift function $s(Z; \delta)$ alters a conditional $\mathbb{P}(W|Z)$ by a shift function $s(Z; \delta)$. We now discuss shift functions that use a larger set Z'. In particular, we consider the setting where Z represents the parents in a graph $\mathcal{G}$ (that is, $Z \coloneqq \mathrm{PA}_{\mathcal{G}}(W)$), and consider shift functions that correspond to adding additional parents in that causal graph. Our definitions and results immediately extend to measuring the impact of shifts that **add edges** to the graph, in the form of shift functions that depend on non-descendants of W.

Building intuition with a simple example: To build intuition, consider the causal graph given in Figure D-1. We consider a shift in X_2, with a shift function which depends not only on the causal parent Y, but also on X_1. Suppose that the distribution $\mathbb{P}(X_2|Y)$ is a conditional exponential family, given by

$$\mathbb{P}(X_2|Y) = g(X_2)\exp(\eta(Y)^\top T(X_2) - h(\eta(Y))).$$

Using that $X_2 \perp\!\!\!\perp X_1 | Y$, we have $\mathbb{P}(X_2|Y) = \mathbb{P}(X_2|Y, X_2)$, and the joint probability factorizes as

$$\mathbb{P}(X_1, X_2, Y) = \mathbb{P}(X_2|Y)\mathbb{P}(Y|X_1)\mathbb{P}(X_1) = \mathbb{P}(X_2|Y, X_1)\mathbb{P}(Y|X_1)\mathbb{P}(X_1).$$

This enables us to consider $Z = (Y, X_1)$ as the conditioning set in the context of Assumption 6.1. This is useful, because it allows us to consider shift functions that depend on Z, which includes X_1 in addition to Y. The δ-perturbation of this conditional distribution under the shift function $s(Y, X_1; \delta)$ is given by

$$\mathbb{P}_\delta(X_2|Y, X_1) = g(X_2) \exp\left(\{\eta(Y) + s(Y, X_1; \delta)\}^\top T(X_2) - h\big(\eta(Y) + s(Y, X_1; \delta)\big) \right),$$

and we can observe that under both graphs, the distribution factorizes in the same fashion, where

$$\mathbb{P}_\delta(X_1, X_2, Y) = \mathbb{P}_\delta(X_2|Y, X_1)\mathbb{P}(Y|X_1)\mathbb{P}(X_1),$$

keeping the same convention that $s(Y, X_1; \delta = 0) = 0$, such that $\mathbb{P}_0 = \mathbb{P}$. This is one example of how our results can be applied with shift functions that effectively add edges to the causal graph. Of course, not all edges are permitted, so we give a more general treatment below.

General guidelines for adding edges: Allowing for the use of non-causal parents in the shift functions is straightforward, and can be done safely as follows, without violating Assumption 6.1: Given knowledge of the directed acyclic graph $\mathcal{G}$ which generates the observed distribution $\mathbb{P}$, we can **add** edges to the graph, as long as they do not create cycles.

Formally, let $\mathcal{G} = (\mathbf{V}, E)$ denote the causal DAG which generates the distribution $\mathbb{P}$, where $\mathbf{V}$ denotes variables and E denotes the set of edges, where we denote a directed edge by $e = (V_i, V_j)$, going from V_i to V_j. Let $\mathcal{G}' = (\mathbf{V}', E')$ denote another DAG (of our creation) with the constraint that we can only add edges, and that the graph must remain acyclic, such that $E' \supseteq E$, and $\mathbf{V}' = \mathbf{V}$.

For any variable $W_i \in \mathbf{V}$, this implies that $\text{PA}_{\mathcal{G}'}(W_i) \supseteq \text{PA}_{\mathcal{G}}(W_i)$. Moreover, any new causal parent V_i of W_i in $\mathcal{G}'$ must have been a non-descendant of W_i in the original graph, as otherwise the graph $\mathcal{G}'$ would have a cycle from $W_i \to V_i \to W_i$. For ease of notation, let $N(W_i) \coloneqq \text{PA}_{\mathcal{G}'}(W_i) \setminus \text{PA}_{\mathcal{G}}(W_i)$ denote the set of new causal parents of

W_i in $\mathcal{G}'$. For any variable W_i such that $N(W_i) \neq \varnothing$, we can write that

$$W_i \perp\!\!\!\perp_{\mathcal{G}} N(W_i) |\operatorname{PA}_{\mathcal{G}}(W_i) \tag{D.1}$$

by the rules of d-separation (Pearl, 2009). As in Assumption 6.1, we use $\mathbf{W} = \{W_1, \ldots, W_m\}$ to denote the set of variables to be intervened upon, and accordingly will assume that in the causal graph $\mathcal{G}'$, we have not added new parents to any other variables, i.e., $N(V_i) = \varnothing$ for any $V_i \subsetneq \mathbf{W}$.

By Equation (D.1), we can write that the distribution $\mathbb{P}$ factorizes as

$$\mathbb{P}(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}(W_i |\operatorname{PA}_{\mathcal{G}'}(W_i))\right) \prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i |\operatorname{PA}_{\mathcal{G}}(V_i))$$

because $\mathbb{P}(W_i |\operatorname{PA}_{\mathcal{G}'}(W_i)) = \mathbb{P}(W_i |\operatorname{PA}_{\mathcal{G}}(W_i)$, and if $\mathbb{P}(W_i |\operatorname{PA}_{\mathcal{G}}(W_i))$ is a conditional exponential family satisfying Definition 6.2, then $\mathbb{P}(W_i |\operatorname{PA}_{\mathcal{G}}(W_i))$ also satisfies this definition, where the function $\eta(\operatorname{PA}_{\mathcal{G}}(W_i), N(W_i))$ is constant with respect to fluctuation in the variables $N(W_i)$. Thus, taking $Z_i := \operatorname{PA}_{\mathcal{G}'}(W_i)$ as the conditioning set satisfies Assumption 6.1, and the rest of our results hold, where the corresponding δ-perturbations in Definition 6.4 are given by

$$\mathbb{P}_{\delta}(\mathbf{V}) = (\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i |\operatorname{PA}_{\mathcal{G}'}(W_i))) \prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i |\operatorname{PA}_{\mathcal{G}}(V_i))$$

with shift function $s_i(\operatorname{PA}_{\mathcal{G}'}(W_i); \delta_i)$ that are parametric functions of causal parents in the modified graph $\mathcal{G}'$.

D.2.3 Domain-preserving parameterizations of shift

For both of the examples considered above, we did not need to restrict the magnitude of the additive change to $\eta(Z)$. However, in some cases, such as changing the variance of a conditional Gaussian, we have the restriction that $\eta_{\delta}(Z) = \eta(Z) + s(Z; \delta)$ must lie in the proper domain, e.g., we cannot consider a shift which causes the conditional variance to become negative. For a conditional Gaussian, we can consider unrestricted shifts in $\eta(Z)_1$, which controls the mean, because the mean has unrestricted domain. On the other hand, $\eta(Z)_2 = (-2\sigma^2(Z))^{-1}$ controls the variance, and must remain negative, such that $\eta(Z)_2 + s(Z;\delta)_2 < 0$ for the shifts we consider.

This can be resolved in one of two ways. First, one can consider parameterizations of $s(Z;\delta)$ which are guaranteed to preserve the correct domain with an additional constraint on the values of δ, such as the multiplicative shift below, which is sign-preserving for $\delta > -1$

$$\eta_\delta(Z)_2 = \eta(Z)_2 + \underbrace{\delta\eta(Z)_2}_{s(Z;\delta)} = (1+\delta)\eta(Z)_2.$$

To handle the general case, at the expense of some additional complexity in the gradients of $s(Z;\delta)$, one can define the shifts as follows for parameters $\eta(Z)$ that have a lower bound L, with an equivalent formulation for shifts where the parameters have an upper bound, for any desired shift function $s'(Z;\delta)$

$$\eta(Z) + \underbrace{s'(Z;\delta) \cdot \text{sigmoid}(\gamma \cdot [(\eta(Z) + s'(Z;\delta)) - (L+\epsilon)])}_{s(Z;\delta)}$$

where $\text{sigmoid}(\gamma \cdot (x - (L+\epsilon)))$ is a smooth relaxation of the indicator function $\mathbf{1}\{x > L+\epsilon\}$, for a sufficiently large temperature parameter $\gamma > 0$ and a small $\epsilon > 0$. This transformation preserves the twice-differentiable nature of $s(Z;\delta)$. In practice, however, we typically evaluate the gradient of $s(Z;\delta)$ at $\delta = 0$, where $\eta(Z)$ does not lie at the boundary of allowable parameter space, such that we can consider simpler parameterizations like

$$\eta(Z) + \underbrace{s'(Z;\delta) \cdot \mathbf{1}\{\eta(Z) + s'(Z;\delta) > L+\epsilon\}}_{s(Z;\delta)}$$

as long as ϵ is taken sufficient small such that $\eta(Z) > L+\epsilon$ almost everywhere in $\mathbb{P}$.

D.3 Considerations and additional results for evaluation of the worst-case loss

In this section, we present additional results on the Taylor approximation and compare how the Taylor approximation compares to the reweighting approach in evaluation and worst-case optimization of the shifted loss.

- In Appendix D.3.1 we give a full treatment of how shift gradients and Hessians are estimated from samples, following Theorem 6.1.

- In Appendix D.3.2, we demonstrate in some cases, one does not need to estimate all of $\eta(Z)$, but only the parts of $\eta(Z)$ that is shifting.

- In Appendix D.3.3, we demonstrate that the second-order Taylor expansion is exact in a linear-Gaussian setting, which gives a conceptual connection between this work and that of Anchor Regression (Rothenhäusler et al., 2021), which considered a restricted type of additive shift intervention in a globally linear structural causal model.

- In Appendix D.3.4, we work out the expression for the shift gradient and Hessian when we condition on binary variables.

- In Appendices D.3.5 to D.3.7, we provide experiments that compare the variance of the importance sampling estimate $\hat{E}_{\delta,\text{IS}}$ (see Equation (6.6)) to the variance of the Taylor estimate $\hat{E}_{\delta,\text{Taylor}}$ (see Equation (6.7)) of the loss in a shifted distribution.

- In Appendix D.3.8, we consider the bound in Theorem 6.2 in a covariate shift setting, and give an explicit expression for this under additional assumptions.

D.3.1 Algorithm for Estimation of Shift Gradients and Hessians

Here, we recall the form of the shift gradients and Hessians in Theorem 6.1, and demonstrate how to compute them in practice using a set of auxiliary regression functions fit to the validation data.

Theorem 6.1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 6.4, with intervened variables $\mathbf{W} = \{W_1, \dots, W_m\}$ and shift functions $s_i(Z_i; \delta_i)$, where $\delta = (\delta_1, \dots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \dots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$\mathrm{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top Cov\Big(\ell,\, T_i(W_i)\Big| Z_i\Big)\right],$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$, where the (i,j)th block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$\{\mathrm{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top Cov\left(\ell,\, \epsilon_{T_i|Z_i}\epsilon_{T_i|Z_i}^\top | Z_i\right) D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \epsilon_{T|Z}\right] & i = j \\ Cov(\ell,\; D_{i,1}^\top \epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i; \delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here, $T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Notation and Dimensions: Let $\mathbf{W} = \{W_1, \dots, W_m\}$ denote the set of m intervened variables, and let $\mathbf{Z} = \{Z_1, \dots, Z_m\}$ denote the conditioning sets. Note that for a single $W_i \in \mathbb{R}^{d_{W_i}}$, we will generally have it that $Z_i \in \mathbb{R}^{d_Z}$, where d_W is the dimension of W (typically 1) and d_Z is the number of conditioning variables, and when considering n samples, W_i will be a matrix in $\mathbb{R}^{n \times d_W}$, and Z_i will be a matrix $\mathbb{R}^{n \times d_Z}$. The sufficient statistic $T_i(W_i)$ maps from $\mathbb{R}^{d_W}$ to $\mathbb{R}^{d_T}$, where d_T is the dimension of the sufficient statistic. For many common distributions, $T_i(W_i) = W_i$, the identity function. For others, like the conditional multi-variate Gaussian, $T_i(W_i) = [W_i, W_i W_i^\top]$, where $W \in \mathbb{R}^{d_W}$ and $W_i W_i^\top \in \mathbb{R}^{d_W \times d_W}$. In these cases, we squeeze $T_i(W_i)$ to be a single vector, so in this case $d_T = d_W + d_W^2$.

Auxiliary models: To estimate the shift gradients and Hessians, we first learn auxiliary predictive models, which are required for computing the relevant conditional covariances. For simplicity, we do not consider sample-splitting in the algorithm given below, but one could employ sample-splitting to learn these predictive models on an independent validation sample.

- For each W_i, we learn $\hat{\mu}_{W_i}(Z_i)$ as a regression model for $\mathbb{E}[T_i(W_i)|Z_i]$. Because $T_i(W_i)$ may have multiple dimensions, this is a function from $\mathbb{R}^{d_Z}$ to $\mathbb{R}^{d_T}$.
- For each conditioning set Z_i, we learn $\hat{\mu}_\ell(Z_i)$ as a regression model for $\mathbb{E}[\ell|Z_i]$. Because the loss is one-dimensional, this is a function from $\mathbb{R}^{d_Z}$ to $\mathbb{R}$.

We then construct the following, which are defined for each data point in the sample.

- For each W_i, we construct $\hat{\epsilon}_{T_i|Z_i} := T_i(W_i) - \hat{\mu}_{W_i}(Z_i)$, which is a vector of length d_{T_i}.
- For each conditioning set Z_i, for the loss ℓ, we construct $\hat{\epsilon}_{\ell|Z_i} := \ell - \hat{\mu}_\ell(Z_i)$, which is a real number.
- For each conditioning set Z_i, we compute $D_{i,1}(Z_i)$ as $\nabla_{\delta_i} s_i(Z_i;\delta_i)\big|_{\delta=0}$, which is a matrix of size $d_T \times d_{\delta_i}$, and a function of Z_i that we can evaluate on each sample.
- For each conditioning set Z_i, we compute $D_{i,2}(Z_i)$ as $\nabla^2_{\delta_i} s_i(Z_i;\delta_i)\big|_{\delta=0}$, which is a tensor of size $d_T \times d_{\delta_i} \times d_{\delta_i}$, and a function of Z_i that we can evaluate on each sample.

Estimating shift gradients The shift gradient and Hessian in Theorem 6.1 are expressed as conditional covariance. Since $\mathbb{E}[\text{Cov}(A,B|C)] = \mathbb{E}[\epsilon_{A|C}\epsilon_{B|C}]$ where $\epsilon_{A|C} := A - \mathbb{E}[A|C]$ and $\epsilon_{B|C} := B - \mathbb{E}[B|C]$, we can use the estimated conditional means above, to compute the shift gradient and Hessian. Suppose that we observe N samples, $n \in \{1, \ldots, N\}$. For each index $i \in [m] := \{1, \ldots, m\}$,

$$\hat{\mathrm{SG}}^{\,1}_{i} = \frac{1}{N} \sum_{n=1}^{N} \hat{\epsilon}^{(n)}_{\ell|Z_i} \cdot D_{i,1}(Z_i^{(n)})^{\top} \hat{\epsilon}^{(n)}_{T_i|Z_i}$$

which yields a vector of length d_{δ_i}, and these are concatenated together for each i to yield the entire shift gradient. The shift Hessian is constructed block-wise, for each index $i, j \in [m] \times [m]$ as follows: If $i = j$, then we construct the corresponding $d_{\delta_i} \times d_{\delta_i}$ block as

$$\hat{\mathrm{SG}}^{\,2}_{i,i} = \frac{1}{N} \sum_{n=1}^{N} \hat{\epsilon}^{(n)}_{\ell|Z_i} \cdot \left[\left(D_{i,1}(Z_i^{(n)})^{\top} \hat{\epsilon}^{(n)}_{T_i|Z_i} \right)^{\otimes 2} - D_{i,2}(Z_i^{(n)})^{\top} \hat{\epsilon}_{T_i|Z_i} \right]$$

where $v^{\otimes 2}$ denotes the outer product so that $v^{\otimes 2} = vv^{\top}$, and the transpose of $D_{i,2}$ refers to a transpose which has dimension $d_{\delta_i} \times d_{\delta_i} \times d_T$. On the other hand, if $i \neq j$ we have

$$\hat{\mathrm{SG}}^{\,2}_{i,j} = \frac{1}{N} \sum_{n=1}^{N} (\ell^{(n)} - \bar{\ell}) \cdot \left(D_{i,1}(Z_i^{(n)})^{\top} \hat{\epsilon}^{(n)}_{T_i|Z_i} \right) \left(D_{j,1}(Z_j^{(n)})^{\top} \hat{\epsilon}^{(n)}_{T_j|Z_j} \right)^{\top}$$

where $\bar{\ell}$ is the average value of ℓ in the validation sample.

D.3.2 Shifts where estimating all of $\eta(Z)$ is not necessary for estimating shift gradient and Hessian

The following example shows that when a shift occurs in an exponential conditional distribution with parameter $\eta(Z)$, we do not necessarily need to model all of $\eta(Z)$ in order to compute the shift gradient and Hessian. In particular, we only need to model the parts of $\eta(Z)$ that shift. This is different from estimating the shifted loss using importance sampling, where $\eta(Z)$ needs to be evaluated to evaluate Equation (6.5).

Example D.3.1. Consider the distribution of W conditioned on variables Z that is a multi-variate Gaussian variable,

$$W|Z = \mathcal{N}(\mu(Z), \Sigma(Z)),$$

for unknown functions μ, Σ. The sufficient statistic for the multivariate Gaussian distribution is $T(W) = (W, WW^\top)$ and the canonical parameter is $\eta(Z) = (\Sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\Sigma(Z)^{-1})$.[1] The first component of $\eta(Z)$ is a signal-to-variance ratio and the second is the inverse covariance matrix. For a shift $(\delta, 0)$ that only affects the first component, we show that we do not need to model $\Sigma(Z)$, but only $\mu(Z)$. This is beneficial, since estimating a conditional covariance from data can be challenging, especially if W is high-dimensional.

For $\delta \in \mathbb{R}^{d_W}$, let $s(Z; \delta) = (\delta, 0)^\top$, and suppose that we wish to estimate $\mathbb{E}_\delta[\ell]$ using Equation (6.7). The derivative of s is given by

$$D_1 = \nabla_\delta^2 s(Z; \delta) = \left(\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right),$$

where the first block is a $d_W \times d_W$ diagonal matrix, and the second is a $d_W \times d_W^2$ matrix of zeros. The second derivative of s is $D_2 = 0$. Hence, using Theorem 6.1, the shift gradient is

$$\mathrm{SG}^1 = \mathbb{E}[D_1 \mathrm{Cov}(\ell, (W, WW^\top)|Z)] = \mathbb{E}[\mathrm{Cov}(\ell, W|Z)],$$

and

$$\mathrm{SG}^2 = \mathbb{E}\left[D_1 \mathrm{Cov}(\ell, \left(W - \mathbb{E}[W|Z], WW^\top - \mathbb{E}[WW^\top|Z]^{\otimes 2} \right) |Z) D_1^\top \right]$$

[1] Or, more formally, $T(W) = \big(W, \mathrm{vec}(WW^\top)\big)$ and $\eta(Z) = \big(\sigma(Z)^{-1}\mu(Z), -\frac{1}{2}\,\mathrm{vec}(\mu(Z))\big)$, where vec denotes the vectorization operation. For a detailed walk through of the exponential family parameterization of multivariate Gaussian distributions, see https://maurocamaraescudero.netlify.app/post/multivariate-normal-as-an-exponential-family-distribution/.

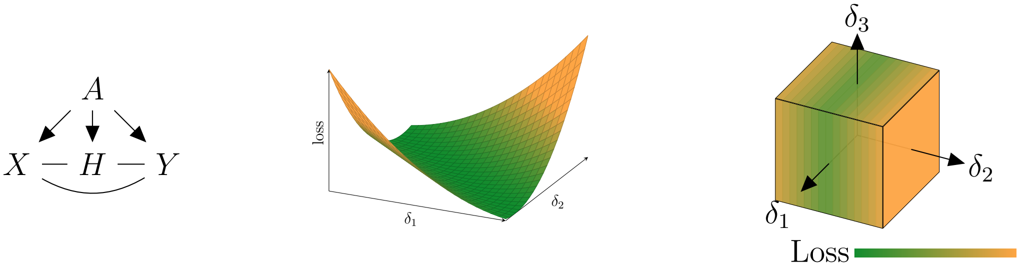


Figure D-2: *(Left) Graphical model assumed by Equation* (D.2)*. The undirected edges represent either any directed configuration of directed edges or the dependence structures arising due to an acyclic SCM (Bongers et al., 2021). (Middle) Plotting* $\mathbb{E}_\delta[(Y-\gamma^\top X)^2]$ *as a function of* $\delta \in \mathbb{R}^2$ *for a fixed predictor* γ*. (Right) Plotting* $\mathbb{E}_\delta[(Y-\gamma^\top X)^2]$ *as a function of* $\delta \in \mathbb{R}^3$*, with the loss indicated by the color. The loss only varies with changes in* δ_2 *(corresponding in Lemma D.3.1 to* $v_\gamma \propto (0,1,0)^\top$*).*

$$= \mathbb{E}\left[\mathrm{Cov}(\ell, \left(W - \mathbb{E}[W|Z]\right)^{\otimes 2}|Z)\right].$$

Conditional covariances can be computed by only residualizing one of the variables: $\mathbb{E}[\mathrm{Cov}(A,B|C)] = \mathbb{E}[A(B-\mathbb{E}[B|C])]$. Thus, if we only residualize ℓ, we get

$$\mathrm{SG}^1 = \mathbb{E}[(\ell - \mathbb{E}[\ell|Z])W] \qquad \text{and} \qquad \mathrm{SG}^2 = \mathbb{E}[(\ell - \mathbb{E}[\ell|Z]) \cdot (W-\mu(Z))^{\otimes 2}].$$

Therefore, given data from $\mathbb{P}$, we can estimate the shift gradients by plugging in estimators $\hat{\mu}(Z)$ of $\mathbb{E}[W|Z]$ and $\hat{L}(Z)$ of $\mathbb{E}[\ell|Z]$. It follows that we do not need to model $\Sigma(Z)$ in order to estimate the shift gradients and Hessian at $\delta = 0$.

The story is different for a reweighting based estimator that seeks to estimate $\mathbb{E}_\delta[\ell]$ using importance sampling (see Section 6.3.1), where the weights are given by

$$w_{\eta,\delta}(Z) = (W-\mu(Z))^\top\delta - \tfrac{1}{2}\delta^\top\Sigma(Z)\delta,$$

and hence estimating $w_{\eta,\delta}(Z)$ requires estimation of $\Sigma(Z)$.

D.3.3 The quadratic approximation is exact, for mean shifts in linear models

We now consider data generated by a linear model, and show that the shifted loss is a quadratic function of δ, meaning that the Taylor approximation $E_{\delta,\mathrm{Taylor}}$ is globally exact. Suppose that data is sampled from a linear structural causal model, and a shift in mean occurs in an variable A that does not have any causal parents. In particular, let A have a normal distribution with mean μ and finite variance and let

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + MA + \epsilon. \tag{D.2}$$

This is the model assumed by Rothenhäusler et al. (2021), and the corresponding graphical model is shown in Figure D-2 (left). We consider the linear predictor $f_\gamma(X) = \gamma^\top X$ and the mean squared loss $\ell(f_\gamma(X), Y) = (Y - f(X))^2$. Due to the linearity of the model, the loss under a mean shift in A is quadratic (Rothenhäusler et al., 2021).

Lemma D.3.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (D.2). For $\gamma \in \mathbb{R}^{d_X}$ define $\ell := (Y - \gamma^\top X)^2$. Then there exist $v_\gamma, u_{\mu,\gamma} \in \mathbb{R}^{d_A}$ such that for all shifts $\delta \in \mathbb{R}^{d_A}$:*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top u_{\mu,\gamma} + \tfrac{1}{2}\delta^\top v_\gamma v_\gamma^\top \delta,$$

where $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$. Further $u_{\mu,\gamma} = 0$ if $\mu = 0$.

Proposition D.3.1 elicits two properties of this linear model: First the loss is described by a quadratic function globally, i.e. also for very large δ. In Figure D-2 (middle), we plot $\mathbb{E}_\delta[\ell]$ as a function of δ. We observe a 'valley' in the loss, in which the expected loss does not at all change with δ. This is a consequence of Lemma D.3.1, and particularly that if δ is orthogonal to both $u_{\mu,\gamma}$ and v_γ then $\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell]$. In higher dimensions $d_A > 2$, since $v_\gamma v_\gamma^\top$ has rank 1, the 'valley' persists in that the loss does not grow at all in $d_A - 2$ dimensions (or $d_A - 1$ if A has mean $\mu = 0$), see Figure D-2 (right).

We now show that coefficients in the quadratic form in Lemma D.3.1 is equal to the shift gradient and Hessian. We use that the Gaussian distribution with known variance Σ can be parameterized as an exponential family with sufficient statistic $T(A) = \Sigma^{-1}A$ and parameter $\eta = \mu$.[2]

Proposition D.3.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (D.2). Then the shift gradient and Hessian are given by*

$$\mathrm{SG}^1 = \mathit{Cov}(\ell, \Sigma^{-1}A) \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathit{Cov}(\ell, \Sigma^{-1}(A-\mu)(A-\mu)^\top\Sigma^{-\top})$$

and the loss under a mean shift of δ in A is given by

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2\,\delta,$$

where $\ell := (Y - \gamma^\top X)^2$ and $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$.

This elicits a connection to anchor regression (Rothenhäusler et al., 2021): Under the generative model Equation (D.2) and using the quadratic loss $\ell = (Y - \gamma^\top X)^2$ for $\gamma \in \mathbb{R}^{d_X}$, they show that for any $\lambda \geq 0$, the worst-case loss $\mathbb{E}_\delta[\ell]$ over a set $\Delta = \{\delta | \delta\delta^\top \preceq \lambda\mathbb{E}[AA^\top]\}$ equals the objective $\ell_{\mathrm{AR}} = \mathbb{E}[\ell] + \lambda\mathbb{E}[\mathbb{E}[Y - \gamma^\top X|A]^2]$, which is computable from the observed distribution.

Because of Proposition D.3.1, ℓ_{AR} *also* equals the solution of the optimization problem Equation (6.9) over the constraint set Δ. Therefore minimizing the anchor regression objective over γ or minimizing Equation (6.9) over γ will lead to the same estimator. Since our proposed Taylor approximation in Equation (6.9) does not assume linearity, one could use the approximation to extend the rationale of anchor regression of

[2] It can also be parameterized as $T(A) = \Sigma^{-1/2}A, \eta = \Sigma^{-1/2}\mu$, which would yield the same result.

minimizing the worst-case loss to non-linear models. This however comes at the cost of not optimizing the exact worst-case loss, but rather an approximation, whose quality is given by Theorem 6.2. Further, this would involving a minimax problem, minimizing Equation (6.9) over models f, and there are questions, such as convexity and tractability, which would need to be solved.

D.3.4 Estimating the shift gradient and Hessian for conditional on binary variables

To build intuition for the shift gradient and Hessian, we here give an example where we condition on variables Z that take a finite number of values and write out explicit expressions for the shift gradient and Hessian. However, we emphasize, that in most practical scenarios, one will not have to work out the shift gradient and Hessian explicitly, but can simply estimate them as covariances from the data (Theorem 6.1).

Example D.3.2 (Shift Function of Discrete Parents)**.** Consider a conditional distribution $W|Z$ where Z takes values in a finite set $\mathcal{Z}$. This is for instance the case if $Z = (Z_1, \ldots, Z_d)$ where each Z_i is binary, so $|\mathcal{Z}| = 2^d$. Instead of a shift $\eta(Z) + \delta$, where the parameter increases by the same amount for all values of Z, we may consider a shift $\eta(Z) + s(Z; \delta)$ where $s(Z; \delta) = \sum_{z \in \mathcal{Z}} \delta_z 1_{Z=z}$, meaning that the shift is different in each category Z. Since $\eta(Z)$ only takes a finite number of variables, this shift corresponds to an arbitrary change in $\eta(Z)$.

$s(Z; \delta)$ is a differentiable function in δ, and if $d_T = 1$ the shift gradient is a (1×2^d)-row vector, $\nabla_\delta s(Z; \delta) = (1_{Z=z})_{z \in \mathcal{Z}}$, and the shift Hessian vanishes, $\nabla^2_\delta s(Z; \delta) = 0$. Enumerating $\mathcal{Z} = \{1, \ldots, 2^d\}$, the i'th entry in the shift gradient becomes

$$(\mathrm{SG}^1)_i = \mathbb{E}\left[1_{Z=i}\mathrm{Cov}\left(\ell, T(W)\middle| Z\right)\right] = \mathbb{P}(Z = i)\mathrm{Cov}(\ell, T(W)|Z = i),$$

and the i, j'th entry of the shift Hessian becomes 0 if $j \neq i$ and else

$$(\mathrm{SG}^2)_{i,i} = \mathbb{E}\left[1_{Z=i}\mathrm{Cov}_\delta\left(\ell, \epsilon^{\otimes 2}_{T|Z}\middle| Z\right)\right] = \mathbb{P}(Z = i)\mathrm{Cov}(\ell, \epsilon^{\otimes 2}_{T|Z}|Z = i).$$

Consider for example the case where both W and Z are binary. Then $T(W) = W$ and $s(Z;\delta) = 1_{Z=0}\delta_0 + 1_{Z=1}\delta_1$ and $s^{(1)} = (1_{Z=0}, 1_{Z=1})$ and $s^{(2)} = 0$. The conditional covariance can be evaluated by residualizing only one of the variables, $\mathbb{E}[\text{Cov}(A,B|C)] = \mathbb{E}[A(B - \mathbb{E}[B|C])]$, so we can chose to residualize only W (for SG^1) or $(W - \mathbb{E}[W|Z=i])^2$ (for SG^2). Finally, if we let $p_i = \mathbb{P}(W = 1|Z = i)$ and use that $\mathbb{E}[W|Z = i] = p_i$ and $\mathbb{E}[(W - p_i)^2|Z = i] = \text{Var}(W|Z = i) = p_i(1 - p_i)$, we get that

$$\text{SG}^1 = \mathbb{E}\left[\begin{pmatrix} p_0 \cdot \ell \cdot (W - p_0) \\ p_1 \cdot \ell \cdot (W - p_1) \end{pmatrix}\right],$$

and

$$\text{SG}^2 = \mathbb{E}\left[\begin{pmatrix} \ell p_0\big\{(W - p_0)^2 - p_0(1 - p_0)\big\} & 0 \\ 0 & \ell p_1\big\{(W - p_1)^2 - p_1(1 - p_1)\big\} \end{pmatrix}\right].$$

D.3.5 Comparison of variance of reweighting and Taylor estimates in the lab ordering example

To compare the bias and variance of the Taylor and the importance sampling estimates of the shifted loss, we simulate data from the following, artificial, generative model (which is the same generative model that was used to construct the loss landscape in Figure 6-1b).

$$\begin{aligned} \text{Age} &\sim \mathcal{N}(0, 0.5^2) \\ \mathbb{P}(\text{Disease} = 1|\text{Age}) &= \text{sigmoid}(0.5 \cdot \text{Age} - 1) \\ \mathbb{P}(\text{Order} = 1|\text{Disease, Age}) &= \text{sigmoid}(2 \cdot \text{Disease} + 0.5 \cdot \text{Age} - 1) \\ \text{Test Result}|\text{Order} = 1, \text{Disease} &\sim \mathcal{N}(-0.5 + \text{Disease}, 1) \end{aligned}$$

where if Order $= 0$, the test result is a placeholder value of zero.

We consider either a shift in the logits of ordering lab tests $\eta_\delta(Z) = \eta(Z) + \delta$ (Figure D-

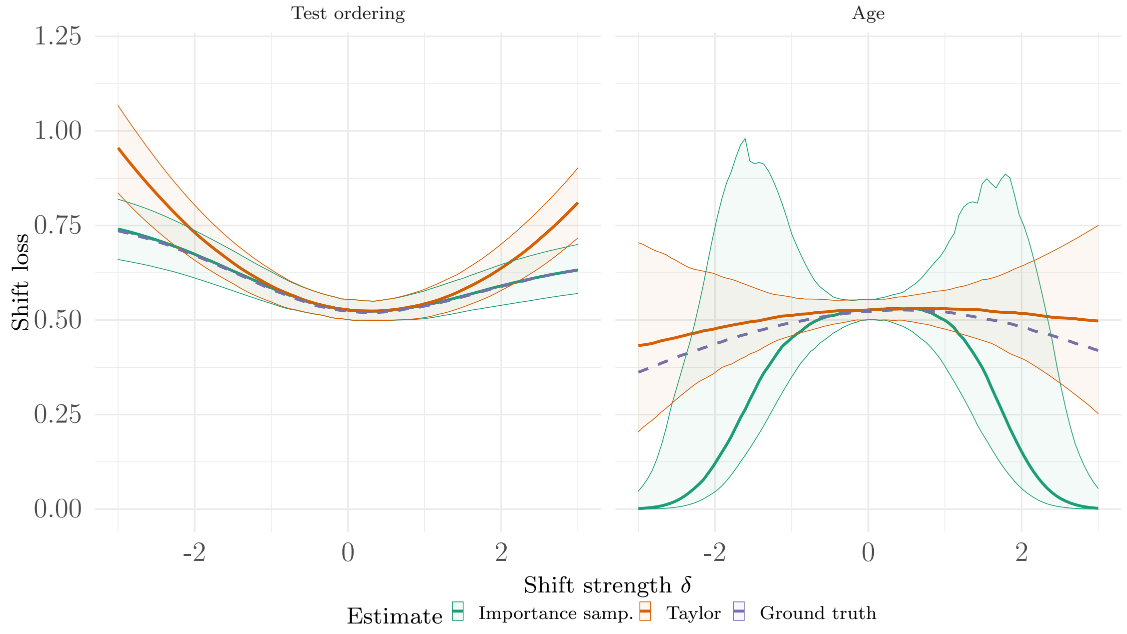


Figure D-3: *We plot the mean and confidence intervals of $\hat{E}_{\delta,Taylor}$ and $\hat{E}_{\delta,IS}$ when the shifted loss as in the lab test ordering example Example 6.1. (Left) We consider a shift in the logits of ordering lab tests from $\eta(Z)$ to $\eta(Z)+\delta_0$. (Right) We consider a shift in the mean of Age. In the observed distribution $\eta = \mu/\sigma = 0$ and we shift to a mean of $\eta = \delta$.*

3 left) or a mean shift in the Gaussian distribution of age $\eta_\delta = \delta$ (Figure D-3 right). For each δ in a grid, we compute estimates $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ of the loss under a shift of size δ, We repeat this $n = 1{,}000$ times, and plot the mean and point-wise prediction intervals (the pointwise 0.05 and 0.95 quantiles) for $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$. We also simulate ground truth data from $\mathbb{P}_\delta$, to compute the actual loss under shift.

For shifts in the binary variable (Figure D-3, left), both estimates capture the loss well for small shifts, but as δ gets larger, the quadratic approximation increasingly deviates from the true mean; the importance sampling estimate remains very close to the ground truth shifted loss. On the contrary, for the Gaussian mean shift (Figure D-3, right), the importance sampling weights are ill-behaved, and the variance dramatically increases as δ becomes larger. This supports the intuition, that while importance sampling tends to work well for binary variables, the variance can be large in continuous distributions, such as the Gaussian distribution.

D.3.6 Comparison of theoretical variance of reweighting and Taylor estimates

Example D.3.3. To demonstrate the reduction in variance obtained from using the Taylor approximation of the importance weights, we consider a simple example where $\mathbb{P}(X) \sim \mathcal{N}(0,1)$ and $\mathbb{P}_\delta(X) \sim \mathcal{N}(\delta,1)$ and we wish to estimate $\mathbb{E}_\delta[\ell(X)]$ for some loss function $\ell(X)$.[3] The importance sampling weights are given by $w_\delta(X) = \exp(-\frac{1}{2}\delta^2 + X\cdot\delta)$, and the shift gradient and Hessians are $\mathrm{SG}^1 = \mathbb{E}[\ell(X)X]$ and $\mathrm{SG}^2 = \mathbb{E}[\ell(X)X^2]$.

Therefore samples $X_1,\dots,X_n$ from $\mathbb{P}$ consider the estimators, for any loss function $\ell(X)$, two estimators of $\mathbb{E}_\delta[\ell]$ are

$$\hat{\mu}_{\text{IS}} = \frac{1}{n}\sum_{i=1}^n w_\delta(X_i)\ell(X_i) \quad \text{and} \quad \hat{\mu}_{\text{Taylor}} = \frac{1}{n}\sum_{i=1}^n \ell(X_i) + \delta\cdot\ell(X_i)X_i + \tfrac{1}{2}\delta^2\ell(X_i)X_i^2,$$

and the variances of the estimators are

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{\text{IS}}) &= \frac{\mathbb{E}[\{\ell(X+2\delta)\}^2]}{n}\exp(\delta^2)\\ \mathrm{Var}(\hat{\mu}_{\text{Taylor}}) &= \frac{\mathrm{Var}\left(\ell(X)+\delta X\ell(X)+\frac{1}{2}\delta^2X^2\ell(X)\right)}{n}.\end{aligned}$$

The variance of $\hat{\mu}_{\text{Taylor}}$ grows like δ^4 and the variance of $\hat{\mu}_{\text{IS}}$ grows exponentially fast (unless $\mathbb{E}[\{\ell(X+2\delta)\}^2]$ also diminishes exponentially fast, which is generally not the case), and so except for small δ, the variance of the importance sampling estimator will be orders of magnitude larger than the variance of the estimator using the Taylor approximation. While, $\hat{\mu}_{\text{IS}}$ is an unbiased estimator of $\mathbb{E}_\delta[\ell(X)]$ and $\hat{\mu}_{\text{Taylor}}$ is a biased, the overall mean squared error will be smaller for the Taylor approximation, unless the bias of the Taylor approximation also grows exponentially.

For the sake of analysis, consider the simple example $\ell(X) = X$. In this case, the Taylor estimate is unbiased because $\mathbb{E}_\delta[X] = \delta$ is a linear function of δ, so the quadratic

[3] In practice one would not use importance sampling estimation for such a simple shift, but use other approaches, such as analytically work out an estimate of $\mathbb{E}_\delta[\ell]$.

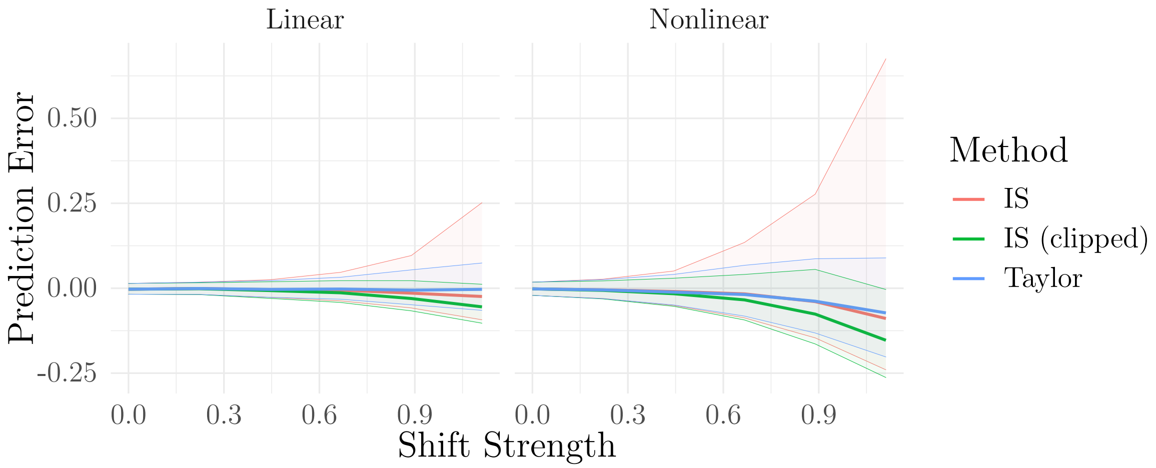


Figure D-4: *Median and quantiles of the error in predicting $\mathbb{E}_\delta[\ell]$ under a shift δ.*

approximation is adequate. Further, the variances are given by

$$\operatorname{Var}(\hat{\mu}_{\mathrm{IS}}) = \frac{\exp(\delta^2)(1+4\delta^2)-\delta^2}{n} \quad \text{and} \quad \operatorname{Var}(\hat{\mu}_{\mathrm{Taylor}}) = \frac{1+5\delta^2+\frac{15}{4}\delta^4}{n}.$$

In particular, the variance of the importance sampling estimate grows like $\exp(\delta^2)$ while that of the Taylor estimate grows like δ^4.

D.3.7 Comparison of variance of reweighting and Taylor estimates in a simple synthetic example

In this experiment, we compare the variance of importance sampling and Taylor estimates in a simple synthetic example. We simulate data from $\mathbb{P}$ where $X \in \mathbb{R}^3$ and $Y \in \mathbb{R}^1$ depend either linearly or quadratically on $W \in \mathbb{R}^3$,

$$W \sim \mathcal{N}(0, \mathrm{Id}_3) \qquad \text{and} \qquad \begin{pmatrix} X \\ Y \end{pmatrix} = (\mathrm{Id}_4 - B)^{-1} M(W + \alpha(W \odot W) + \epsilon),$$

where $\odot$ refers to entrywise multiplication, $\epsilon \sim \mathcal{N}(0, \mathrm{Id}_4)$, α is either 0 (linear) or $\frac{1}{2}$ (nonlinear) and

$$B := \begin{pmatrix} 2 & 1 & 0 & 1 \\ 2 & 2 & 0 & 3 \\ 3 & 3 & 0 & 2 \\ 4 & 2 & 4 & 0 \end{pmatrix} \qquad \text{and} \qquad M := \begin{pmatrix} 2 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 2 & 0 \\ 4 & 1 & 1 \end{pmatrix}.$$

On the simulated data from $\mathbb{P}$, we then fit a linear predictor $f(X)$ of Y, and consider a shift in the mean of W from $\mathbb{P}(W) \sim \mathcal{N}(0, \mathrm{Id}_3)$ to $\mathbb{P}_\delta(W) \sim \mathcal{N}(\delta, \mathrm{Id}_3)$, where $\delta = [s, s, s]^\top$ for some shift strength $s > 0$. We then compute the shift gradient $\mathrm{SG}^1 = \mathrm{Cov}(\ell, W)$ and Hessian $\mathrm{SG}^2 = \mathrm{Cov}(\ell, WW^\top)$, and approximate $\mathbb{E}_\delta[\ell]$ by $\hat{E}_{\delta,\mathrm{Taylor}}$ (see Equation (6.7)). In the linear data, the Taylor approximation is exact (see Appendix D.3.3), such that any prediction error can be attributed to finite-sample fluctuation, whereas both model misspecification and finite-sample fluctuation contribute to the error in the nonlinear setting.

Similarly, we estimate $\mathbb{E}_\delta[\ell]$ by importance sampling, $\mathbb{E}_\delta[\ell] = \mathbb{E}[w_\delta(W)\ell] \approx \frac{1}{n}\sum w_\delta(W)\ell$, where $w_\delta(W) = \frac{\mathbb{P}_\delta(W)}{\mathbb{P}(W)} = \delta^\top W - \frac{1}{2}\delta^\top\delta$, and compare this to ground truth data sampled from $\mathbb{P}_\delta$; we do the same for an importance sampling estimator with weights 'clipped' at the 99% quantile.

We compare the predicted loss $\mathbb{E}_\delta[\ell]$ by actually simulating data from $\mathbb{P}_\delta$ and evaluating $\mathbb{E}_\delta[\ell]$ (where ℓ is still the model trained on data from $\mathbb{P}$). We then compute the prediction error, as the difference $\mathbb{E}_\delta[\ell] - \hat{E}_{\delta,\mathrm{Taylor}}$ or $\mathbb{E}_\delta[\ell] - \hat{E}_{\delta,\mathrm{IS}}$.

For a number of different shift strengths s, we repeat this procedure $M = 1{,}000$ times, and in Figure D-4 we plot the median and a confidence interval defined by the 2.5 and the 97.5% quantiles of the prediction error.

In the linear case, both the importance sampling and the Taylor approximation retains a median error close to 0, with the variance of $\hat{E}_{\delta,\mathrm{IS}}$ being larger than $\hat{E}_{\delta,\mathrm{Taylor}}$. The clipped importance sampling estimate has a smaller variance than that of ordinary importance sampling, though the median deviates further from 0, and the variance is not smaller than that of the Taylor estimate.

In the non-linear cases, all three models underestimate the shifted loss. For $\hat{E}_{\delta,\text{Taylor}}$, this happens because as the mean of W shift, the mean shift is amplified by the non-linearity, such that the quadratic approximation of the loss is an underestimate. While the variance of the clipped importance sampling is smaller than the variance of the ordinary importance sampling estimate and comparable to the variance of the Taylor estimate, this prediction is further from 0 than the Taylor estimate.

Since importance sampling methods are known to produce very large outliers, the use of the median and quantiles, as opposed to the mean an confidence intervals based on the standard deviation, is favouring importance sampling; the Taylor method looks even more favourable if we instead plot the mean and standard deviations.

D.3.8 The bound in Theorem 6.2 under covariate shift

The bound in Theorem 6.2 is in a general form that applies to any shift in the CEF framework. In concrete cases, the bound can be made simpler, as we now demonstrate.

Suppose that X is a covariate that is Gaussian distributed $\mathcal{N}(0,1)$. Also consider a prediction target $Y := f_0(X) + \epsilon$ for some function f_0 and noise variable ϵ that is independent of X.

Suppose we consider a predictor $\hat{Y} = f(X)$ and apply our proposed methodology to estimate the mean squared prediction error when predicting $Y \approx f(X)$ under a mean shift of size $\delta \in \mathbb{R}$ to X. When we only consider shifts in the mean (and not the variance), the sufficient statistic is $T(X) = X$. We can use Theorem 6.2 to bound the prediction error. In this setting,

$$\ell = (Y - \hat{Y})^2 = (f_0(X) - f(X) + \epsilon)^2 \quad \text{and} \quad \epsilon_{t\cdot\delta T} = X - t \cdot \delta,$$

such that the bound in Theorem 6.2 becomes

$$\begin{aligned}&\Big|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\Big| \\ &\leq \tfrac{1}{2} \sup_{t\in[0,1]} \Big|\text{Cov}_{t\cdot\delta}\big((f_0(X) - f(X) + \epsilon)^2, (X - t\cdot\delta)^2\big) \\ &\qquad - \text{Cov}\big((f_0(X) - f(X) + \epsilon)^2, (f_0(X) - f(X) + \epsilon)^2, X^2\big)\Big| \cdot \delta^2.\end{aligned}$$

The subscript $\text{Cov}_{t\cdot\delta}$ indicates that the covariance is taken in the distribution $\mathcal{N}(t\cdot\delta, 1)$; instead we can write this in the observed distribution, and add $t\cdot\delta$ to X. Further, the terms relating to ϵ disappear, as they are independent of X. Thus, if we define the modelling error $g(x) = f_0(x) - f(x)$, we can write

$$\Big|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\Big| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \Big|\text{Cov}\big(g(X + t\cdot\delta)^2 - g(X)^2, X^2\big)\Big| \cdot \delta^2.$$

We can bound the covariance using the inequality $\text{Cov}(A, B) \leq \sqrt{\text{Var}(A)\text{Var}(B)}$,

$$\Big|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\Big| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \Big|\sqrt{\text{Var}\left((g(X + t\cdot\delta)^2 - g(X)^2\right)}\Big| \cdot \Big|\sqrt{\text{Var}(X^2)}\Big| \cdot \delta^2.$$

The first term on the right hand side is the variance of the difference of approximation error in X and in $X + t\delta$. If we are willing to make assumptions on the quality of the approximation f, we can simplify this further. For example, we can assume that $|g(x)^2 - g(y)^2| \leq C \cdot |x - y|^2$, meaning that the squared error of $f_0(x) - f(x)$ does not change faster than quadratically in x. In that case, we get

$$\Big|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\Big| \leq \tfrac{1}{2} C \Big|\sqrt{\text{Var}(X^2)}\Big| \cdot \delta^4.$$

In some cases, one can sharpen this bound by using prior knowledge about the data generating mechanism (for example, the data generating function f_0 may be bounded).

D.4 Limitations of worst-case conditional subpopulation shift for defining plausible robustness sets

For the example in Section 6.4.1, we can contrast the type of shift we consider with the worst-case $(1-\alpha)$-conditional subpopulation shift considered by Subbaswamy et al. (2021).

In this section, we will make the following points: First, worst-case conditional $(1-\alpha)$-subpopulation shifts can be too pessimistic, with even moderate values of α leading to implausible conditional distributions. Second, we will argue that parametric robustness sets enable more fine-grained control over the set of plausible shifts, leading to more informative estimates of worst-case risk. Overall, we argue that the two approaches are complementary, with different strengths.

Before we proceed, we define a conditional $(1-\alpha)$ subpopulation shift. A $(1-\alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ is defined by a weighting function $h : \mathcal{O} \times \mathcal{Y} \mapsto [0, 1]$, which has the property that $\mathbb{E}[h(O, Y)|Y] = 1 - \alpha$ for all values of Y. This can be used to construct a worst-case objective, which measures the worst-case loss under such a shift:

$$\begin{aligned} \sup_{h:\{0,1\}^2 \mapsto [0,1]} \quad & \frac{1}{(1-\alpha)} \mathbb{E}[h(O,Y)\mu(O,Y)] \\ \text{s.t.} \quad & \mathbb{E}[h(O,Y)|Y=y] = 1-\alpha, \quad \text{for } y \in \{0,1\} \end{aligned} \tag{D.3}$$

where $\mu(O, Y) \coloneqq \mathbb{E}[\ell(Y, f)|O, Y]$, for a predictor f and loss ℓ. This has the effect of leaving the distribution $\mathbb{P}(Y)$ untouched, while changing the conditional distribution $\mathbb{P}(O|Y)$. Throughout this section, we will use the same predictor $f(O, L)$ described in Section 6.4.1. The rest of this section is structured as follows:

In Appendix D.4.1, we derive the feasible set of conditional distributions $\mathbb{P}(O|Y)$ implicitly considered by this objective in the simple generative model of Section 6.4.1, which only involves variables O, L and Y. We do so by showing that (for discrete O, Y), maximizing Equation (D.3) over h is equivalent to solving a linear program, where we can characterize the constraints on h exactly, and translate them into constraints on $\mathbb{P}(O = 1|Y = 1), \mathbb{P}(O = 1|Y = 0)$. Here, we show that the resulting feasible set is quite large, even for moderately large subpopulations. In particular, whenever $(1 - \alpha) < \min\{\mathbb{P}(O = 1|Y = 0), \mathbb{P}(O = 0|Y = 1)\}$, all conditional distributions are possible.

In Appendix D.4.2, we derive the value of h that maximizes Equation (D.3), and show that, as we vary α, the worst-case shift is always in the same "direction" probability space: Healthy patients ($Y = 0$) are tested more, and sick patients ($Y = 1$) are tested less, and for $\alpha < 0.27$, the worst-case subpopulation shift is the (unrealistic) scenario where healthy patients are always tested, and sick patients are never tested.

In Appendix D.4.3, we illustrate how this type of behavior can be avoided with our approach. We first give a parameterized shift function $s(Z; \delta_0, \delta_1)$ such that we can reach any conditional distribution of $\mathbb{P}(O|Y)$, for sufficiently large values of δ_0, δ_1. We then demonstrate how an iterative process might play out with domain experts, where we consider different constraint sets until we find a constraint set that contains plausible shifts.

D.4.1 Feasible conditional subpopulations in Section 6.4.1

For the simple example in Section 6.4.1, we give a self-contained derivation of the feasible region for $1 - \alpha$ conditional subpopulations in the distribution $\mathbb{P}(O|Y)$. The advantage of working with this simple generative model is that the conditional distribution can be described by only two numbers, $\mathbb{P}(O = 1|Y = 1)$ and $\mathbb{P}(O = 1|Y = 0)$, and so we can visualize the resulting conditional distribution.

Because O, Y are discrete, the worst-case subpopulation in this simple example can be solved via a linear program, for a fixed α. We have an optimization problem in two variables, since $h_{11}\mathbb{P}(O = 1|Y = 1) + h_{01}\mathbb{P}(O = 0|Y = 1) = 1 - \alpha$, and likewise for h_{10}, h_{00}, where $h_{ij} = h(O = i, Y = j)$. We also have the constraint that each variable must live in $[0, 1]$. Meanwhile, the loss to maximize is a linear function, as an expectation of $\mathbb{E}[h(O,Y)\mu(O,Y)]$, where $\mu(O,Y)$ takes on four possible values, where we write $p_{ij} = \mathbb{P}(O = i|Y = j)$, and μ_{ij} similarly.

$$\begin{aligned} \max_{h \in \mathbb{R}^{2\times 2}} \quad & h_{00}\mu_{00} + h_{10}\mu_{10} + h_{01}\mu_{01} + h_{11}\mu_{11} \\ \text{s.t.,} \quad & h_{11}p_{11} + h_{01}(1-p_{11}) = 1-\alpha \\ & h_{10}p_{10} + h_{00}(1-p_{10}) = 1-\alpha \\ & 0 \leq h_{ij} \leq 1, \forall i,j \end{aligned} \tag{D.4}$$

This linear program is simple enough to solve by hand, and we will do here to build intuition. In this section, we begin by characterizing the feasible region of h, and then translating that into a feasible region for $\mathbb{P}_h(O|Y)$, which we can plot in two dimensions.

Characterizing feasible values of h: Here, we focus on characterizing the feasible set that h can lie in, as a way of characterizing the feasible set for $\mathbb{P}(O|Y)$. From the constraints, we can write that

$$\begin{aligned} h_{11}p_{11} + h_{01}(1-p_{11}) = 1-\alpha \qquad & \implies \qquad h_{01} = \frac{1-\alpha-h_{11}p_{11}}{1-p_{11}} \\ h_{10}p_{10} + h_{00}(1-p_{10}) = 1-\alpha \qquad & \implies \qquad h_{00} = \frac{1-\alpha-h_{10}p_{10}}{1-p_{10}} \end{aligned}$$

There are only two constraints on h_{11}: Those directly imposed by $0 \leq h_{11} \leq 1$, and those which are imposed by the equality constraint with h_{01} and the fact that $0 \leq h_{01} \leq 1$. For the latter, with some algebra we can write that

$$0 \leq \frac{1-\alpha-h_{11}p_{11}}{1-p_{11}} \leq 1 \qquad \implies \qquad \frac{p_{11}-\alpha}{p_{11}} \leq h_{11} \leq \frac{1-\alpha}{p_{11}}$$

So that the constraints on h_{11} become

$$\max\left\{0, \frac{p_{11}-\alpha}{p_{11}}\right\} \leq h_{11} \leq \min\left\{1, \frac{1-\alpha}{p_{11}}\right\} \tag{D.5}$$

which recovers our intuition that if $\alpha = 0$, it must be that $h_{11} = 1$ and $h_{01} = 1$.

Bounding feasible values of $\mathbb{P}_h(O|Y)$ The parameters h can be understood as importance weights whose expectation is $1 - \alpha$ instead of 1, that reweight $\mathbb{P}$ to a new distribution $\mathbb{P}_h$ when appropriately normalized. To compute conditional probabilities $\mathbb{P}_h(O = i|Y = j)$ under the new distribution, we can compute the expectation of $\mathbf{1}\{O = i, Y = j\}$, and normalize by $\mathbb{P}(Y = j)$.

$$\mathbb{P}_h(O = i, Y = j) = \frac{1}{1-\alpha}\mathbb{E}[h(O, Y)\mathbf{1}\{O = i, Y = j\}] = \frac{h_{ij}}{1-\alpha}\mathbb{P}(O = i, Y = j)$$
$$\implies \mathbb{P}_h(O = i|Y = j) = \frac{h_{ij}}{1-\alpha}\mathbb{P}(O = i|Y = j)$$

where the implication follows from the fact that $\mathbb{P}_h(Y) = \mathbb{P}(Y)$. This allows us to translate bounds on h_{ij} directly into bounds on $\mathbb{P}_h(O = i|Y = j)$. Making use of Equation (D.5), we can write that

$$\max\left\{0, \frac{p_{11} - \alpha}{p_{11}}\right\} \cdot \frac{p_{11}}{1-\alpha} \leq \mathbb{P}_h(O = 1|Y = 1) \leq \min\left\{1, \frac{1-\alpha}{p_{11}}\right\} \cdot \frac{p_{11}}{1-\alpha}$$

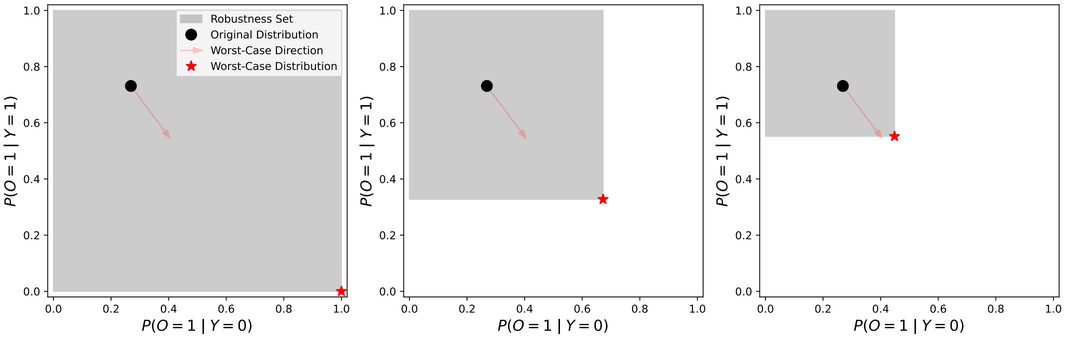
which yields

$$\max\left\{0, \frac{p_{11} - \alpha}{1-\alpha}\right\} \leq \mathbb{P}_h(O = 1|Y = 1) \leq \min\left\{\frac{p_{11}}{1-\alpha}, 1\right\}$$

We can apply a similar logic to h_{10}, which is identical except for p_{11} being replaced by p_{10}, yielding

$$\max\left\{0, \frac{p_{10} - \alpha}{1-\alpha}\right\} \leq \mathbb{P}_h(O = 1|Y = 0) \leq \min\left\{\frac{p_{10}}{1-\alpha}, 1\right\}$$

Visualizing the constraint set: Figure D-5 gives feasible conditional distributions under different values of α. We can observe that when $\alpha = 0.8$, all conditional distributions are feasible, including the distribution where $\mathbb{P}(O = 1|Y = 0) = 1$ and $\mathbb{P}(O = 1|Y = 1) = 0$, representing the case where every healthy patient gets tested, and no sick patients receive a test. This is generally possible in this example whenever $1 - \alpha < \min\{\mathbb{P}(O = 1|Y = 0), \mathbb{P}(O = 0|Y = 1)\}$, as it permits the following



(a) $(1-\alpha) = 0.2$ **(b)** $(1-\alpha) = 0.4$ **(c)** $(1-\alpha) = 0.6$

Figure D-5: *Feasible sets, worst-case directions, and worst-case solutions for a $(1-\alpha)$ subpopulation shift in the conditional distribution $\mathbb{P}(O|Y)$ for differing values of α. Worst-case directions are computed using Equation (D.6), as unit-norm vectors rescaled to fit in the plot, and the colored dots give the worst-case solutions, all of which lie in the lower-right corner of the constraint set. The original conditional distribution is given by the black dot.*

subpopulation function, which yields this result.

$$h(O=o, Y=y) = \frac{1-\alpha}{\mathbb{P}(O=o|Y=y)} \mathbf{1}\{o \neq y\}$$

D.4.2 Worst-case conditional subpopulation shifts

Given the constraint set which describes the feasible set of conditional distributions under the $(1-\alpha)$-conditional subpopulation objective, we can derive the worst-case conditional distribution. Here, since Y, O are both binary, the expected loss under a new distribution $\mathbb{P}_h$ is given by

$$\mathbb{E}_h[\ell] = \sum_{y,o} \mu(o,y)\mathbb{P}_h(O=o|Y=y)\mathbb{P}(Y=y)$$

which we can write in terms of the constrained probabilities $\mathbb{P}_h$ as follows, where $q_{11} \coloneqq \mathbb{P}_h(O=1|Y=1)$ and $q_{10} \coloneqq \mathbb{P}_h(O=1|Y=0)$

$$\mathbb{P}(Y=1)[\mu(1,1)q_{11} + \mu(0,1)(1-q_{11})] + \mathbb{P}(Y=0)[\mu(1,0)q_{10} + \mu(0,0)(1-q_{10})]$$

which also gives us a direction in which the loss is maximized, since the loss is given by

$$\mathbb{E}_h[\ell] = q_{11} \cdot \mathbb{P}(Y = 1) \cdot (\mu(1,1) - \mu(0,1)) + q_{10}\mathbb{P}(Y = 0) \cdot (\mu(1,0) - \mu(0,0)) + C \quad \text{(D.6)}$$

where $C = \mathbb{P}(Y = 1)\mu(0,1) + \mathbb{P}(Y = 0)\mu(0,0)$. Since q_{11}, q_{10} can be optimized independently, the worst-case solution is given by taking the maximum value of q_{11} if $\mu(1,1) > \mu(0,1)$ and the minimum value if $\mu(1,1) < \mu(0,1)$, and likewise taking the maximum value of q_{10} if $\mu(1,0) > \mu(0,0)$, and the minimum value otherwise. If $\mu(1,1) = \mu(0,1)$ or $\mu(1,0) = \mu(0,0)$, then the objective is unaffected by the choice of q_{11} or q_{10} respectively.

Visualizing the worst-case conditional distributions The worst-case directions on the probability scale, and the resulting worst-case conditional distribution obtained by solving Equation (D.4), are given in Figure D-5. The red line arrow visualizes the direction from Equation (D.6), and the worst-case distribution is the point which is furthest in this direction in the constraint set. Here, we are finding the worst-case accuracy of the same predictive model $f(O, L)$ described in Section 6.4.1. We can observe that the worst-case loss is obtained by seeking to reverse the correlation between Y and O, decreasing the probability that a sick patient ($Y = 1$) gets a test ordered, and increasing the probability that a healthy patient ($Y = 0$) gets a test ordered.

D.4.3 Iterating with domain experts to define realistic parametric robustness sets

In the previous sections, we saw that $(1-\alpha)$-conditional subpopulation shift does not always produce realistic worst-case conditional distributions. Moreover, given only the parameter α, there is limited ability to control the nature of the resulting worst-case conditional distribution $\mathbb{P}(O|Y)$. In this section, we contrast this limitation with the finer-grained control enabled by considering parametric robustness sets. In particular, we argue that parametric shifts allow for end-users to customize robustness sets, ruling

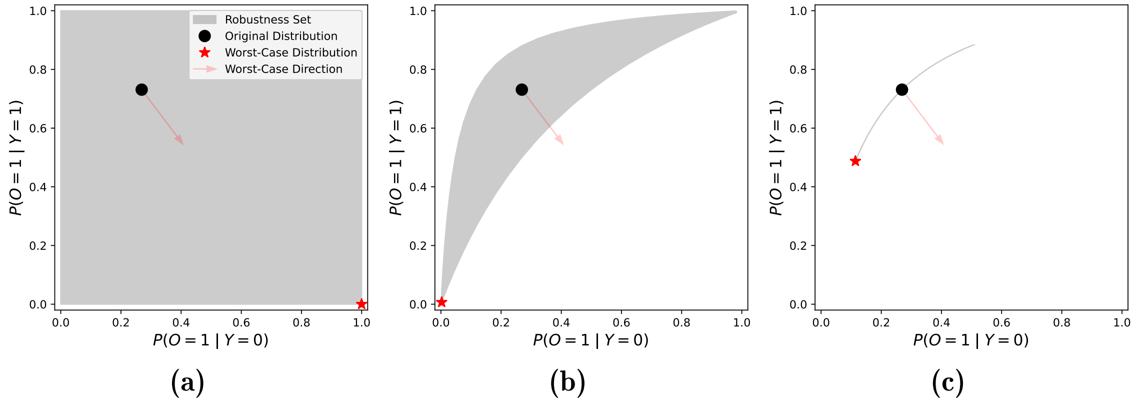


Figure D-6: *Each figure shows the set of conditional probability distributions ("CPDs") $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the worst-case distribution (given by the red star) for the 0–1 loss. In this example, the expected loss under $\mathbb{P}_\delta$ is a linear function of the two conditional probabilities (see Appendix D.4.2), where the loss increases along the red arrow. (a) captures (nearly) all conditional probability distributions, with Δ_0, Δ_1 unconstrained. (b) shows a set of CPDs with Δ_0 unconstrained, and $\Delta_1 = [-1, 1]$, with resulting worst-case accuracy of 50%. (c) shows a more restrictive set of shifts, where $\Delta_0 = [-1.05, 1.05], \Delta_1 = \{0\}$. The worst-case accuracy in this case is 69%, comparable to the accuracy of 75% on the original distribution.*

out shifts that represent unrealistic changes.

In practice, we imagine that the following iterative process could be a useful tool in model development: (i) Define a class of shifts with an appropriate $s(Z; \delta)$ and constraint set Δ, and search for a worst-case shift δ. (ii) Present to domain experts **both** the worst-case shift δ (in terms of summary statistics of the resulting distribution $\mathbb{P}_\delta$) alongside the associated estimate of the worst-case loss. For instance, report both the worst-case loss, as well as corresponding rate of testing among sick and healthy patients. (iii) If the shift itself is unrealistic, further the constrain parameter set or shift function, and repeat the process.

In Figure D-6, we give a concrete example. Each sub-figure shows the set of conditional probability distributions $\mathbb{P}(O|Y)$ that can be represented by a shift of $(\delta_0, \delta_1) \in \Delta_0 \times \Delta_1$, along with the worst-case conditional distribution (given by the red star) for the 0–1 loss. Recall that we use the shift function $s(Y; \delta) = \delta_0 + \delta_1 Y$, where δ_0 controls a general increase or decrease in testing, while δ_1 controls a shift in the testing rate for only sick patients, and allows for a different change in the testing rate of sick vs healthy patients.

Iteration 1: We might imagine starting with a relatively unconstrained robustness set, where δ_0 and δ_1 are unconstrained. Figure D-6a shows the resulting robustness set of conditional distributions, and finds a shift with with a worst-case accuracy of 16%, compared to accuracy of 75% on the original distribution. However, the corresponding δ-perturbation $\mathbb{P}_\delta$ is unrealistic, where all healthy patients (and no sick patients) are tested. Luckily, because we have parameterized the shift, we can constrain the robustness set to exclude these types of results.

Iteration 2: A benefit of our approach is that we can refine the robustness set, with this type of feedback in mind. In Figure D-6b, we restrict the support of δ_1 to $[-1, 1]$, to avoid large changes in the relative probability of testing sick vs healthy patients. Here, the resulting worst-case accuracy is much higher (50%), but the corresponding worst-case conditional probability distribution is perhaps still unrealistic: No patients undergo laboratory testing at all!

Iteration 3: Finally, we consider only shifts that affect all patients in a similar way, generally raising or lowering the conditional probability of a lab test, represented by shifts in δ_0 alone. This may correspond to a more realistic scenario where (in a new hospital) laboratory testing use is more or less constrained. Additionally, we can specify that this shift should decrease testing rates by at most 20%, which translates directly into a lower-bound on δ_0.[4] Figure D-6c shows the resulting robustness set of distributions, where the worst-case shift may seem more plausible: A reduction in testing rates for both populations. The worst-case accuracy in this case is 69%, comparable to the accuracy of 75% on the original distribution.

[4] In Proposition D.2.1, we prove that for binary random variables with a shift $\eta(Z) + \delta$, there is a one-to-one mapping between a new marginal distribution ($\mathbb{P}(O = 1)$ in this case) and the value of the parameter δ.

D.5 CelebA: Experiment details and additional results

In this section, we give details of the computer vision experiment in Section 6.4.2.

D.5.1 Details for the experiment

Creating the training distribution To construct the training distribution $\mathbb{P}$, we use the conditional GAN in Kocaoglu et al. (2018). In particular, we use their CausalBEGAN, which is an extends the boundary equillibrium GAN (Berthelot et al., 2017) to also take attributes as inputs. We train the CausalBEGAN using the default hyper parameters in the implementation provided by Kocaoglu et al. (2018), available under the MIT license. The model is trained for 250,000 iterations on a single GPU, taking around approximately 16 hours.

Similar to Kocaoglu et al. (2018), we use the CelebA dataset (Liu et al., 2015), which contains approximately 200,000 images of faces, along 40 binary attributes. Of those, we use the following 9 attributes {Male, Young, Wearing Lipstick, Bald, Mustache, Eyeglasses, Narrow Eyes, Smiling, Mouth Slightly Open}. The CelebA dataset is licensed for non-commercial research purposes only, and consists of publicly available images of celebrities, which were collected from the internet. Although the data set has been widely used, Liu et al. (2015) do not make any mention of consent by the individuals to have the images included in the data set, and it is therefore likely that those celebrities did not provide consent.

Training distribution over attributes For the training distribution, we simulate binary attributes according to the structural causal model in Figure 6-4 (for convenience also copied to Figure D-7), where the model parameters are

$$\mathbb{P}(\text{Young} = 1) = \sigma(0.0)$$

$$\mathbb{P}(\text{Male} = 1) = \sigma(0.0)$$

$$\mathbb{P}(\text{Eyeglasses} = 1|\text{Young}) = \sigma(0.0 - 0.4 \cdot \text{Young})$$

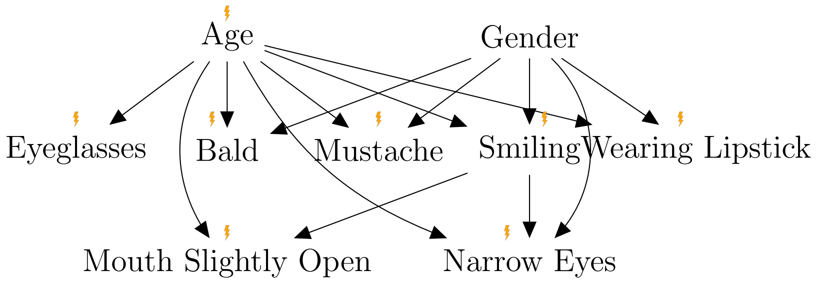


Figure D-7: *Causal graph over attributes, where lightning bolts indicate changes in mechanisms. Also displayed in Figure 6-4.*

$$\begin{aligned}
\mathbb{P}(\text{Bald} = 1|\text{Young, Male}) &= \sigma(-3.0 + 3.5 \cdot \text{Male} - \text{Young}) \\
\mathbb{P}(\text{Mustache} = 1|\text{Young, Male}) &= \sigma(-2.5 + 2.5 \cdot \text{Male} - \text{Young}) \\
\mathbb{P}(\text{Smiling} = 1|\text{Young, Male}) &= \sigma(0.25 - 0.5 \cdot \text{Male} + 0.5 \cdot \text{Young}) \\
\mathbb{P}(\text{Wearing Lipstick} = 1|\text{Young, Male}) &= \sigma(3.0 - 5.0 \cdot \text{Male} - 0.5 \cdot \text{Young}) \\
\mathbb{P}(\text{Mouth Slightly Open} = 1|\text{Young, Smiling}) &= \sigma(-1.0 + 0.5 \cdot \text{Young} + \text{Smiling}) \\
\mathbb{P}(\text{Narrow Eyes} = 1|\text{Male, Young, Smiling}) &= \sigma(-0.5 + 0.3 \cdot \text{Male} + 0.2 \cdot \text{Young} + \text{Smiling}),
\end{aligned}$$

where each variable either takes the value 0 or 1 and σ indicates the sigmoid. To generate data, we first simulate attributes from this binary Bayesian network, which we then pass as inputs to the GAN to simulate images (in addition to the random noise used by the GANs to simulate different images). In Figures D-8 and D-9, we plot examples of the training images that were generated.

Predictive model We simulate a training set of 12,000 attribute-image pairs, and a validation set of 2,000 pairs. The training set is used to fit a classifier f, and the validation set is used for model selection. To build a classifier f, we use the ResNet-50 (He et al., 2016) model implemented in the python package `torch`. We add a final fully connected layer to adapt the ResNet model to a binary classification task, and fine-tune the model on the training data by (only) learning the weights and bias of the final layer. The model is trained using the negative log-likelihood criterion and an ADAM optimizer. The model is trained for 25 epochs and we select the model which after a full epoch had the best validation set performance. Given the learned model f, we simulate a separate validation dataset of $n = 1{,}000$ samples, and make model predictions $f(X)$. We then compute the model accuracy as $\ell = \mathbf{1}\{f(X) = Y\}$, which is the input to computing the shift gradient and Hessian.

Estimation of shifted loss We apply the methods in Section 6.3.2 to estimate the worst-case shift to the distribution $\mathbb{P}$ (given by the binary probabilities above). For each conditional $\mathbb{P}(W_i | \operatorname{PA}(W_i))$, we consider a shift $\eta_{\delta_i}(\operatorname{PA}(W_i)) = \eta(\operatorname{PA}(W_i)) + \sum_{z \in \mathcal{Z}} \mathbf{1}\{\operatorname{PA}(W_i) = z\}\, \delta_i$, which corresponds to arbitrarily shifting the conditional distribution (see Appendix D.3.4). For example, for $W_i =$ Bald, where $\eta(\text{Young, Male}) = -3.0 + 3.5 \cdot \text{Male} - 1.0 \cdot \text{Young}$, the shift would be

$$\eta_{\delta_{\text{Bald}}}(\text{Young, Male}) = \eta(\text{Young, Male}) + \begin{cases} \delta_{\text{Bald},0}, & \text{Young} = 0, \text{Male} = 0 \\ \delta_{\text{Bald},1}, & \text{Young} = 0, \text{Male} = 1 \\ \delta_{\text{Bald},2}, & \text{Young} = 1, \text{Male} = 0 \\ \delta_{\text{Bald},3}, & \text{Young} = 1, \text{Male} = 1. \end{cases} \tag{D.7}$$

For each W_i, this means that δ_i is $\mathbb{R}^{2^{|\operatorname{PA}(W_i)|}}$, and in total $\delta = (\delta_1, \ldots, \delta_8) \in \mathbb{R}^{31}$ (we do not consider shifts in the distribution of gender, since this is the label we are predicting).

We compute the shift gradient and Hessian using Theorem 6.1. In particular, since W_i is binary, the sufficient statistic is $T(W_i) = W_i$, so the shift gradients and Hessians given by Appendix D.3.4. See Appendix D.3.1 for a detailed walk through of computing the shift gradient and Hessian from a sample.

For any given δ, the shifted distribution of W_i is given by $\mathbb{P}_\delta(W_i = 1 | \operatorname{PA}(W_i)) = \sigma(\eta_{\delta_i})$, where η_{δ_i} is computed similar to Equation (D.7), and σ is the sigmoid function. Then the importance sampling weights are given by

$$w_\delta = \prod_{i=1}^{8} \frac{\sigma(\eta_{\delta_i}(\operatorname{PA}(W_i)))}{\sigma(\eta(\operatorname{PA}(W_i))}.$$

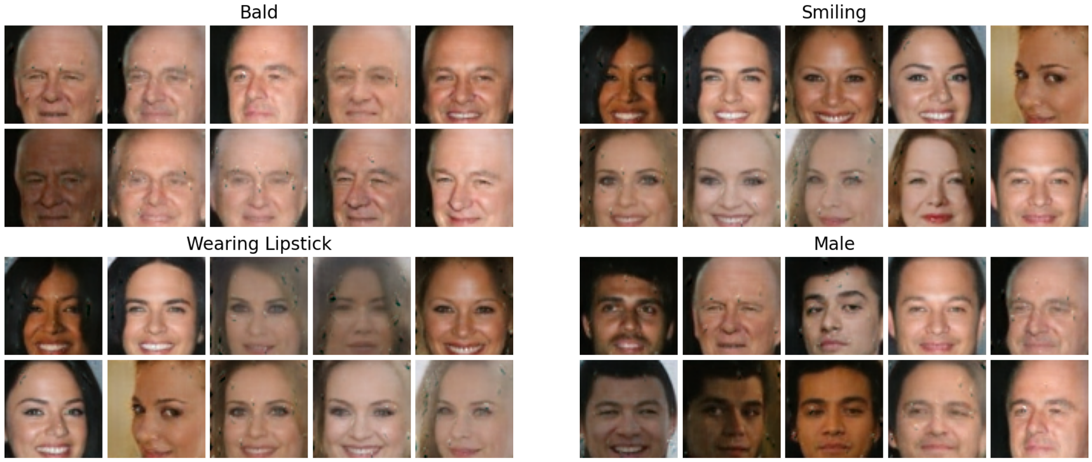


Figure D-8: *Examples of images from the training distribution $\mathbb{P}$. Each of the four groups (Bald, Smiling, Wearing Lipstick, Male) show training images who have that characteristic.*

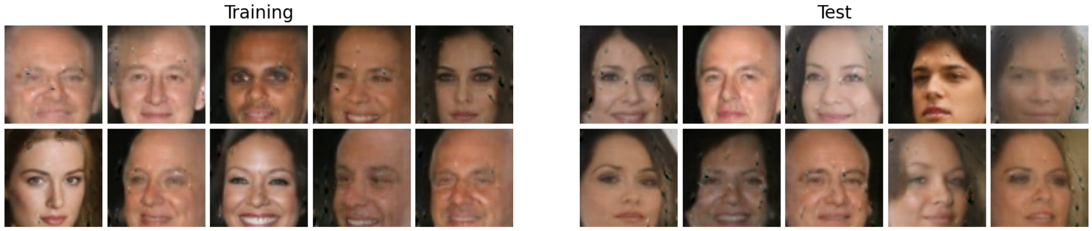


Figure D-9: *Examples of images from the training distribution $\mathbb{P}$ and the test distribution $\mathbb{P}_\delta$ that is characterized by the worst-case shift δ, see Figure 6-4.*

Using these weights, for any δ, we can estimate $\mathbb{E}_\delta[\ell]$ by $\hat{E}_{\delta,\text{IS}}$ and $\hat{E}_{\delta,\text{Taylor}}$ using Equations (6.6) and (6.8), respectively.

D.5.2 Full table of worst-case shift in Section 6.4.2

In Section 6.4.2, we find the worst-case shift δ, and display the 5 largest components. In Table D.2, we display the full vector $\delta \in \mathbb{R}^{31}$, sorted by absolute value of the size of the component.

Conditional	δ_i
Bald | Male= 0, Young= 0	0.899
Bald | Male= 1, Young= 1	-0.800
Bald | Male= 1, Young= 0	-0.680
Wearing Lipstick | Male= 0, Young= 1	-0.618
Wearing Lipstick | Male= 0, Young= 0	-0.543
Eyeglasses | Young= 1	0.507
Mustache | Male= 1, Young= 0	-0.476
Mustache | Male= 0, Young= 0	0.449
Mustache | Male= 1, Young= 1	-0.415
Eyeglasses | Young= 0	0.399
Smiling | Male= 0, Young= 0	-0.261
Wearing Lipstick | Male= 1, Young= 0	0.205
Narrow Eyes | Male= 0, Smiling= 0, Young= 0	0.192
Mouth Slightly Open | Smiling= 1, Young= 1	0.191
Smiling | Male= 1, Young= 0	0.183
Narrow Eyes | Male= 1, Smiling= 1, Young= 1	0.179
Mouth Slightly Open | Smiling= 0, Young= 1	-0.153
Mustache | Male= 0, Young= 1	0.133
Bald | Male= 0, Young= 1	0.128
Mouth Slightly Open | Smiling= 1, Young= 0	-0.127
Narrow Eyes | Male= 0, Smiling= 1, Young= 0	-0.125
Wearing Lipstick | Male= 1, Young= 1	0.123
Narrow Eyes | Male= 1, Smiling= 1, Young= 0	-0.117
Narrow Eyes | Male= 0, Smiling= 0, Young= 1	0.106
Young | No parents	0.092
Narrow Eyes | Male= 0, Smiling= 1, Young= 1	0.057
Narrow Eyes | Male= 1, Smiling= 0, Young= 1	-0.050
Narrow Eyes | Male= 1, Smiling= 0, Young= 0	-0.039
Mouth Slightly Open | Smiling= 0, Young= 0	0.028
Smiling | Male= 1, Young= 1	0.028
Smiling | Male= 0, Young= 1	0.017

Table D.2: *Worst case shift in the $\delta \in \mathbb{R}^{31}$ identified by the Taylor approach in Section 6.4.2. Each entry corresponds to a shift in a conditional distribution given a particular outcome, and the squared sum of the entries equal $\lambda^2 = 4$.*

D.5.3 Sample images from training distribution in Section 6.4.2

In Figure D-8, for the 4 attributes {Bald, Smiling, Wearing Lipstick, Male}, we display images generated from the training distribution $\mathbb{P}$ (i.e. by the GAN) with that particular attribute. In Figure D-9 we show 10 randomly drawn images from the training distribution $\mathbb{P}$ as well as the test distribution $\mathbb{P}_\delta$ corresponding to the worst-case δ found in Section 6.4.2.

D.5.4 Impact of changing λ

The shift considered in the main text yields a relatively small drop in accuracy. To demonstrate that larger drops in accuracy are possible, we repeated our experimental setup over the same 100 initial validation datasets, while varying the size of the constraint $\|\delta\|_2 \leq \lambda$. We report results in Table D.3 for $\lambda \in [2, 4, 6, 8, 10]$, where $\lambda = 2$ corresponds to the setting of Table 6.1b.

Table D.3: *Performance of the Taylor and IS approaches over different values of λ, where $\lambda = 2$ corresponds to the setting of Table 6.1b. Averages taken over 100 simulations.*

	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$	$\lambda = 8$	$\lambda = 10$
Original Acc. ($\mathbb{E}[\mathbf{1}\{f(X) = Y\}]$)	0.912	0.912	0.912	0.912	0.912
Acc. under Taylor shift ($\mathbb{E}_{\delta_{\text{Taylor}}}[\mathbf{1}\{f(X) = Y\}]$)	0.874	0.812	0.736	0.681	0.648
IS est. of acc. under Taylor shift ($\hat{E}_{\delta_{\text{Taylor}},\text{IS}}$)	0.863	0.795	0.715	0.658	0.625
Taylor est. of acc. under Taylor shift ($\hat{E}_{\delta_{\text{Taylor}},\text{Taylor}}$)	0.863	0.798	0.711	0.601	0.466
Acc. under IS shift ($\mathbb{E}_{\delta_{\text{IS}}}[\mathbf{1}\{f(X) = Y\}]$)	0.889	0.830	0.746	0.670	0.596
IS est. of acc. under IS Shift ($\hat{E}_{\delta_{\text{IS}},\text{IS}}$)	0.821	0.670	0.463	0.264	0.130

Recall that we have two complementary goals: First, we would like to **find** a shift that results in a large drop in accuracy. Second, we would like to reliably **evaluate** the impact of the shift that we find, using only the training data. These two goals can be tackled with different approaches, such as using the Taylor approximation to find a shift, but using importance sampling (IS) to estimate the loss under that shift. Table D.3 allows us to compare three different strategies: (i) using the Taylor approximation for both finding and evaluating the shift, (ii) using IS for both finding and evaluating, and (iii) using Taylor to find, but IS to evaluate the shift.

From Table D.3, we can observe that **using Taylor to find, but IS to evaluate, consistently performs best in terms of reliable evaluation** (i.e., predicting the shifted accuracy), across all values of λ. For $\lambda = 2$, the bias in evaluation is 1% (predicting 86% vs ground truth of 87% on average), and for $\lambda = 10$, the bias of this approach is still only 2% (predicting 63% vs ground truth of 65% on average). In contrast, for $\lambda = 10$, the first strategy (using Taylor to find and evaluate) over-predicts the impact by 18%, and the second strategy (using IS to find and evaluate) over-predicts the impact by 47%.

This strategy **also tends to find the most impactful shifts, for moderate values of λ**. For $\lambda \leq 6$, the shifts found by the Taylor approach are more impactful than those found by the IS approach. Moreover, the drop in accuracy remains substantial (e.g., a drop of around 17% at $\lambda = 6$). For $\lambda > 6$, the story is more subtle: The third approach (using IS to find and evaluate shifts) finds more impactful shifts, but (as noted previously) dramatically over-estimates their impact.

D.6 Relationship to other approaches

In this section, we give a more detailed discussion of how our work relates to other approaches for evaluation of distributional robustness and learning of robust models. Much of the content from Section 6.1.1 is duplicated here, but expanded upon to include other relevant work and detailed discussion.

Distributionally Robust Optimization/Evaluation with divergence measures: Distributionally robust optimization (DRO) seeks to learn models that minimize objectives of the form of Equation (6.1) (Duchi and Namkoong, 2021; Duchi et al., 2020b; Sagawa et al., 2020). We focus on proactive worst-case evaluation of a fixed model, not optimization, similar to Subbaswamy et al. (2021); Li et al. (2021), but major differences between our work and prior work lie in the **definition of the set of plausible future distributions** $\mathcal{P}$, often called an "uncertainty set" in the optimization literature, where the goal is to specify a set that captures expected shifts, without being overly conservative.

Shifts in $\mathbb{P}(X, Y)$: A conservative approach is to include all joint distributions $\mathbb{P}(X, Y)$ within a certain neighborhood of the training distribution. Many coherent risk measures can be written as a worst-case loss of this form. For instance, the Entropic Value-at-Risk (EVaR), with confidence level $1 - \alpha$, corresponds to the worst-case loss over a set of distributions $\mathcal{P} = \{P \ll P_0 : D_{KL}(P \| P_0) \leq -\ln \alpha\}$, where P_0 is the original distribution (Ahmadi-Javid, 2012). Similarly, the Conditional Value-at-Risk (CVaR) with parameter α can be seen as the worst-case loss over an uncertainty set obtained from a limiting f-divergence (see Example 3 of Duchi and Namkoong (2021)), including all α-fractions of the original distribution. These measures are appealing, in that they are straightforward to compute, but can be very conservative.

Indeed, such measures often reduce to only considering the distribution of the loss itself. CVaR, for instance is equivalent to sorting the training examples by their loss, and taking the average loss of the top α-fraction. To illustrate these limitations, it is straightforward to see that, using the 0-1 loss and a classifier with 80% accuracy, the worst-case loss under both of these measures is 1.0 for any $\alpha \leq 0.2$. This is intuitive for CVaR (since over 20% of samples are misclassified in the original distribution), and follows for EVaR from the fact that the binary distribution with probability $q = 1$ has a KL-divergence to the original distribution $p = 0.2$ of $-\ln 0.2$.

Lam (2016) consider a more general problem of estimating the worst-case performance of stochastic systems over infinitesimal changes in distribution, measured by Kullback-Leibler divergence. Their approach is applicable beyond machine-learning settings, and generalizes to e.g., worst-case waiting times in a queueing system. They demonstrate that for a sufficiently small neighborhood of distributions, this worst-case performance can be well-approximated by a Taylor expansion whose coefficients can be estimated from the original distribution.

Shifts in $\mathbb{P}(X)$ *alone*: Partially due to this overly-conservative behavior, there has been a line of work incorporating additional restrictions on the allowable shift (i.e., adding more assumptions). For instance, Duchi et al. (2020b) considers learning predictive models that optimize a worst-case loss similar to CVaR (a “worst-case subpopulation shift”), but where only $\mathbb{P}(X)$ is allowed to change, and $\mathbb{P}(Y \mid X)$ is assumed to be constant. For similar shifts, Li et al. (2021) considers only the task of evaluation, but provides a novel estimation procedure with dimension-free finite-sample guarantees. However, many real-world shifts do not fit this framework: In Example 6.1, both $\mathbb{P}(X)$ and $\mathbb{P}(Y \mid X)$ are changing, where $X = (A, O, L)$, as a result of a shift in $\mathbb{P}(O \mid Y, A)$.

Shifts in a conditional distribution: Closer to our work is Subbaswamy et al. (2021) who consider evaluating the loss under worst-case changes in a conditional distribution, but while we consider parametric shifts, they estimates the loss under worst-case $(1 - \alpha)$ conditional subpopulation shifts. However, it is not obvious how to choose an appropriate level of α: in some settings, seemingly plausible values of α (e.g., a 20% subpopulation) correspond to entirely implausible shifts. We give a simple lab-testing example in Appendix D.4, where the worst-case subpopulation is one where healthy patients are always tested, and sick patients never tested.

In contrast to these methods, our approach uses explicit parametric perturbations to define shifts, as opposed to distributional distances or subpopulations. In addition, our approach allows for shifts in multiple marginal or conditional distributions simultaneously: In Example 6.1, for instance, we can model a simultaneous change in both the marginal distribution of age, as well as the conditional distribution of lab testing, while other conditionals are unchanged. Our main requirement is that each shifting distribution is exponential family, and that the shift can be represented via the natural parameters: For continuous variables this is a non-trivial restriction, but for discrete variables it is true by definition.

Causality-motivated methods for learning robust models: Several approaches seek to learn models that perform well under arbitrarily large causal interventions (which result in arbitrary changes in selected conditional distributions). Several approaches proactively specify shifting mechanisms/conditional distributions, and then seek to learn predictors that have good performance under arbitrarily large changes in these mechanisms (Subbaswamy et al., 2019; Veitch et al., 2021; Makar et al., 2022; Puli et al., 2022). Other approaches use auxiliary information, such as environments (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2019) or identity indicators (Heinze-Deml and Meinshausen, 2021) to learn models that rely on invariant conditional distributions. The worst-case optimality of these approaches is often restricted to cases where the shifts are arbitrarily large: In Example 6.1, worst-case optimality under arbitrarily large shifts would correspond to minimizing the worst-case loss under all possible lab testing policies.

However, when the causal interventions (i.e., changes in causal mechanisms) are bounded (i.e., not arbitrary), then these approaches are not necessarily optimal. Closest to our work in motivation is prior work on robustness to bounded shift interventions in linear causal models (Rothenhäusler et al., 2021; Oberst et al., 2021b; Kook et al., 2022). Our work can be seen as extending those ideas to general non-linear causal models, where our focus is on evaluation rather than learning robust models. We discuss this point in more detail in Appendix D.6.1 below.

Our work can serve as an aid to deploying these causality-motivated methods in a few ways, by comparing their worst-case performance under bounded shifts: First, our work can inform whether such methods should be deployed at all, as for sufficiently small shifts, it may be the case that standard training yields better performance. Second, our work can inform hyperparameter selection for several of these approaches, which include regularization terms that implicitly trade off between robustness and in-distribution performance. More broadly, our approach is useful for probing (and comparing) the reliability of specific learned models under shift, regardless of the algorithm that produced them.

Evaluating out-of-distribution performance with unlabelled samples: A recent line of work has focused on predicting model performance in out-of-distribution settings, where unlabelled data is available from the target distribution (Garg et al., 2022; Jiang et al., 2022; Chen et al., 2021). In contrast, our method operates using only samples from the original source distribution, and seeks to estimate the worst-case loss over a set of possible target distributions.

D.6.1 The importance of considering restricted shifts in causal mechanisms

In Figure D-10 we revisit Example 6.1, adopting the perspective of a model developer, who is aware that laboratory testing policies (i.e., $P(O \mid A, Y)$) may change. As this change may impact the correlation between laboratory testing features (O, L) and the label Y, how should the model developer proceed?

From a causal perspective, one way to approach model development is to learn a predictive model that is "causal" in the sense that it only relies on the causal parents of the label Y. In this example, A is the full set of causal parents of Y, and the conditional distribution $P(Y = 1 \mid A)$ does not change under changes in laboratory testing policy. This conditional distribution is an example of an "invariant" conditional distribution (Rojas-Carulla et al., 2018), reflecting the unchanging causal mechanisms that generate Y which are not affected by changes in laboratory testing policy. With this in mind, we consider the choice between two models:

- **Age-based model**: $f(A) \approx P(Y = 1 \mid A)$, predicting disease using age alone.[5]
- **Full model**: $f(A, O, L) \approx P(Y = 1 \mid A, O, L)$, predicting disease using all features.

We now demonstrate the utility of incorporating additional knowledge, considering not only "what" can change (i.e., $P(O \mid A, Y)$), but also considering "how" and "how much" it can change, and translating that knowledge into a quantitative comparison between these modelling choices. The question of "how" corresponds to our choice of shift function, and "how much" corresponds to our choice of constraints on shift parameters. We consider changes in testing that correspond to a uniform increase/decrease in testing rates, parameterized as

$$P_\delta(O = 1 \mid A, Y) = \text{sigmoid}(\eta(A, Y) + \delta) \tag{D.8}$$

Other details of the underlying distribution are given in Appendix D.1.

In Figure D-10 (right), we plot the loss of each model under distributions[6] that corre-

[5] Details of how the full model $f(A, O, L)$ is trained are described in Appendix D.1. The model $f(A)$ is trained using unregularized logistic regression. Both models are trained on data drawn from the original distribution, where the marginal testing rate is 50%.

[6] In this case, every choice of δ maps to a unique marginal testing rate in the distribution P_δ (see Proposition D.2.1), so we plot the loss as a function of testing rate, instead of δ directly.

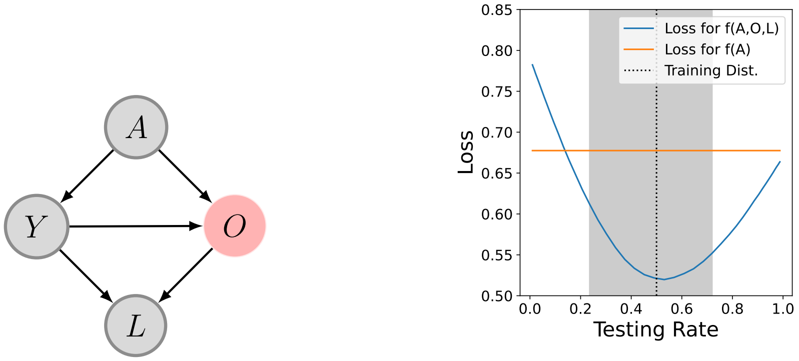


Figure D-10: *(Left) Causal graph for Example 6.1, where the variables are $Y \in \{0, 1\}$ for the label (Disease), $A \in \mathbb{R}$ for Age, $O \in \{0, 1\}$ for whether a laboratory test is ordered (Test Order), and $L \in \mathbb{R}$ for the lab result (Test Result), if available. (Right) Using the same generative model as in Appendix D.1, we contrast the performance of the full model $f(A, O, L)$ and a model $f(A)$ that only uses age, across distributions which differ in testing rates according to $P_\delta(O = 1 \mid A, Y) = \text{sigmoid}(\eta(A, Y) + \delta)$. Comparing performance on a range of distributions where we vary δ, we observe that $f(A)$ has invariant loss, but $f(A, O, L)$ has better performance for a wide range of shifts δ. In particular, if we compare the worst-case loss under shifts $|\delta| \leq 1.5$ (corresponding to marginal testing rates in the grey region), we can observe that the worst-case loss of $f(A, O, L)$ is lower than that of $f(A)$.*

spond to different choices of δ, and observe that despite having invariant performance, the age-based model only out-performs the full model under substantial changes in testing policy. In this case, the model $f(A)$ (throwing away laboratory testing information) yields better performance if testing rates drop substantially, but for a large set of changes in testing rates, the full model $f(A, O, L)$ is superior.

Considering the worst-case performance of each model can guide model selection. If a substantial change in testing rates is not plausible (which can be expressed as constraints on δ), and the worst-case loss (over plausible changes) of $f(A, O, L)$ is lower than that of $f(A)$, the model developer may decide to use the full model $f(A, O, L)$ in any case.

D.7 Proofs

D.7.1 Proof of Proposition 6.1

Proposition 6.1. *For any* $\mathbb{P}_\delta(\mathbf{V}), \mathbb{P}(\mathbf{V})$ *that satisfy Definition 6.4,* $\operatorname{supp}(\mathbb{P}) = \operatorname{supp}(\mathbb{P}_\delta)$ *and the density ratio* $w_\delta := \mathbb{P}_\delta/\mathbb{P}$ *is given by*

$$w_\delta(\mathbf{V}) = \exp\left(\sum_{i=1}^m s_i(Z_i;\delta_i)^\top T_i(W_i)\right)\exp\left(\sum_{i=1}^m h(\eta_i(Z_i)) - h(\eta(Z_i) + s_i(Z_i;\delta_i))\right).$$

Proof. By Definition 6.4 and Assumption 6.1, we have that

$$\mathbb{P}_\delta(\mathbf{V}) = \prod_{i=1}^m \mathbb{P}_{\delta_i}(W_i|Z_i) \prod_{V_j\in\mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j)$$
$$\mathbb{P}(\mathbf{V}) = \prod_{i=1}^m \mathbb{P}(W_i|Z_i) \prod_{V_j\in\mathbf{V}\setminus\mathbf{W}} \mathbb{P}(V_j|U_j).$$

It follows that the supports of $\mathbb{P}_\delta$ and $\mathbb{P}$ are the same: Since the exponential family density is given by the base measure $g_i(W_i)$ times a exponential term (which is always strictly positive), and since the terms $\prod_{V_j\in\mathbf{V}\setminus\mathbf{W}}\mathbb{P}(V_j|U_j)$ are shared between $\mathbb{P}_\delta$ and $\mathbb{P}$, their supports agree.

To get the density ratio, we take the ratio of $\mathbb{P}_\delta(\mathbf{V})$ and $\mathbb{P}(\mathbf{V})$, and the terms $V_j \in \mathbf{V}\setminus\mathbf{W}$ cancel:

$$\begin{aligned} w_\delta(\mathbf{V}) &= \frac{\mathbb{P}_\delta(\mathbf{V})}{\mathbb{P}(\mathbf{V})} \\ &= \prod_{i=1}^m \frac{\mathbb{P}_{\delta_i}(W_i|Z_i)}{\mathbb{P}(W_i|Z_i)}. \end{aligned}$$

By Definition 6.4 and Assumption 6.1, each $\mathbb{P}_{\delta_i}(W_i|Z_i)$ is a δ_i-perturbation around the CEF distribution $\mathbb{P}(W_i|Z_i)$, so plugging in the exponential family densities, we get

$$
\begin{aligned}
w_\delta(\mathbf{V}) &= \prod_{i=1}^{m} \frac{g(W_i)\exp\Big(\{\eta_i(Z_i)+s_i(Z_i;\delta_i)\}^\top T_i(W_i) - h_i(\eta_i(Z_i)+s_i(Z_i;\delta_i))\Big)}{g(W_i)\exp\Big(\eta_i(Z_i)^\top T_i(W_i) - h_i(\eta_i(Z_i))\Big)} \\
&= \prod_{i=1}^{m} \exp\Big(s_i(Z_i;\delta_i)T_i(W_i) - h_i(\eta_i(Z_i)+s_i(Z_i;\delta_i)) + h_i(\eta_i(Z_i))\Big) \\
&= \exp\Big(\sum_{i=1}^{m} s_i(Z_i;\delta_i)T_i(W_i)\Big)\exp\Big(\sum_{i=1}^{m} h_i(\eta_i(Z_i)) - h_i(\eta_i(Z_i)+s_i(Z_i;\delta_i))\Big).
\end{aligned}
$$

□

D.7.2 Proof of Theorem 6.1

Theorem 6.1 (Shift gradients and Hessians as covariances)**.** *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy Definition 6.4, with intervened variables $\mathbf{W} = \{W_1, \dots, W_m\}$ and shift functions $s_i(Z_i;\delta_i)$, where $\delta = (\delta_1, \dots, \delta_m)$. Then the shift gradient is given by $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \dots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_\delta}$ where*

$$
\mathrm{SG}^1_i = \mathbb{E}\left[D_{i,1}^\top\, Cov\Big(\ell,\, T_i(W_i)\Big| Z_i\Big)\right],
$$

and the shift Hessian is a matrix of size $(d_\delta \times d_\delta)$, where the (i,j)th block of size $d_{\delta_i} \times d_{\delta_j}$ equals

$$
\{\mathrm{SG}^2\}_{i,j} = \begin{cases} \mathbb{E}\left[D_{i,1}^\top Cov\Big(\ell,\, \epsilon_{T_i|Z_i}\epsilon_{T_i|Z_i}^\top | Z_i\Big)\, D_{i,1}\right] - \mathbb{E}\left[\ell \cdot D_{i,2}^\top \epsilon_{T|Z}\right] & i = j \\ Cov(\ell,\; D_{i,1}^\top \epsilon_{T_i|Z_i}\epsilon_{T_j|Z_j}^\top D_{j,1}) & i \neq j, \end{cases}
$$

where $D_{i,k} := \nabla^k_{\delta_i} s_i(Z_i;\delta_i)|_{\delta=0}$, is the gradient of the shift function for $k = 1$, and the Hessian for $k = 2$. Here, $T_i(W_i)$ is the sufficient statistic of $\mathbb{P}(W_i|Z_i)$ and $\epsilon_{T_i|Z_i} := T_i(W_i) - \mathbb{E}[T(W_i)|Z_i]$.

Proof. For simplicity throughout, we use $h_i^{(1)}$ to denote the gradient of the log-partition function $\nabla h_i(\cdot)$ with respect to the arguments, which is a column vector of length d_{T_i}, and we use $h_i^{(2)}$ to denote the Hessian $\nabla^2 h_i(\cdot)$, which is a matrix of size $d_{T_i} \times d_{T_i}$. We also use $\eta_{\delta_i}(z_i)$ as short-hand for $\eta_i(z_i) + s_i(z_i; \delta_i)$.

Shift Gradient: By Definition 6.4, the probability density / mass function $\mathbb{P}_\delta$ factorizes as follows, where $\delta = (\delta_1, \dots, \delta_m)$

$$\mathbb{P}_\delta(\mathbf{V}) = \left(\prod_{W_i \in \mathbf{W}} \mathbb{P}_{\delta_i}(W_i | Z_i) \right) \left(\prod_{V_i \in \mathbf{V} \setminus \mathbf{W}} \mathbb{P}(V_i | \operatorname{PA}(V_i)) \right), \tag{D.9}$$

and the gradient with respect to shift parameters δ_i is given by

$$\nabla_{\delta_i} p_\delta(v) = p_\delta(v) \nabla_{\delta_i} \log p_\delta(v) = p_\delta(v) \nabla_{\delta_i} \log p_{\delta_i}(w_i | z_i)$$

where the last equality follows from additivity of the log-likelihood in the conditionals, the factorization above, and the fact that δ_i only enters into the given conditional distribution. Given the assumed form of $\log p_{\delta_i}(w_i|z_i)$ given in Definition 6.3, we can observe that

$$\begin{aligned}
\nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i) &= \nabla_{\delta_i} \left[(\eta_i(z_i) + s_i(z_i; \delta_i))^\top T_i(w_i) - h_i(\eta(z_i) + s_i(z_i; \delta_i)) \right] \\
&= (\nabla_{\delta_i} s_i(z_i; \delta_i))^\top T_i(w_i) - (\nabla_{\delta_i} s_i(z_i; \delta_i))^\top \nabla h_i(\eta(z_i) + s_i(z_i; \delta_i)) \\
&= (\nabla_{\delta_i} s_i(z_i; \delta_i))^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \qquad \text{(D.10)}
\end{aligned}$$

where $\nabla_{\delta_i} s_i(z_i; \delta_i) \in \mathbb{R}^{d_{T_i} \times d_{\delta_i}}$, and $\nabla h_i(\eta(z_i) + s_i(z_i; \delta_i))$ is the gradient of the function $h_i : \mathbb{R}^{d_{T_i}} \to \mathbb{R}$, which is a column vector of length d_{T_i}. It follows from known properties of the log-partition function (Wainwright et al., 2008, Proposition 3.1), that $h_i^{(1)}(\eta_{\delta_i}(z_i)) = \mathbb{E}_\delta[T_i(W_i)|z_i]$. This gives us that

$$\begin{aligned}
\nabla_{\delta_i} \mathbb{E}_\delta[\ell] &= \mathbb{E}_\delta \left[\ell \cdot (\nabla_{\delta_i} s_i(Z_i; \delta_i))^\top (T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]) \right] \\
&= \mathbb{E}_\delta \left[(\nabla_{\delta_i} s_i(Z_i; \delta_i))^\top \mathbb{E}_\delta[\ell \cdot (T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]) | Z_i] \right] \\
&= \mathbb{E}_\delta \left[(\nabla_{\delta_i} s_i(Z_i; \delta_i))^\top \operatorname{Cov}_\delta(\ell, T_i(W_i)|Z_i) \right],
\end{aligned}$$

where the second equality follows from the tower property and Z_i-measurability of $\nabla_{\delta_i} s_i(Z_i; \delta_i)$, and the final equality follows from the definition of the conditional covariance. This expression, evaluated at $\delta = 0$, gives us the desired result, that

$$\mathrm{SG}^1_i := \nabla_{\delta_i} \mathbb{E}_\delta[\ell]\big|_{\delta=0} = \mathbb{E}\left[D_{i,1}^\top \mathrm{Cov}(\ell, T_i(W_i)|Z_i)\right],$$

where $D_{i,1} = \nabla_{\delta_i} s_i(Z_i, \delta_i)|_{\delta=0}$. The result follows from the definition that gradients are taken entry-wise, giving $\mathrm{SG}^1 = (\mathrm{SG}^1_1, \ldots, \mathrm{SG}^1_m) \in \mathbb{R}^{d_{\delta_1} + \cdots d_{\delta_m}}$.

Shift Hessian (Diagonal): For the shift Hessian, we first compute the diagonal entries of $\nabla^2_\delta \mathbb{E}_\delta[\ell]|_{\delta=0}$, which are blocks of size $\mathbb{R}^{d_{\delta_i} \times d_{\delta_i}}$. We begin by computing the Hessian of the likelihood.

$$\begin{aligned}
&\nabla^2_{\delta_i} p_\delta(v) \\
&= \nabla_{\delta_i}\Bigg(p_\delta(v) \nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i)\Bigg) \\
&= p_\delta(v)\Bigg((\nabla_{\delta_i} \log p_{\delta_i}(w_i|z_i))^{\otimes 2} + \nabla^2_{\delta_i} \log p_{\delta_i}(w_i|z_i)\Bigg) \\
&= p_\delta(v)\Bigg(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)^{\otimes 2} \{\nabla_{\delta_i} s_i(z_i;\delta_i)\} \\
&\qquad\qquad - \{\nabla^2_{\delta_i} s_i(z_i;\delta_i)\}^\top (T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))) \\
&\qquad\qquad - \{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top h_i^{(2)}(\eta_{\delta_i}(z_i)) \{\nabla_{\delta_i} s_i(z_i;\delta_i)\}\Bigg), \\
&= p_\delta(v)\Bigg(\{\nabla_{\delta_i} s_i(z_i;\delta_i)\}^\top \Bigg(\big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)^{\otimes 2} - h_i^{(2)}(\eta_{\delta_i}(z_i))\Bigg) \{\nabla_{\delta_i} s_i(z_i;\delta_i)\} \\
&\qquad\qquad - \{\nabla^2_{\delta_i} s_i(z_i;\delta_i)\}^\top \big(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i))\big)\Bigg)
\end{aligned}$$

where we use the notation $v^{\otimes 2} := vv^\top$, and we note that $\nabla^2_{\delta_i} s(z_i; \delta_i)$ is a tensor of size $d_{T_i} \times d_{\delta_i} \times d_{\delta_i}$, and $\{\nabla^2_{\delta_i} s_i(z_i; \delta_i)\}^\top h_i^{(1)}(\cdot)$ is a matrix of size $d_{\delta_i} \times d_{\delta_i}$, where the (m, n)'th entry is $\{\frac{\partial}{\partial \delta_{im}} \frac{\partial}{\partial \delta_{in}} s(z_i; \delta_i)\}^\top h^{(1)}(\cdot)$.

Now, using the fact that $h^{(1)}(\eta_{\delta_i}(z)) = \mathbb{E}_\delta[T_i(W_i)|z_i]$ and $h^{(2)}(\eta_{\delta_i}(z_i)) = \mathrm{Var}_\delta[T_i(W_i)|z_i]$ (Wainwright et al., 2008, Proposition 3.1), and the definition $\epsilon_{T_i|Z_i} = T_i(W_i) -$

$\mathbb{E}_\delta[T_i(W_i)|Z_i]$, we obtain

$$
\begin{aligned}
&\nabla^2_{\delta_i}\mathbb{E}_\delta[\ell] \\
&= \mathbb{E}_\delta\left[\ell\cdot\{\nabla_{\delta_i}s_i(Z_i;\delta_i)\}^\top\left(\epsilon^{\otimes 2}_{T|Z_i} - \mathrm{Var}_\delta(T_i(W_i)|Z_i)\right)\{\nabla_{\delta_i}s_i(Z_i;\delta_i)\}\right] \\
&\quad - \mathbb{E}_\delta\left[\ell\cdot\{\nabla^2_{\delta_i}s_i(Z_i;\delta_i)\}^\top\epsilon_{T_i|Z_i}\right] \\
&= \mathbb{E}_\delta\left[\{\nabla_{\delta_i}s_i(Z_i;\delta_i)\}^\top\mathrm{Cov}_\delta\left(\ell,\epsilon^{\otimes 2}_{T_i|Z_i}\middle|Z_i\right)\{\nabla_{\delta_i}s_i(Z_i;\delta_i)\}\right] \\
&\qquad\qquad - \mathbb{E}_\delta\left[\ell\cdot\{\nabla^2_{\delta_i}s_i(Z_i;\delta_i)\}^\top\epsilon_{T_i|Z_i}\right]
\end{aligned}
$$

which gives the desired result when we evaluate at $\delta = 0$.

Shift Hessian (Off-Diagonal) For $i \neq j$, we have that

$$
\begin{aligned}
&\nabla_{\delta_i}\nabla_{\delta_j}p_\delta(v) \\
&= \nabla_{\delta_i}(p_\delta(v)\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) \\
&= \nabla_{\delta_i}(p_\delta(v)\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) \\
&= p_\delta(v)\nabla_{\delta_i}\log p_{\delta_i}(w_i|z_i)\big(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)\big)^\top \\
&= p_\delta(v)\left(\{\nabla_{\delta_i}s_i(z_i;\delta_i)\}^\top(T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i)))\right) \\
&\qquad\qquad\left(\{\nabla_{\delta_j}s_j(z_j;\delta_j)\}^\top(T_j(w_j) - h_j^{(1)}(\eta_{\delta_j}(z_j)))\right)^\top
\end{aligned}
$$

where the third line follows from the fact that $\nabla_{\delta_i}(\nabla_{\delta_j}\log p_{\delta_j}(w_j|z_j)) = 0$, and the last line follows from the derivation of the gradient of the log-likelihood in Equation (D.10). We can again use the fact that $h_i^{(1)}(\eta_{\delta_i}(Z_i)) = \mathbb{E}_\delta[T_i(W_i)|Z_i]$ and the shorthand $\epsilon_{T_i|Z_i} \coloneqq T_i(W_i) - \mathbb{E}_\delta[T_i(W_i)|Z_i]$ to write that

$$
\begin{aligned}
&\nabla_{\delta_i}\nabla_{\delta_j}\mathbb{E}_\delta[\ell] \\
&= \mathbb{E}_\delta\Bigg[\ell\cdot\{\nabla_{\delta_i}s_i(z_i;\delta_i)\}^\top\left((T_i(w_i) - h_i^{(1)}(\eta_{\delta_i}(z_i)))\right) \\
&\qquad\qquad\qquad\left((T_j(w_j) - h_j^{(1)}(\eta_{\delta_j}(z_j)))\right)^\top\{\nabla_{\delta_j}s_j(z_j;\delta_j)\}\Bigg]
\end{aligned}
$$

and when we evaluate this expression at $\delta = 0$, we obtain

$$\nabla_{\delta_i} \nabla_{\delta_j} \mathbb{E}_\delta[\ell]\big|_{\delta=0} = \mathbb{E}\left[\ell \cdot D_{i,1}^\top \epsilon_{T_i|Z_i} (\epsilon_{T_j|Z_j})^\top D_{j,1}\right] = \mathrm{Cov}(\ell, D_{i,1}^\top \epsilon_{T_i|Z_i} \epsilon_{T_j|Z_j}^\top D_{j,1}).$$

Where the last equality follows because $\mathbb{E}[D_{i,1}^\top \epsilon_{T_i|Z_i} \epsilon_{T_j|Z_j}^\top D_{j,i}] = 0$. To see this, note that one of W_i, W_j must be a non-descendant of the other, and we will assume without loss of generality that W_j is a non-descendant of W_i in the causal graph consistent with the factorization given in Equation (D.9), which implies that Z_j (the parents of W_j in the underlying graph) are also non-descendants of W_i.Thus, $W_i \perp\!\!\!\perp (W_j, Z_j)|Z_i$, because (W_j, Z_j) are both non-descendants of W_i. Then, observe that $D_{i,1}$ is a function of Z_i, and $\epsilon_{T_i|Z_i}$ is a variable with zero-mean conditioned on Z_i. Thus, $\mathbb{E}[D_{i,1}^\top \epsilon_{T_i|Z_i}|Z_i] = 0$, for all Z_i. Moreover, given Z_i, we have that $D_{i,1}^\top \epsilon_{T_i|Z_i}$ is independent of $D_{j,1}^\top \epsilon_{T_j|Z_j}$. As a result, we can write that

$$\begin{aligned}
\mathbb{E}[D_{i,1}^\top \epsilon_{T_i|Z_i} \epsilon_{T_j|Z_j}^\top D_{j,1}] &= \mathbb{E}[\mathbb{E}[D_{i,1}^\top \epsilon_{T_i|Z_i} \epsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]] \\
&= \mathbb{E}[\mathbb{E}[D_{i,1}^\top \epsilon_{T_i|Z_i}|Z_i]\mathbb{E}[\epsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]] \\
&= \mathbb{E}[0 \cdot \mathbb{E}[\epsilon_{T_j|Z_j}^\top D_{j,1}|Z_i]] \\
&= 0
\end{aligned}$$

□

D.7.3 Proof of Corollary 6.1

Corollary 6.1 (Simple shift in a single variable)**.** *Assume the setup of Theorem 6.1, restricted to a shift in a single variable W, and that $s(Z;\delta) = \delta$. Then $D_1 = 1$, $D_2 = 0$, and*

$$\mathrm{SG}^1 = \mathbb{E}\left[Cov\Big(\ell, T(W)\Big| Z\Big)\right] \qquad \textit{and} \qquad \mathrm{SG}^2 = \mathbb{E}\left[Cov\Big(\ell, \epsilon_{T|Z}\epsilon_{T|Z}^\top \Big| Z\Big)\right],$$

where $T(W)$ is the sufficient statistic of W and $\epsilon_{T|Z} := T(W) - \mathbb{E}[T(W)|Z]$.

Proof. We have $\nabla_\delta s(Z;\delta) = \nabla_\delta \delta = 1$ and $\nabla^2_\delta s(Z;\delta) = \nabla^2_\delta \delta = 0$. The result now follows from Theorem 6.1. □

D.7.4 Proof of Theorem 6.2

Theorem 6.2. *Assume that $\mathbb{P}_\delta, \mathbb{P}$ satisfy the conditions of Theorem 6.1, with a shift in a single variable W, where $s(Z;\delta) = \delta$. Let $E_{\delta,Taylor}$ be the population Taylor estimate (Equation (6.7)) and let $\sigma(M)$ denote the largest absolute value of the eigenvalues of a matrix M. Then*

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,Taylor}\right| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \sigma\left(Cov_{t\cdot\delta}(\ell, \epsilon_{t\cdot\delta,T|Z}\epsilon^\top_{t\cdot\delta,T|Z}) - Cov(\ell, \epsilon_{0,T|Z}\epsilon^\top_{0,T|Z})\right) \cdot \|\delta\|^2,$$

where $T(W)$ is the sufficient statistic of $W|Z$ and $\epsilon_{t\cdot\delta,T|Z} = T(W|Z) - \mathbb{E}_{t\cdot\delta}[T(W|Z)]$.

Proof. The expectation is continuous and twice-differentiable with respect to δ, because of the smoothness of the exponential family in the parameter, the fact that the shift function s is twice-differentiable, and because the support does not change. Thus, applying Taylors remainder theorem to the function $t \mapsto \mathbb{E}_{t\cdot\delta}[\ell]$, it follows that there exist a $t_0 \in [0,1]$ such that

$$\mathbb{E}_{1\cdot\delta}[\ell] - \mathbb{E}_{0\cdot\delta}[\ell] - \left(\tfrac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \left(\tfrac{1}{2}\tfrac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0}. \tag{D.11}$$

We have $\left(\tfrac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=0} = \mathrm{SG}^1$ and by the same arguments (see the proof of Theorem 6.1), it follows that $\left(\tfrac{1}{2}\tfrac{\mathrm{d}^2}{\mathrm{d}^2 t}\mathbb{E}_{t\cdot\delta}[\ell]\right)\Big|_{t=t_0} = \delta^\top \mathrm{Cov}_{t_0\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t_0\cdot\delta,T|Z})\delta$. Plugging this in, and subtracting $\frac{1}{2}\delta^\top\,\mathrm{SG}^2\,\delta$ on both sides of Equation (D.11) yields

$$\begin{aligned}\left|\mathbb{E}_\delta[\ell] - E_{\delta,\text{Taylor}}\right| &= \tfrac{1}{2}\left|\delta^\top\left(\mathrm{Cov}_{t_0\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t_0\cdot\delta,T|Z}) - \mathrm{Cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)\delta\right| \\ &\leq \tfrac{1}{2}\sup_{t\in[0,1]}\left|\delta^\top\left(\mathrm{Cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \mathrm{Cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)\delta\right|.\end{aligned}$$

Let $K := \left(\mathrm{Cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \mathrm{Cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)$. Since K is symmetric and real valued, it is diagonalizeable, $K = U^\top \Lambda U$ for an orthonormal matrix U and diagonal matrix $\Lambda = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$. We then have

$$\begin{aligned}
|\delta^\top K \delta| &= |\delta^\top U^\top \Lambda U \delta| \\
&= |(\Lambda^{1/2} U\delta)^\top (\Lambda^{1/2} U\delta)| \\
&= \|\Lambda^{1/2} U\delta\|_2^2 \\
&\leq \|\Lambda^{1/2}\|_2^2 \|U\delta\|_2^2 \\
&= \sigma(K)\|\delta\|_2^2,
\end{aligned}$$

where $\Lambda^{1/2} = \mathrm{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_d})$, $\|\cdot\|_2$ denotes the supremum-norm when applied to matrices and the 2-norm when applied to vectors and $\|U\delta\|_2 = \|\delta\|_2$ because $\|U\delta\|_2^2 = \delta^\top U^\top U \delta = \delta^\top \delta = \|\delta\|_2^2$, using orthonormality of U. Plugging in this inequality, we get that

$$\left|\mathbb{E}_\delta[\ell] - E_{\delta,\mathrm{Taylor}}\right| \leq \tfrac{1}{2} \sup_{t\in[0,1]} \sigma\left(\mathrm{Cov}_{t\cdot\delta}(\ell, \epsilon^{\otimes 2}_{t\cdot\delta,T|Z}) - \mathrm{Cov}(\ell, \epsilon^{\otimes 2}_{0,T|Z})\right)\|\delta\|_2^2,$$

which concludes the proof. □

D.7.5 Proof of Proposition D.2.1

Proposition D.2.1. *Consider a binary random variable W with conditional distribution*

$$\mathbb{P}_\delta(W = 1|Z) = \sigma(\eta(Z) + \delta)$$

for an arbitrary measurable function $\eta(Z)$ whose range is the extended real numbers $\eta(Z) \in \mathbb{R} \cup \{+\infty, -\infty\}$. Let $p_+ := \mathbb{P}(\eta(Z) = +\infty)$, $p_- := \mathbb{P}(\eta(Z) = -\infty)$, and assume that $p_+ + p_- < 1$. Then, the marginal probability

$$p_\delta = \mathbb{P}_\delta(W = 1)$$

is a strictly monotonically increasing function of $\delta \in \mathbb{R}$ whose range is $(p_+, 1 - p_-)$,

Proof. Let F denote the event that $\eta(Z)$ is finite (i.e., $\eta(Z) \notin \{-\infty, +\infty\}$). Under F, the conditional probability function $\sigma(\eta(Z) + \delta)$ is a strictly monotonically increasing function of δ, and if $\eta(Z) \in \{-\infty, +\infty\}$, then the conditional probability is a constant function of δ (zero or one, respectively). Hence, we can write that

$$\mathbb{P}_\delta(W = 1) = \mathbb{P}_\delta(W = 1|F)(1 - p_+ - p_-) + p_+$$

and by assumption, $1 - p_+ - p_- > 0$. The marginal probability $\mathbb{P}_\delta(W = 1|F)$ is a strictly monotonically increasing function of δ, with a limit of 1 as $\delta \to \infty$, and a limit of 0 as $\delta \to -\infty$. As a result, it is bounded in $(p_+, 1 - p_-)$. □

D.7.6 Proof of Lemma D.3.1

Lemma D.3.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (D.2). For $\gamma \in \mathbb{R}^{d_X}$ define $\ell := (Y - \gamma^\top X)^2$. Then there exist $v_\gamma, u_{\mu,\gamma} \in \mathbb{R}^{d_A}$ such that for all shifts $\delta \in \mathbb{R}^{d_A}$:*

$$\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top u_{\mu,\gamma} + \tfrac{1}{2}\delta^\top v_\gamma v_\gamma^\top \delta,$$

where $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$. Further $u_{\mu,\gamma} = 0$ if $\mu = 0$.

Proof. It follows from Equation (D.2) that one can write $(X^\top, Y^\top, H^\top) = (1 - B)^{-1}(MA + \epsilon)$, and for a given γ, there exist b_γ, κ_γ such that $Y - \gamma^\top X = b_\gamma^\top A + \kappa_\gamma^\top \epsilon$ (Rothenhäusler et al., 2021). In $\mathbb{P}_\delta$, we can write $A = \mu + \delta + \epsilon_A$, where $\epsilon_A \sim \mathcal{N}(0, \Sigma)$, for all values of μ and δ. Plugging this in yields

$$\begin{aligned}\mathbb{E}_\delta[(Y - \gamma^\top X)^2] &= \mathbb{E}_\delta[(b_\gamma^\top A + \kappa_\gamma^\top \epsilon)^2] \\ &= \mathbb{E}_\delta[(b_\gamma^\top(\mu + \delta + \epsilon_A) + \kappa_\gamma^\top \epsilon)^2]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[(b_\gamma^\top(\mu + \epsilon_A) + \kappa^\top \epsilon)^2] + (2b_\gamma^\top \mu)\delta^\top b_\gamma + \delta^\top b_\gamma b_\gamma^\top \delta \\
&= \mathbb{E}[(Y - \gamma^\top X)^2] + (2b_\gamma^\top \mu)\delta^\top b_\gamma + \delta^\top b_\gamma b_\gamma^\top \delta.
\end{aligned}
$$

where we do not put a subscript on the expectation in the third line because it is taking expectations over ϵ_A and ϵ, both which do not depend on the choice of μ and δ. The statement of the lemma follows by letting $u_{\mu,\gamma} = 2b_\gamma^\top \mu$ and $v_\gamma = \sqrt{2} b_\gamma$. □

D.7.7 Proof of Proposition D.3.1

Proposition D.3.1. *Suppose $A \sim \mathcal{N}(\mu, \Sigma)$ and that (X, Y, H) are generated according to Equation (D.2). Then the shift gradient and Hessian are given by*

$$
\mathrm{SG}^1 = Cov(\ell, \Sigma^{-1} A) \qquad \textit{and} \qquad \mathrm{SG}^2 = Cov(\ell, \Sigma^{-1}(A-\mu)(A-\mu)^\top \Sigma^{-\top})
$$

and the loss under a mean shift of δ in A is given by

$$
\mathbb{E}_\delta[\ell] = \mathbb{E}[\ell] + \delta^\top \mathrm{SG}^1 + \tfrac{1}{2}\delta^\top \mathrm{SG}^2 \delta,
$$

where $\ell := (Y - \gamma^\top X)^2$ and $\mathbb{E}_\delta$ corresponds to taking the mean in the distribution where $A \sim \mathcal{N}(\mu + \delta, \Sigma)$.

Proof. Similar to Lemma D.3.1, we rewrite $Y - \gamma^\top X = b_\gamma^\top A + \kappa^\top \epsilon$, and by rewriting $A = \mu + \delta + \epsilon_A$, where $\epsilon_A \sim \mathcal{N}(0, \Sigma)$, we obtain

$$
\begin{aligned}
\mathbb{E}_\delta[(Y - \gamma^\top X)^2] &= \mathbb{E}(b_\gamma^\top(\mu + \epsilon_A) + \kappa^\top \epsilon)^2 && \text{(D.12)} \\
&\quad + (2b_\gamma^\top \mu)\delta^\top b && \text{(D.13)} \\
&\quad + \delta^\top b b^\top \delta. && \text{(D.14)}
\end{aligned}
$$

We recognize that Equation (D.12) equals $\mathbb{E}(Y - \gamma^\top X)^2$. Similarly, we now show that Equations (D.13) and (D.14) match the shift gradients (multiplied appropriately with δ).

First, we assume that $\Sigma = \mathrm{Id}$. Since A is a Gaussian with (known) mean Id, the sufficient statistic is $T(A) = A$. Hence, according to Theorem 6.1, we can compute the shift gradient as

$$\mathrm{SG}^1 = \mathrm{Cov}(A, \ell) = \mathrm{Cov}(A, (Y - \gamma^\top X)^2) = \mathrm{Cov}(A, (b_\gamma^\top A)^2).$$

We can calculate the i'th entrance of this vector as:

$$\begin{aligned}\mathrm{SG}^1 = \mathrm{Cov}(A_i, (b_\gamma^\top A)^2) &= \mathrm{Cov}(A_i - \mu_i, (b_\gamma^\top A)^2)) \\ &= \mathrm{Cov}(A_i - \mu_i, b_{\gamma,i}^2 A_i^2 + 2\sum_{j \neq i} b_i b_j A_i A_j) \\ &= b_{\gamma,i}^2 \mathrm{Cov}(A_i - \mu_i, A_i^2) + 2b_{\gamma,i}\sum_{j\neq i} b_j \mathrm{Cov}(A_i - \mu_i, A_i A_j),\end{aligned}$$

where in the first equality we use that subtracting a constant doesn't change the covariance, and we use independence of A_i from $A_j A_{j'}$ when $i \notin \{j, j'\}$. Using the assumption that A_i has unit variance, we now get that

$$\begin{aligned}\mathrm{Cov}(A_i - \mu_i, A_i^2) = \mathbb{E}[A_i^3 - \mu_i A_i^2] = (\mu_i^3 + 3\mu_i) - \mu_i(\mu_i^2 + 1) = 2\mu_i \\ \mathrm{Cov}(A_i - \mu_i, A_i A_j) = \mathbb{E}[A_i^2 - A_i\mu_i]\mathbb{E}[A_j] = (\mu_i^2 + 1 - \mu_i^2)\mu_j = \mu_j.\end{aligned}$$

By plugging in, we obtain

$$\begin{aligned}\mathrm{SG}^1(\mu_i) &= 2b_{\gamma,i}^2\mu_i + 2b_{\gamma,i}\sum_{j\neq i} b_j \mu_j \\ &= 2b_{\gamma,i} b_\gamma^\top \mu.\end{aligned}$$

Since this was element-wise, we obtain that the full vector is $\mathrm{SG}^1 = 2b_\gamma b_\gamma^\top \mu$, which, when multiplied with δ yields Equation (D.13).

We compute SG^2 similarly. The diagonal entries are given by

$$\mathrm{SG}_{i,i}^2 = \mathrm{Cov}((A_i - \mu_i)^2, (b_\gamma^\top A)^2)$$

$$
\begin{aligned}
&= \operatorname{Cov}((A_i - \mu_i)^2, b_{\gamma,i}^2 A_i^2 + b_{\gamma,i} \sum_{j \neq i} b_{\gamma,j} A_i A_j) \\
&= b_{\gamma,i}^2 \operatorname{Cov}((A_i - \mu_i)^2, A_i^2) + b_{\gamma,i} \sum_{j \neq i} b_{\gamma,j} \operatorname{Cov}((A_i - \mu_i)^2, A_i A_j).
\end{aligned}
$$

Because $\Sigma = \mathrm{Id}$, the second through fourth moments of A_i are given by $\mathbb{E}[A_i^2] = \mu_i^2 + 1$, $\mathbb{E}[A_i^3] = \mu_i^3 + 3\mu_i$ and $\mathbb{E}[A_i^4] = \mu_i^4 + 6\mu_i^2 + 3$. Using this, we get

$$
\begin{aligned}
\operatorname{Cov}((A_i - \mu_i)^2, A_i^2) &= \mathbb{E}[A_i^4 - 2\mu_i A_i^3 + \mu_i^2 A_i^2] - \mathbb{E}[(A_i - \mu_i)^2]\mathbb{E}[A_i^2] \\
&= (\mu_i^4 + 6\mu_i^2 + 3) - 2\mu_i(\mu_i^3 + 3\mu_i) + \mu_i^2(\mu_i^2 + 1) - 1 \cdot (\mu_i^2 + 1) \\
&= 2,
\end{aligned}
$$

and for $j \neq i$:

$$
\begin{aligned}
\operatorname{Cov}((A_i - \mu_i)^2, A_i A_j) &= \operatorname{Cov}((A_i - \mu_i)^2, (A_i - \mu_i) A_j) + \operatorname{Cov}((A_i - \mu_i)^2, \mu_i A_j) \\
&= \operatorname{Cov}((A_i - \mu_i)^2, (A_i - \mu_i) A_j) \\
&= \mathbb{E}[(A_i - \mu_i)^3]\mathbb{E}[A_j] - \mathbb{E}[(A_i - \mu_i)^2]\mathbb{E}[(A_i - \mu_i)]\mathbb{E}[A_j] \\
&= 0 - 0,
\end{aligned}
$$

using linearity of the covariance, that $A_i \perp\!\!\!\perp A_j$ and that the first and third moments are zero for a centered Gaussian $A_i - \mu_i$. Plugging this in, we get that the diagonal entries are given by

$$
\mathrm{SG}_{i,i}^2 = 2b_{\gamma,i}^2.
$$

We can compute the off-diagonal entries similarly. For $i \neq j$, we have:

$$
\begin{aligned}
\mathrm{SG}_{i,j}^2 = \operatorname{Cov}\Big(&(A_i - \mu_i)(A_j - \mu_j), \\
&b_{\gamma,i}^2 A_i^2 + b_{\gamma,j}^2 A_j^2 + 2b_{\gamma,i} b_{\gamma,j} A_i A_j + 2 \sum_{v \notin \{i,j\}} b_{\gamma,i} b_{\gamma,v} A_i A_v + b_{\gamma,j} b_{\gamma,v} A_j A_v\Big).
\end{aligned} \tag{D.15}
$$

Using the independence of A_i and A_j, we have

$$\begin{aligned}
&\mathrm{Cov}((A_i-\mu_i)(A_j-\mu_j),A_i^2)\\
&=\mathbb{E}[A_i^2(A_i-\mu_i)]\underbrace{\mathbb{E}[A_j-\mu_j]}_{=0}-\underbrace{\mathbb{E}[A_i-\mu_i]}_{=0}\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_i^2]\\
&=0,
\end{aligned}$$

and similarly $\mathrm{Cov}((A_i-\mu_i)(A_j-\mu_j),A_j^2)=0$. Using the same reasoning, for $v\notin\{i,j\}$

$$\begin{aligned}
&\mathrm{Cov}((A_i-\mu_i)(A_j-\mu_j),A_iA_v)\\
&=\mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_v]-\mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_i]\mathbb{E}[A_j-\mu_j]\mathbb{E}[A_v]\\
&=0,
\end{aligned}$$

and the same for $\mathrm{Cov}((A_i-\mu_i)(A_j-\mu_j),A_jA_v)$. Finally, we have

$$\begin{aligned}
&\mathrm{Cov}((A_i-\mu_i)(A_j-\mu_j),A_iA_j)\\
&=\mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[(A_j-\mu_j)A_j]-\mathbb{E}[(A_i-\mu_i)]\mathbb{E}[A_i]\mathbb{E}[(A_j-\mu_j)]\mathbb{E}[A_j]\\
&=\mathbb{E}[(A_i-\mu_i)A_i]\mathbb{E}[(A_j-\mu_j)A_j]\\
&=\mathbb{E}[A_i^2-\mu_iA_i]\mathbb{E}[A_j^2-\mu_jA_j]\\
&=[(\mu_i^2+1)-\mu_i^2][[(\mu_j^2+1)-\mu_j^2]]\\
&=1.
\end{aligned}$$

Plugging into Equation (D.15), we get that

$$\mathrm{SG}^2_{i,j}=2b_{\gamma,i}b_{\gamma,j},$$

and hence for both diagonal and off-diagonal entries, $\mathrm{SG}^2_{i,j}=2b_{\gamma,i}b_{\gamma,j}$, implying that

$$\mathrm{SG}^2=2b_\gamma b_\gamma^\top.$$

In particular $\frac{1}{2}\delta^\top\,\mathrm{SG}^2\,\delta$ matches Equation (D.14).

Finally, we consider the case $\Sigma \neq \mathrm{Id}$. Let $\Sigma^{-1/2}$ be the 'square-root' of Σ^{-1}, such that $\Sigma^{-1/2}\Sigma^{-\top/2}$ (where the latter denotes $(\Sigma^{-1/2})^\top$.[7]

The sufficient statistics for the mean in a multivariate Gaussian distribution with known variance is given by $T(A) = \Sigma^{-1}A$. We then have

$$\begin{aligned}
\mathrm{SG}^1 &= \mathrm{Cov}(\Sigma^{-1}A, (b_\gamma^\top A)^2) \\
&= \Sigma^{-1/2}\mathrm{Cov}(\Sigma^{-1/2}A, ((\Sigma^{1/2}b_\gamma)^\top \Sigma^{-1/2}A)^2) \\
&= \Sigma^{-1/2}\mathrm{Cov}_{\tilde{\mu}}(\tilde{A}, (\tilde{b}_\gamma^\top \tilde{A})^2),
\end{aligned}$$

where $\tilde{A} = \Sigma^{-1/2}A =\sim \mathcal{N}(\tilde{\mu}, \mathrm{Id})$, $\tilde{\mu} = \Sigma^{-1/2}\mu$ and $\tilde{b}_\gamma = \Sigma^{1/2}b_\gamma$. In particular, since $\tilde{A}$ has unit variance, we can use the above derivations to obtain

$$\mathrm{SG}^1 = 2\Sigma^{-1/2}(\tilde{b}_\gamma\tilde{b}_\gamma^\top\tilde{\mu}) = 2b_\gamma b_\gamma^\top \mu.$$

In particular, the first shift gradient is the when $\Sigma \neq \mathrm{Id}$ as when $\Sigma = \mathrm{Id}$. Similarly,

$$\begin{aligned}
\mathrm{SG}^2 &= \mathrm{Cov}(\Sigma^{-1}(A-\mu)(A-\mu)^\top\Sigma^{-\top}, (b_\gamma^\top A)^2) \\
&= \mathrm{Cov}(\Sigma^{-1/2}\Sigma^{-1/2}(A-\mu)(A-\mu)^\top\Sigma^{-\top/2}\Sigma^{-\top/2}, (\Sigma^{1/2}b_\gamma)^\top\Sigma^{-1/2}A)^2) \\
&= \Sigma^{-1/2}\mathrm{Cov}_{\tilde{\mu}}((\tilde{A}-\tilde{\mu})(\tilde{A}-\tilde{\mu})^\top, (\tilde{b}_\gamma^\top\tilde{A})^2)\Sigma^{-\top/2} \\
&= \Sigma^{-1/2}2\tilde{b}_\gamma\tilde{b}_\gamma^\top\Sigma^{-\top/2} \\
&= 2b_\gamma b_\gamma^\top.
\end{aligned}$$

Hence, also when $\Sigma \neq \mathrm{Id}$, the terms of Equations (D.13) and (D.14) matches the expression given by SG^1 and SG^2. This concludes the proof. □

[7] Formally, if $\Sigma^{-1} = U\Lambda U^\top$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d_A})$, define $\Sigma^{-1/2} := U\,\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{d_A}})$.

Appendix E

Appendix for Chapter 7

E.1 Details: Setup and Definition of Shifts

In the CelebA dataset, there are 40 distinct binary attributes for each image. To evaluate the sensitivity of models to distribution shift, we first define a factorization over the entire conditional distribution, based on our own categorization of these attributes. We consider a factorization of the joint distribution as follows, where X is the image, and A is the set of all attributes (including the label Y). We use red to indicate portions of the factorization which change.

$$P(A, X) = P(A)P(X \mid A), \tag{E.1}$$

where P denotes the probability mass function for discrete variables, and the probability density function for continuous variables. We consider a **structured shift** in the distribution of $P(A)$, which we further factorize as follows, without assuming any conditional independences, and where $A_{1:i-1} := (A_1, \dots, A_{i-1})$, and where we adopt the convention that $A_{1:0} = \varnothing$

$$P(A) = \prod_{i=1}^{d_A} P(A_i \mid A_{1:i-1}). \tag{E.2}$$

For each conditional distribution $P(A_i \mid A_{1:i})$, we define a set of possible conditional distribution indexed by a parameter $\delta \in \mathbb{R}^{d_A}$, where

$$P_\delta(A_i = 1 \mid A_{1:i}) = \sigma(\eta_i(A_{1:i}) + \delta_i). \tag{E.3}$$

where $\eta_i(A_{1:i-1})$ denotes the conditional log-odds $\log(p/(1-p))$ for $p := P(A_i = 1 \mid A_{1:i-1})$. This in turn defines a set of joint distributions

$$P_\delta(A, X) = P_\delta(A) P(X \mid A) \tag{E.4}$$

where $P_\delta(A)$ is equal to the product of the shifted distributions. For a fixed predictive model, we then seek to estimate the worst-case zero-one loss under a set of distributions defined by δ, recalling that Y is just one of the attributes

$$\sup_{\|\delta\|_2 \leq \lambda} \mathbb{E}_{X,Y \sim P_\delta}[f(X) \neq Y]. \tag{E.5}$$

Defining the order of the factorization: We partition the CelebA attributes into the following categories, and construct our factorization according to this order: Demographics (Gender, Age), Facial Features (e.g., Narrow Eyes), Facial Hair (e.g., Mustache), Non-Facial Hair (e.g., Baldness), Items Worn (e.g., Lipstick), Expression (e.g., Smiling). Within each category we also give an ordering of the attributes, which defines the full factorization. The full ordering of attributes is given in Table E.4.

Finding Candidate Shifts: Using the CelebA training set, we construct a Taylor approximation of the loss in Equation (E.5) as laid out in Thams et al. (2022), which requires the use of plug-in estimators for the conditional means $\mu_{A_i}(A_{1:i-1}) := \mathbb{E}[A_i \mid A_{1:i-1}]$ and $\mu_\ell(A_{1:i-1}) = \mathbb{E}[f(X) \neq Y \mid A_{1:i-1}]$.[1]

For both of these auxiliary functions, we use a gradient boosted classifier `XGBClassifier`

[1] We require one such estimator for each conditional distribution that is changing, but we note that this is unaffected by the complexity of the shift functions themselves. In this experiment, we consider simple shifts in the conditional log-odds of $\eta_i(\cdot) + \delta_i$, but for more complex shift functions $\eta_i(\cdot) + s(\cdot; \delta_i)$, the required estimators are the same.

from the `xgboost` python package, with hyperparameters (maximum tree depth in $\{2, 4, 6\}$) selected using cross-validation. Using these plug-in estimates, we solve for the maximizer of the quadratic approximation with respect to the training set

$$\hat{\delta^*} = \arg\max_{\|\delta\|_2 \leq \lambda} E_n[\ell(f(X), Y)] + \delta^\top \hat{\mathrm{SG}}_1 + \frac{1}{2}\delta^\top \hat{\mathrm{SG}}_2 \delta \tag{E.6}$$

where $\hat{\mathrm{SG}}_1, \hat{\mathrm{SG}}_2$ are the estimated gradient and Hessian of the loss $\mathbb{E}_{P_\delta}[\ell(f(X), Y)]$ with respect to δ, evaluated at $\delta = 0$, and $E_n[\ell]$ is the empirical loss on the training distribution.

Evaluating shifted performance We use the CelebA validation set for estimating shifted performance using importance sampling, for the candidate $\hat{\delta^*}$ that is found in the previous step. To construct these importance weights, we require estimates of the original conditional probabilities $P(A_i = 1 \mid A_{1:i})$. Here we fit an `XGBoost` classifier for each conditional probability, using cross-validation to select the maximum depth separately for each conditional distribution. These models are fit directly on the validation dataset, and are then used to construct weights on the same dataset.

Table E.1: *Candidate worst-case δ found for the original blond hair classification prompt, corresponding to Figure 7-3. Marginal proportions of each attribute before and after the shift are given based on the validation dataset, and estimated (via importance sampling on the validation dataset). Note that the importance sampling estimates are not exact: For instance, the marginal distribution of "No Beard" is estimated to be slightly over 100%.*

	Delta	Pre-shift	Post-shift	Difference
Male	-3.792	0.426	0.017	-0.409
Young	-1.401	0.747	0.565	-0.181
Bags Under Eyes	-0.749	0.207	0.074	-0.134
Big Lips	-0.159	0.153	0.157	0.004
Big Nose	-0.583	0.249	0.079	-0.170
Chubby	0.086	0.061	0.027	-0.034
Double Chin	0.027	0.049	0.019	-0.030
High Cheekbones	-0.648	0.449	0.446	-0.004
Narrow Eyes	-0.165	0.075	0.049	-0.026
Oval Face	0.060	0.280	0.293	0.013
Pale Skin	0.863	0.043	0.148	0.105
Pointy Nose	1.523	0.285	0.717	0.432
Rosy Cheeks	0.247	0.068	0.150	0.082
5 o Clock Shadow	-0.009	0.118	0.004	-0.114
Mustache	0.213	0.050	0.003	-0.048
Sideburns	0.074	0.069	0.003	-0.065
Goatee	0.073	0.074	0.004	-0.070
No Beard	-0.036	0.822	1.008	0.186
Bushy Eyebrows	-0.555	0.142	0.026	-0.117
Bald	0.148	0.021	0.001	-0.019
Receding Hairline	-0.290	0.072	0.027	-0.045
Bangs	0.773	0.147	0.409	0.263
Straight Hair	-0.423	0.206	0.104	-0.102
Blond Hair	-3.426	0.154	0.013	-0.140
Wearing Earrings	-0.937	0.191	0.206	0.015
Wearing Hat	-0.119	0.047	0.025	-0.023
Wearing Lipstick	0.672	0.446	0.844	0.398
Wearing Necklace	-0.158	0.121	0.187	0.066
Wearing Necktie	0.124	0.073	0.004	-0.069
Heavy Makeup	0.589	0.390	0.766	0.376
Eyeglasses	0.041	0.070	0.043	-0.027
Arched Eyebrows	-0.470	0.258	0.313	0.054
Mouth Slightly Open	0.434	0.482	0.588	0.106
Smiling	0.135	0.483	0.532	0.049

Table E.2: ***Restriction to shifts that do not change the causal mechanisms of the label****: Candidate worst-case δ found for the original blond hair classification prompt, where the shift is restricted to exclude direct interventions on hair related features. For these features, we write "0*" to indicate that the value of δ is effectively set to zero. That said, the proportions can still change, due to changes in the upstream features such as gender and age (e.g., there are more individuals with blond hair in the shifted distribution, likely due to the increased prevalence of women).*

	Delta	Pre-shift	Post-shift	Difference
Male	-4.144	0.426	0.012	-0.414
Young	-3.101	0.747	0.195	-0.551
Bags Under Eyes	-0.709	0.207	0.098	-0.109
Big Lips	-0.118	0.153	0.147	-0.006
Big Nose	-0.232	0.249	0.153	-0.096
Chubby	0.133	0.061	0.056	-0.005
Double Chin	0.164	0.049	0.042	-0.007
High Cheekbones	-1.680	0.449	0.287	-0.162
Narrow Eyes	-0.507	0.075	0.047	-0.028
Oval Face	-0.222	0.280	0.153	-0.127
Pale Skin	0.758	0.043	0.146	0.103
Pointy Nose	1.301	0.285	0.587	0.302
Rosy Cheeks	-0.282	0.068	0.081	0.012
5 o Clock Shadow	0.205	0.118	0.002	-0.116
Mustache	0.418	0.050	0.003	-0.048
Sideburns	0.461	0.069	0.004	-0.065
Goatee	0.443	0.074	0.004	-0.070
No Beard	-0.432	0.822	0.985	0.162
Bushy Eyebrows	-0.061	0.142	0.019	-0.124
Bald	0*	0.021	0.001	-0.020
Receding Hairline	0*	0.072	0.033	-0.039
Bangs	0*	0.147	0.284	0.137
Straight Hair	0*	0.206	0.110	-0.096
Blond Hair	0*	0.154	0.249	0.095
Wearing Earrings	-0.939	0.191	0.211	0.021
Wearing Hat	0.209	0.047	0.044	-0.003
Wearing Lipstick	0.282	0.446	0.677	0.231
Wearing Necklace	-0.167	0.121	0.210	0.089
Wearing Necktie	0.094	0.073	0.003	-0.070
Heavy Makeup	0.442	0.390	0.569	0.179
Eyeglasses	0.157	0.070	0.083	0.013
Arched Eyebrows	-0.755	0.258	0.266	0.008
Mouth Slightly Open	0.676	0.482	0.581	0.099
Smiling	-0.120	0.483	0.409	-0.074

Table E.3: ***Restriction to shifts that do not change the marginal distribution of the label****: Candidate worst-case δ found for the original blond hair classification prompt, where the shift is restricted to exclude direct interventions on hair related features, and where the hair related features come first, to avoid any change in their distribution. For these features, we write "0*" to indicate that the value of δ is effectively set to zero.*

	Delta	Pre-shift	Post-shift	Difference
Blond Hair	0*	0.154	0.145	-0.009
Bald	0*	0.021	0.013	-0.008
Receding Hairline	0*	0.072	0.059	-0.012
Bangs	2.219	0.147	0.615	0.469
Straight Hair	-1.070	0.206	0.084	-0.122
Male	-4.076	0.426	0.019	-0.406
Young	-2.323	0.747	0.277	-0.469
Bags Under Eyes	-0.594	0.207	0.114	-0.093
Big Lips	-0.596	0.153	0.120	-0.034
Big Nose	-0.629	0.249	0.112	-0.136
Chubby	-0.086	0.061	0.036	-0.025
Double Chin	0.084	0.049	0.028	-0.021
High Cheekbones	-0.356	0.449	0.600	0.151
Narrow Eyes	-0.235	0.075	0.052	-0.024
Oval Face	-0.110	0.280	0.220	-0.060
Pale Skin	0.733	0.043	0.111	0.068
Pointy Nose	1.999	0.285	0.781	0.496
Rosy Cheeks	0.208	0.068	0.169	0.101
5 o Clock Shadow	-0.094	0.118	0.003	-0.115
Mustache	0.224	0.050	0.003	-0.047
Sideburns	0.222	0.069	0.003	-0.066
Goatee	0.231	0.074	0.004	-0.070
No Beard	-0.181	0.822	1.006	0.183
Bushy Eyebrows	-0.614	0.142	0.014	-0.128
Wearing Earrings	-0.789	0.191	0.277	0.086
Wearing Hat	0.004	0.047	0.015	-0.032
Wearing Lipstick	0.430	0.446	0.824	0.378
Wearing Necklace	-0.075	0.121	0.267	0.146
Wearing Necktie	0.038	0.073	0.007	-0.066
Heavy Makeup	0.353	0.390	0.703	0.313
Eyeglasses	0.069	0.070	0.066	-0.003
Arched Eyebrows	-0.594	0.258	0.269	0.011
Mouth Slightly Open	0.403	0.482	0.659	0.178
Smiling	-0.065	0.483	0.637	0.154

Table E.4: *Blond Hair Classification: List of all attributes in the assumed causal order, with indications for which variables (i.e., conditional distributions) are allowed to shift. All variables are binary: We exclude the attributes "Black Hair", "Brown Hair", and "Grey Hair", using the attribute "Blond Hair" as a single binary attribute. We similarly exclude "Wavy Hair", treating "Straight Hair" as a binary attribute. We exclude the attribute "Attractive" from this list.*

Attribute Group	Attribute	Full Shift	Restricted Shift
N/A	Blurry		
Demographics	Male	*	*
Demographics	Young	*	*
Facial Features	Bags Under Eyes	*	*
Facial Features	Big Lips	*	*
Facial Features	Big Nose	*	*
Facial Features	Chubby	*	*
Facial Features	Double Chin	*	*
Facial Features	High Cheekbones	*	*
Facial Features	Narrow Eyes	*	*
Facial Features	Oval Face	*	*
Facial Features	Pale Skin	*	*
Facial Features	Pointy Nose	*	*
Facial Features	Rosy Cheeks	*	*
Facial Hair	5 o Clock Shadow	*	*
Facial Hair	Mustache	*	*
Facial Hair	Sideburns	*	*
Facial Hair	Goatee	*	*
Facial Hair	No Beard	*	*
Facial Hair	Bushy Eyebrows	*	*
Non-Facial Hair	Bald	*	
Non-Facial Hair	Receding Hairline	*	
Non-Facial Hair	Bangs	*	
Non-Facial Hair	Straight Hair	*	
Non-Facial Hair	Blond Hair	*	
Items Worn	Wearing Earrings	*	*
Items Worn	Wearing Hat	*	*
Items Worn	Wearing Lipstick	*	*
Items Worn	Wearing Necklace	*	*
Items Worn	Wearing Necktie	*	*
Items Worn	Heavy Makeup	*	*
Items Worn	Eyeglasses	*	*
Expression	Arched Eyebrows	*	*
Expression	Mouth Slightly Open	*	*
Expression	Smiling	*	*

Bibliography

Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., Klein, E. Y., Chen, E. S., Saheed, M. O., Henley, M., Miranda, S., Houston, K., Linton, R. C., Ahluwalia, A. R., Wu, A. W., and Saria, S. (2022). Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature medicine*, 28(7):1455–1460.

Adler-Milstein, J. and Jha, A. K. (2017). Hitech act drove large gains in hospital electronic health record adoption. *Health Affairs*, 36:1416–1422.

Agniel, D., Kohane, I. S., and Weber, G. M. (2018). Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361.

Aguirregabiria, V. and Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67.

Ahmadi-Javid, A. (2012). Entropic Value-at-Risk: A new coherent risk measure. *Journal of optimization theory and applications*, 155(3):1105–1123.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 35–44.

Anglemyer, A., Horvath, H. T., and Bero, L. (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane database of systematic reviews*.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv (1907.02893)*.

Athey, S., Imbens, G. W., and Wager, S. (2016). Approximate residual balancing: De-Biased inference of average treatment effects in high dimensions. *arXiv preprint (1604.07125)*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Bal, B. S. (2009). An introduction to medical malpractice in the United States. *Clinical Orthopaedics and Related Research*, 467(2):339–347.

Balke, A. and Pearl, J. (1994). Counterfactual Probabilities: Computational Methods, Bounds and Applications. *Proceedings of the Tenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-94)*, pages 46–54.

Banack, H. R., Kaufman, J. S., Wactawski-Wende, J., Troen, B. R., and Stovitz, S. D. (2019). Investigating and remediating selection bias in geriatrics research: the selection bias toolkit. *Journal of the American Geriatrics Society*, 67(9):1970–1976.

Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., Kuo, P.-C., Lungren, M. P., Palmer, L., Price, B. J., Purkayastha, S., Pyrros, A., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., Trivedi, H., Wang, R., Zaiman, Z., Zhang, H., and Gichoya, J. W. (2021). Reading race: AI recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*.

Bang, H. and Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61:962–972.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and others (2017). Network dissection: Quantifying interpretability of deep visual representations. *CVPR*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2011a). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Belloni, A., Chernozhukov, V., and Wang, L. (2011b). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Bellot, A. and van der Schaar, M. (2020). Accounting for Unobserved Confounding in Domain Generalization. *arXiv (2007.10653)*.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Benavidez, G. and Frakt, A. B. (2019). Fixing clinical practice guidelines.

Berthelot, D., Schumm, T., and Metz, L. (2017). BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*.

Bhattacharya, R., Nabi, R., and Shpitser, I. (2020). Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint (2003.12659)*.

Bibaut, A., Malenica, I., Vlassis, N., and Van Der Laan, M. (2019). More Efficient Off-Policy Evaluation through Regularized Targeted Learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 654–663, Long Beach, California, USA. PMLR.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2018). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.

Boominathan, S., Oberst, M., Zhou, H., Kanjilal, S., and Sontag, D. (2020). Treatment policy learning in multiobjective settings with fully observed outcomes. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 1937–1947, New York, NY, USA. ACM.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Chapter 59: Measurement Error In Survey Data. *Handbook of Econometrics*, 5:3705–3843.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brat, G. A., Agniel, D., Beam, A., Yorkgitis, B., Bicket, M., Homer, M., Fox, K. P., Knecht, D. B., McMahill-Walraven, C. N., Palmer, N., et al. (2018). Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study. *Bmj*, 360:j5790.

Brooks-Gunn, J., Liaw, F.-r., and Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359.

Buesing, L., Weber, T., Zwols, Y., Heess, N., Racaniere, S., Guez, A., and Lespiau, J.-B. (2019). Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. In *International Conference on Learning Representations*.

Bühlmann, P. and Ćevid, D. (2020). Deconfounding and causal regularisation for stability and external validity. *International Statistical Review*, 88(S1):S114–S134.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

CDC (2022). What is sepsis? https://www.cdc.gov/sepsis/what-is-sepsis.html. Accessed: 2023-06-07.

Chen, M., Goel, K., Sohoni, N. S., Poms, F., Fatahalian, K., and Re, C. (2021). Mandoline: Model evaluation under distribution shift. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1617–1629. PMLR.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The econometrics journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint (1712.04802)*.

Chow, Y., Petrik, M., and Ghavamzadeh, M. (2015). Robust Policy Optimization with Baseline Guarantees. *arXiv preprint*, pages 1–25.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. (2020). The Difficult Task of Distribution Generalization in Nonlinear Models. *arXiv*, pages 1–48.

Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada.

Cloyd, J. M., Heh, V., Pawlik, T. M., Ejaz, A., Dillhoff, M., Tsung, A., Williams, T., Abushahin, L., Bridges, J. F., and Santry, H. (2020). Neoadjuvant therapy for resectable and borderline resectable pancreatic cancer: a meta-analysis of randomized controlled trials. *Journal of clinical medicine*, 9(4):1129.

Conn, A. R., Gould, N. I., and Toint, P. L. (2000). *Trust region methods*. SIAM.

Corvelo Benz, N. L. and Gomez Rodriguez, M. (2022). Counterfactual inference of second opinions. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Cuellar, M. and Kennedy, E. H. (2018). A nonparametric projection-based estimator for the probability of causation, with application to water sanitation in Kenya. *Journal of the Royal Statistical Society: Series A (Special Issue on Causal Inference)*, pages 1–24.

Cummings, B. C., Ansari, S., Motyka, J. R., Wang, G., Medlin, R. P., Kronick, S. L., Singh, K., Park, P. K., Napolitano, L. M., Dickson, R. P., Mathis, M. R., Sjoding, M. W., Admon, A. J., Blank, R., McSparron, J. I., Ward, K. R., and Gillies, C. E. (2021). Predicting intensive care transfers and other unforeseen events: Analytic model validation study and comparison to existing methods. *JMIR medical informatics*, 9.

Dagan, N., Barda, N., Biron-Shental, T., Makov-Assif, M., Key, C., Kohane, I. S., Hernán, M. A., Lipsitch, M., Hernandez-Diaz, S., Reis, B. Y., and Balicer, R. D. (2021). Effectiveness of the BNT162b2 mRNA COVID-19 vaccine in pregnancy. *Nature medicine*, 27(10):1693–1695.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. (2019). Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *arXiv preprint (1908.09230).*

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2017). Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582.*

Dash, S., Gunluk, O., and Wei, D. (2018). Boolean decision rules via column generation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 4660–4670. Curran Associates, Inc.

Dawid, P., Faigman, D. L., and Fienberg, S. E. (2015). On the Causes of Effects: Response to Pearl. *Sociological Methods and Research*, 44(1):165–174.

Dawid, P., Musio, M., and Fienberg, S. E. (2016). From statistical evidence to evidence of causality. *Bayesian Analysis*, 11(3):725–752.

Degtiar, I. and Rose, S. (2021). A review of generalizability and transportability. *arXiv preprint (2102.11904).*

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. (2021). An automated approach to causal inference in discrete settings. *arXiv preprint arXiv:2109.13471.*

Duchi, J., Hashimoto, T., and Namkoong, H. (2020a). Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982.*

Duchi, J. C., Hashimoto, T., and Namkoong, H. (2020b). Distributionally robust losses for latent covariate mixtures. *arXiv (2007.13982)*, pages 1–39.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Eddy, D. M., Hasselblad, V., and Shachter, R. (1990). An introduction to a bayesian method for meta-analysis: the confidence profile method.

Encyclopedia, W. (2008). *West's Encyclopedia of American Law*. The Gale Group, 2nd edition.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More Robust Doubly Robust Off-policy Evaluation. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456, Stockholmsmässan, Stockholm Sweden. PMLR.

Farrell, M. H. (2018). Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint (1309.4686)*.

FDA (2018). Framework for fda's real-world evidence program. https://www.fda.gov/media/120060/download. Accessed: 2023-5-31.

FDA (2020). Memorandum explaining basis for revocation of emergency use authorization for emergency use of chloroquine phosphate and hydroxychloroquine sulfate. https://www.fda.gov/media/138945/download. Accessed: 2023-05-19.

FDA (2021). FDA approves new use of transplant drug based on real-world evidence. https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-new-use-transplant-drug-based-real-world-evidence. Accessed: 2023-1-2.

FDA (2022). Clinical decision support software — guidance for industry and food and drug administration staff. https://www.fda.gov/media/109618/download. Accessed: 2022-10-19.

Fehrenbacher, L., Ackerson, L., and Somkin, C. (2009). Randomized clinical trial eligibility rates for chemotherapy (ct) and antiangiogenic therapy (aat) in a population-based cohort of newly diagnosed non-small cell lung cancer (nsclc) patients. *Journal of Clinical Oncology*, 27:6538–6538.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283–286.

Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F., and Small, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514):447–458.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10.

Frost, C. and Thompson, S. G. (2000). Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable. *Journal of the Royal Statistical Society: Series A*, 163(2):173–189.

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2052–2062. PMLR.

Fuller, W. A. (1987). *Measurement error models*. John Wiley and Sons Inc.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. (2022). Leveraging unlabeled data to predict Out-of-Distribution performance. In *ICLR*.

Gautret, P., Lagier, J.-C., Parola, P., Hoang, V. T., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V. E., Tissot Dupont, H., Honoré, S., Colson, P., Chabrière, E., La Scola, B., Rolain, J.-M., Brouqui, P., and Raoult, D. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *International journal of antimicrobial agents*, 56(1):105949.

Goh, S. T. and Rudin, C. (2015). Cascaded high dimensional histograms: A generative approach to density estimation. *arXiv preprint arXiv:1510.06779*.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019a). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18.

Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., and Doshi-Velez, F. (2019b). Combining Parametric and Nonparametric Models for Off-Policy Evaluation. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306.

Guez, A., Vincent, R. D., Avoli, M., and Pineau, J. (2008). Adaptive Treatment of Epilepsy via Batch-mode Reinforcement Learning. *AAAI*.

Guo, R., Zhang, P., Liu, H., and Kiciman, E. (2021). Out-of-distribution Prediction with Invariant Risk Minimization: The Limitation and An Effective Fix. *arXiv (2101.07732)*.

Gupta, K., Hooton, T. M., Naber, K. G., Wullt, B., Colgan, R., Miller, L. G., Moran, G. J., Nicolle, L. E., Raz, R., Schaeffer, A. J., and Soper, D. E. (2011). International clinical practice guidelines for the treatment of acute uncomplicated cystitis and pyelonephritis in women: A 2010 update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. *Clinical Infectious Diseases*, 52(5):e103–20.

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., and Schünemann, H. J. (2008a). What is "quality of evidence" and why is it important to clinicians? *Bmj*, 336(7651):995–998.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008b). Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj*, 336(7650):924–926.

Hanna, J. P., Stone, P., and Niekum, S. (2017). Bootstrapping with models: Confidence intervals for off-policy evaluation. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 1(May):538–546.

Hazan, T. and Jaakkola, T. (2012). On the partition function and random maximum a-posteriori perturbations. In *ICML*.

Hazan, T., Papandreou, G., and Tarlow, D. (2016). *Perturbation, Optimization, and Statistics*. MIT Press.

He, B., Kwan, A. C., Cho, J. H., Yuan, N., Pollick, C., Shiota, T., Ebinger, J., Bello, N. A., Wei, J., Josan, K., Duffy, G., Jujjavarapu, M., Siegel, R. b., Cheng, S., Zou, J. Y., and Ouyang, D. (2023). Blinded, randomized trial of sonographer versus ai cardiac function assessment. *Nature*, 616:520–524.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heinze-Deml, C. and Meinshausen, N. (2021). Conditional variance penalties and domain shift robustness. *Machine learning*, 110(2):303–348.

Hernan, M. and Robbins, J. (2019). *Causal Inference*. Chapman & Hall/CRC, forthcoming, Boca Raton.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372.

Herrera, F., Carmona, C. J., González, P., and Del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR.

Hussain, Z., Oberst, M., Shih, M.-C., and Sontag, D. (2022). Falsification before extrapolation in causal effect estimation. In *Advances in Neural Information Processing Systems*.

Hussain, Z., Shih, M.-C., Oberst, M., Demirel, I., and Sontag, D. (2023). Falsification of internal and external validity in observational studies via conditional moment restrictions. In *26th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hyslop, D. R. and Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business and Economic Statistics*, 19(4):475–481.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, pages 46–60.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jacob, D. (2019). Group average treatment effects for observational studies. *arXiv preprint (1911.02688)*.

Jain, S., Lawrence, H., Moitra, A., and Madry, A. (2022). Distilling model failures as directions in latent space. *arXiv preprint (2206.14754)*.

Jakobsen, M. E. and Peters, J. (2020). Distributional Robustness of K-class Estimators and the PULSE. *arXiv (2005.03353)*.

Jeter, R., Josef, C., Shashikumar, S., and Nemati, S. (2019). Does the Artificial Intelligence Clinician learn optimal treatment strategies for sepsis in intensive care? *arXiv preprint*.

Ji, C. X., Oberst, M., Kanjilal, S., and Sontag, D. (2021). Trajectory inspection: A method for iterative clinician-driven design of reinforcement learning studies. In *AMIA Virtual Informatics Summit*.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. *33rd International Conference on Machine Learning, ICML 2016*, 2:1022–1035.

Jiang, Y., Nagarajan, V., Baek, C., and Zico Kolter, J. (2022). Assessing generalization of SGD via disagreement. In *ICLR*.

Johansson, F., Sontag, D., and Ranganath, R. (2019). Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536.

Johansson, F. D., Shalit, U., and Sontag, D. (2016). Learning Representations for Counterfactual Inference. In *ICML*.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Jung, Y., Tian, J., and Bareinboim, E. (2021). Estimating identifiable causal effects through double machine learning. In *AAAI*. aaai.org.

Kahn, H. (1955). Use of Different Monte Carlo Sampling Techniques. Technical report, RAND Corporation, Santa Monica, California.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*.

Kallus, N. (2019). Classifying treatment responders under causal effect monotonicity. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3201–3210, Long Beach, California, USA. PMLR.

Kallus, N., Puli, A. M., and Shalit, U. (2018). Removing hidden confounding by experimental grounding. In S. Bengio and H. Wallach and H. Larochelle and K. Grauman and N. Cesa-Bianchi and R. Garnett, editor, *Advances in Neural Information Processing Systems*, 31. Curran Associates, Inc.

Kallus, N. and Zhou, A. (2018a). Confounding-Robust Policy Improvement. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9269–9279. Curran Associates, Inc.

Kallus, N. and Zhou, A. (2018b). Policy evaluation and optimization with continuous treatments. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 84:1243–1251.

Kamath, P., Tangella, A., Sutherland, D. J., and Srebro, N. (2021). Does Invariant Risk Minimization Capture Invariance? *arXiv (2101.01134)*.

Kanjilal, S., Oberst, M., Boominathan, S., Zhou, H., Hooper, D. C., and Sontag, D. (2020). A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science translational medicine*, 12(568).

Keum, N., Lee, D., Greenwood, D., Manson, J., and Giovannucci, E. (2019). Vitamin d supplementation and total cancer incidence and mortality: a meta-analysis of randomized controlled trials. *Annals of Oncology*, 30(5):733–743.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2018). CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720.

Kook, L., Sick, B., and Bühlmann, P. (2022). Distributional anchor regression. *Statistics and Computing*, 32(3):1–19.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. (2020). Out-of-Distribution Generalization via Risk Extrapolation (REx). *arXiv (2003.00688)*.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684. ACM.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620.

Lam, H. (2016). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275.

Landgren, O., Siegel, D. S., Auclair, D., Chari, A., Boedigheimer, M., Welliver, T., Mezzi, K., Iskander, K., and Jakubowiak, A. (2018). Carfilzomib-lenalidomide-dexamethasone versus bortezomib-lenalidomide-dexamethasone in patients with newly diagnosed multiple myeloma: results from the prospective, longitudinal, observational commpass study. *Blood*, 132:799.

Lane, J. C. E., Weaver, J., Kostka, K., Duarte-Salles, T., Abrahao, M. T. F., Alghoul, H., Alser, O., Alshammari, T. M., Biedermann, P., Banda, J. M., Burn, E., Casajust, P., Conover, M. M., Culhane, A. C., Davydov, A., DuVall, S. L., Dymshyts, D., Fernandez-Bertolin, S., Fišter, K., Hardin, J., Hester, L., Hripcsak, G., Kaas-Hansen, B. S., Kent, S., Khosla, S., Kolovos, S., Lambert, C. G., van der Lei, J., Lynch, K. E., Makadia, R., Margulis, A. V., Matheny, M. E., Mehta, P., Morales, D. R., Morgan-Stewart, H., Mosseveld, M., Newby, D., Nyberg, F., Ostropolets, A., Park, R. W., Prats-Uribe, A., Rao, G. A., Reich, C., Reps, J., Rijnbeek, P., Sathappan, S. M. K., Schuemie, M., Seager, S., Sena, A. G., Shoaibi, A., Spotnitz, M., Suchard,

M. A., Torre, C. O., Vizcaya, D., Wen, H., de Wilde, M., Xie, J., You, S. C., Zhang, L., Zhuk, O., Ryan, P., Prieto-Alhambra, D., and OHDSI-COVID-19 consortium (2020). Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. *The Lancet. Rheumatology*, 2(11):e698–e711.

Li, B., Ren, K., Shen, L., Hou, P., Su, Z., Di Bacco, A., Hong, J.-L., Galaznik, A., Dash, A. B., Crossland, V., et al. (2018a). Comparing bortezomib-lenalidomide-dexamethasone (vrd) with carfilzomib-lenalidomide-dexamethasone (krd) in the patients with newly diagnosed multiple myeloma (ndmm) in two observational studies. *Blood*, 132:3298.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018b). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Li, M., Namkoong, H., and Xia, S. (2021). Evaluating model performance under worst-case subpopulations. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. (2016). PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres*, 121.

Lim, J., Ji, C., Oberst, M., Blecker, S., Horwitz, L., and Sontag, D. (2021). Finding regions of heterogeneity in decision-making via expected conditional covariance. In *Advances in Neural Information Processing Systems*.

Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. (2018). Representation Balancing MDPs for Off-policy Policy Evaluation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 2644–2653. Curran Associates, Inc.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lodi, S., Sharma, S., Lundgren, J. D., Phillips, A. N., Cole, S. R., Logan, R., Agan, B. K., Babiker, A., Klinker, H., Chu, H., Law, M., Neaton, J. D., and Hernán, M. A. (2016). The per-protocol effect of immediate versus deferred antiretroviral therapy initiation. *AIDS*, 30(17).

Lorberbom, G., Johnson, D. D., Maddison, C. J., Tarlow, D., and Hazan, T. (2021). Learning generalized gumbel-max causal mechanisms. In *Advances in Neural Information Processing Systems*.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010.

Lovell, M. C. (2008). A simple proof of the FWL theorem. *Journal of Economic Education*, 39(1):88–91.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* Wiley, New York.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations (ICLR).*

Maddison, C. J. and Tarlow, D. (2017). Gumbel Machinery.

Maddison, C. J., Tarlow, D., and Minka, T. (2014). A* Sampling. *Advances in Neural Information Processing Systems.*

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS).*

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., and D'Amour, A. (2022). Causally motivated shortcut removal using auxiliary labels. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 739–766. PMLR.

Miao, W. and Tchetgen, E. T. (2018). A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects. *arXiv (1808.04945).*

Miettinen, O. S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology*, 99(5):325–332.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press, Cambridge, 2 edition.

National Cancer Institute (2012). Bortezomib in treating patients with newly diagnosed multiple myeloma. ClinicalTrials.gov Identifier NCT00075881.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in medicine*, 28(28):3562–3566.

NIH (2016). Relating clinical outcomes in multiple myeloma to personal assessment of genetic profile (com-mpass). *Clinical Trials website. https://clinicaltrials.gov/ct2/show/NCT01454297.*

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Re, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 151–159, New York, NY, USA. Association for Computing Machinery.

Oberst, M. (2019). Counterfactual policy introspection using structural causal models. Master's thesis, Massachusetts Institute of Technology.

Oberst, M., D'Amour, A., Chen, M., Wang, Y., Sontag, D., and Yadlowsky, S. (2022). Bias-robust integration of observational and experimental estimators. *arXiv preprint (2205.10467)*.

Oberst, M., Johansson, F., Wei, D., Gao, T., Brat, G., Sontag, D., and Varshney, K. (2020). Characterization of overlap in observational studies. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

Oberst, M. and Sontag, D. (2019). Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models. *Proceedings of the 36th International Conference on Machine Learning*, 97.

Oberst, M., Thams, N., Peters, J., and Sontag, D. (2021a). Regularizing towards causal invariance: Linear models with proxies. In *Proceedings of the 38th International Conference on Machine Learning*.

Oberst, M., Thams, N., Peters, J., and Sontag, D. (2021b). Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4932–4941. PMLR.

OSTP (2022). Blueprint for an ai bill of rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/. Accessed: 2023-6-1.

Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., and Doshi-Velez, F. (2017). Combining Kernel and Model Based Learning for HIV Therapy Selection. *AMIA Joint Summits on Translational Science Proc*, 2017:239–248.

Park, C. and Kang, H. (2019). A groupwise approach for inferring heterogeneous treatment effects in causal inference. *arXiv preprint (1908.04427)*.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1/2):93–149.

Pearl, J. (2000). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2nd edition.

Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference*, 3(2):259–266.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):247–254.

Pearl, J. and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-W. H., Ross, A., Faisal, A., and Doshi-Velez, F. (2018). Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning. *AMIA Annual Symposium*, pages 887–896.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, Cambridge, MA.

Pólik, I. and Terlaky, T. (2007). A survey of the S-lemma. *SIAM review*, 49(3):371–418.

Powell, M. J. (2006). The NEWUOA software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*, pages 255–297. Springer.

Precup, D., Sutton, R. S., and Singh, S. P. (2000). Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 759–766, San Francisco, CA, USA.

Prete, F. P., Pezzolla, A., Prete, F., Testini, M., Marzaioli, R., Patriti, A., Jimenez-Rodriguez, R. M., Gurrado, A., and Strippoli, G. F. (2018). Robotic versus laparoscopic minimally invasive surgery for rectal cancer: a systematic review and meta-analysis of randomized controlled trials. *Annals of surgery*, 267(6):1034–1046.

Prevost, T. C., Abrams, K. R., and Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in medicine*, 19(24):3359–3376.

Puli, A., Zhang, L. H., Oermann, E. K., and Ranganath, R. (2022). Out-of-distribution generalization in the presence of Nuisance-Induced spurious correlations. In *International Conference on Learning Representations.*

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2008). *Dataset Shift in Machine Learning.* The MIT Press.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors (2009). *Dataset Shift in Machine Learning.* MIT Press.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint (2103.00020).*

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach. In *Machine Learning for Healthcare*.

Raghu, A., Komorowski, M., and Singh, S. (2018). Model-Based Reinforcement Learning for Sepsis Treatment. *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*.

Ram, P. and Gray, A. G. (2011). Density estimation trees. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–635. ACM.

RECOVERY Collaborative Group, Horby, P., Mafham, M., Linsell, L., Bell, J. L., Staplin, N., Emberson, J. R., Wiselka, M., Ustianowski, A., Elmahi, E., Prudon, B., Whitehouse, T., Felton, T., Williams, J., Faccenda, J., Underwood, J., Baillie, J. K., Chappell, L. C., Faust, S. N., Jaki, T., Jeffery, K., Lim, W. S., Montgomery, A., Rowan, K., Tarning, J., Watson, J. A., White, N. J., Juszczak, E., Haynes, R., and Landray, M. J. (2020). Effect of hydroxychloroquine in hospitalized patients with covid-19. *The New England journal of medicine*, 383(21):2030–2040.

Rivest, R. L. (1987). Learning decision lists. *Machine learning*, 2(3):229–246.

Robins, J. M. (1986). A New Approach to Causal Inference In Mortality Studies with a Sustained Exposure Period - Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant Models for Causal Transfer Learning. *Journal of Machine Learning Research*, 19(36):1–34.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenbaum, P. R., Rosenbaum, P., and Briskman (2010). *Design of observational studies*, volume 10. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 45(2):212–218.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rosenfeld, E. and Risteski, A. (2020). The Risks of Invariant Risk Minimization. *arXiv (2010.05761v1)*, pages 1–36.

Rosenman, E., Basse, G., Owen, A., and Baiocchi, M. (2020). Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*.

Rosenman, E. T., Owen, A. B., Baiocchi, M., and Banack, H. R. (2021). Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*.

Ross, C. (2021). Epic's sepsis algorithm is going off the rails in the real world. the use of these variables may explain why. https://www.statnews.com/2021/09/27/epic-sepsis-algorithm-antibiotics-model. Accessed: 2022-10-19.

Ross, C. (2022). Epic overhauls popular sepsis algorithm criticized for faulty alarms. https://www.statnews.com/2022/10/03/epic-sepsis-algorithm-revamp-training. Accessed: 2022-10-19.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A., Howard, B. V., Johnson, K. C., et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Jama*, 288(3):321–333.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246.

Rubin, D. B. and Rosenbaum, P. R. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of American Statistical Association*, 79:516–524.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Saag, M. S. (2020). Misguided use of hydroxychloroquine for COVID-19: The infusion of politics into science. *JAMA: the journal of the American Medical Association*, 324(21):2161–2162.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally robust neural networks. In *International Conference on Learning Representations*.

Sanchez, G. V., Babiker, A., Master, R. N., Luu, T., Mathur, A., and Bordon, J. (2016). Antibiotic Resistance among Urinary Isolates from Female Outpatients in the United States in 2003 and 2012. *Antimicrobial Agents and Chemotherapy*, 60(5):2680–2683.

Schnatz, P. F., Jiang, X., Aragaki, A. K., Nudy, M., O'Sullivan, D. M., Williams, M., LeBlanc, E. S., Martin, L. W., Manson, J. E., Shikany, J. M., et al. (2017). Effects of calcium, vitamin d, and hormone therapy on cardiovascular disease risk factors in the women's health initiative: a randomized controlled trial. *Obstetrics and gynecology*, 129(1):121.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., and Suchard, M. A. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2571–2577.

Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., and Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine*, 33(2):209–218.

Schulam, P. and Saria, S. (2017). Reliable Decision Support using Counterfactual Models. In *31st Conference on Neural Information Processing Systems*.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., and Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius : sociological research for a dynamic world*, 6.

Scott, C. D. and Nowak, R. D. (2006). Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704.

Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The econometrics journal*, 24(2):264–289.

Shalit, U., Johansson, F. D., and Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. In *33rd International Conference on Machine Learning (ICML)*.

Shapiro, D. J., Hicks, L. A., Pavia, A. T., and Hersh, A. L. (2013). Antibiotic prescribing for adults in ambulatory care in the USA, 2007–09. *Journal of Antimicrobial Chemotherapy*, 69(1):234–240.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen, E. J. T. (2018). Multiply Robust Causal Inference with Double Negative Control Adjustment for Categorical Unmeasured Confounding. *arXiv (1808.04906).*

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144.

Singh, K., Valley, T. S., Tang, S., Li, B. Y., Kamran, F., Sjoding, M. W., Wiens, J., Otles, E., Donnelly, J. P., Wei, M. Y., McBride, J. P., Cao, J., Penoza, C., Ayanian, J. Z., and Nallamothu, B. K. (2021). Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with covid-19. *Annals of the American Thoracic Society*, 18:1129–1137.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353.

SPRINT Research Group, Wright, Jr, J. T., Williamson, J. D., Whelton, P. K., Snyder, J. K., Sink, K. M., Rocco, M. V., Reboussin, D. M., Rahman, M., Oparil, S., Lewis, C. E., Kimmel, P. L., Johnson, K. C., Goff, Jr, D. C., Fine, L. J., Cutler, J. A., Cushman, W. C., Cheung, A. K., and Ambrosius, W. T. (2015). A randomized trial of intensive versus standard Blood-Pressure control. *The New England journal of medicine*, 373(22):2103–2116.

Srivastava, M., Hashimoto, T., and Liang, P. (2020). Robustness to Spurious Correlations via Human Annotations. *37th International Conference on Machine Learning.*

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-Estimation. *The American statistician*, 56(1):29–38.

Stewart, A. K., Rajkumar, S. V., Dimopoulos, M. A., Masszi, T., Špička, I., Oriol, A., Hájek, R., Rosiñol, L., Siegel, D. S., Mihaylov, G. G., et al. (2015). Carfilzomib, lenalidomide, and dexamethasone for relapsed multiple myeloma. *New England Journal of Medicine*, 372(2):142–152.

Su, G., Wei, D., Varshney, K. R., and Malioutov, D. M. (2016). Learning sparse two-level Boolean rules. In *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pages 1–6.

Subbaswamy, A., Adams, R., and Saria, S. (2021). Evaluating model robustness and stability to dataset shift. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR.

Subbaswamy, A., Schulam, P., and Saria, S. (2019). Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Suchard, M. A., Schuemie, M. J., Krumholz, H. M., You, S. C., Chen, R., Pratt, N., Reich, C. G., Duke, J., Madigan, D., Hripcsak, G., and Ryan, P. B. (2019). Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826.

Sun, J., Peng, L., Li, T., Adila, D., Zaiman, Z., Melton-Meaux, G. B., Ingraham, N. E., Murray, E., Boley, D., Switzer, S., Burns, J. L., Huang, K., Allen, T., Steenburg, S. D., Gichoya, J. W., Kummerfeld, E., and Tignanelli, C. J. (2022). Performance of a chest radiograph AI diagnostic tool for COVID-19: A prospective observational study. *Radiology. Artificial intelligence*, 4(4):e210217.

Sutton, R. S. and Barto, A. G. (2017). *Reinforcement Learning: An Introduction.* MIT Press, 2nd edition.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An Introduction to Proximal Causal Learning. *arXiv (2009.10982)*.

Thams, N., Oberst, M., and Sontag, D. (2022). Evaluating robustness to dataset shift via parametric robustness sets. In *Advances in Neural Information Processing Systems*.

Thomas, P. S. and Brunskill, E. (2016). Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *33rd International Confernece on Machine Learning (ICML)*, volume 48.

Tian, J. and Pearl, J. (2000). Probabilities of causation : Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313.

Train, K. (2002). *Discrete choice methods with simulation.* Cambridge University Press.

Travers, J., Marsh, S., Williams, M., Weatherall, M., Caldwell, B., Shirtcliffe, P., Aldington, S., and Beasley, R. (2007). External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*, 62:219–223.

Veitch, D'Amour, Yadlowsky, and Eisenstein (2021). Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *Advances in neural information processing systems*, 34.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Visconti, G. and Zubizarreta, J. R. (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4:217–249.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, F. and Rudin, C. (2015). Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and MacNeille, P. (2017). A Bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18(70):1–37.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, NY.

Wei, D., Dash, S., Gao, T., and Gunluk, O. (2019). Generalized linear rule models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Welton, N. J., Ades, A. E., Carlin, J., Altman, D., and Sterne, J. (2009). Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):119–136.

Wolpert, R. L. and Mengersen, K. L. (2004). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science*, 19(3):450–471.

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., and Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, 181(8):1065–1070.

Xie, C., Ye, H., Chen, F., Liu, Y., Sun, R., and Li, Z. (2020). Risk Variance Penalization. *arXiv (2006.07544)*.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

Yamada, K. and Kuroki, M. (2017). Counterfactual-Based Prevented and Preventable Proportions. *Journal of Causal Inference*, 5(2).

Yang, H., Rudin, C., and Seltzer, M. (2017). Scalable Bayesian rule lists. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 1013–1022.

Yau, S. T. and Zhang, L. (2006). An upper estimate of integral points in real simplices with an application to singularity theory. *Math. Res. Lett.*, 13(6):911–921.

Yellott, J. I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144.

Yuille., G. P. and L, A. (2011). Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*.

Zhang, F. (2006). *The Schur complement and its applications*, volume 4. Springer Science & Business Media.

Zhang, J., Iyengar, V., Wei, D., Vinzamuri, B., Bastani, H. S., Macalalad, A. R., Fischer, A. E., Yuen-Reed, G., Mojsilovic, A., and Varshney, K. R. (2017). Exploring the causal relationships between initial opioid prescriptions and outcomes. In *AMIA Workshop on Data Mining for Medical Informatics*, Washington, DC.

Zhang, Y., Young, J. G., Thamer, M., and Hernán, M. A. (2018). Comparing the Effectiveness of Dynamic Treatment Strategies Using Electronic Health Records: An Application of the Parametric g-Formula to Anemia Management Strategies. *Health services research*, 53(3):1900–1918.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.