Machine learning and causality: Building efficient, and reliable models for decision-making

by

Maggie Makar

B.Sc, Mathematics and Economics, University of Massachusetts, Amherst (2013)
S.M., Electrical Engineering and Computer Science, MIT (2017)

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

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2021

Certified by .....
John V. Guttag
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

Machine learning and causality: Building efficient, and reliable models for decision-making

by

Maggie Makar

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We explore relationships between machine learning (ML) and causal inference. We focus on improvements in each by borrowing ideas from one another.

ML has been successfully applied to many problems, but the lack of strong theoretical guarantees has led to many unexpected failures. Models that perform well on the training distribution tend to break down when applied to different distributions; small perturbations can "fool" the trained model and drastically change its predictions; arbitrary choices in the training algorithm lead to vastly different models; and so forth. On the other hand, while there has been tremendous progress in developing causal inference methods with strong theoretical guarantees, existing methods typically do not apply in practice since they assume an abundance of data. Working at the intersection of ML and causal inference, we directly address the lack of robustness in ML, and improve the statistical efficiency of causal inference techniques.

The motivation behind the work presented in this thesis is to improve methods for building predictive, and causal models that are used to guide decision making. Throughout, we focus mostly on decision making in the healthcare context. On the ML for causality side, we use ML tools and analysis techniques to develop statistically efficient causal models that can guide clinicians when choosing between two treatments. On the causality for ML side, we study how knowledge of the causal mechanisms that generate observed data can be used to efficiently regularize predictive models without introducing biases. In a clinical context, we show how causal knowledge can be used to build robust, and accurate models to predict the spread of contagious infections. In a non-clinical setting, we study how to use causal knowledge to train models that are robust to distribution shifts in the context of image classification.

Thesis Supervisor: John V. Guttag
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

None of this work would have been possible without my advisor, John Guttag's, guidance and technical feedback, for which I am very grateful. Over the years John has become a trusted confidant, mentor, and friend. John's calm, and encouraging words in the face of every setback, and his excitement about good progress have kept me going when the going got tough. More importantly, I am thankful for John's encouragement to pursue meaningful research even if it's difficult and his constant reminders that life is not a task to be optimized but a journey to be enjoyed.

I am also deeply thankful for all the members of my committee. David Sontag's relentless pursuit of hard questions, unfettered enthusiasm, and unbridled scientific curiosity made me more bold in the questions I ask, and more rigorous in my journey to get answers. Jenna Wiens has taught me how to formulate scientific questions that have a deep and meaningful impact on both healthcare, and machine learning. Jenna's ardent support for her students inspires me to be a better mentor. I am eternally grateful for her support for me, personally, which entailed her bearing through a number of tearful Zoom calls.

I am grateful for the many collaborators that I have worked with and learned from over the years. Fredrik Johansson, Alexandar D'Amour, Uri Shalit, and Francesca Dominici have taught me a lot of what I know about causal inference. Caleb Miles, my unofficial collaborator, has be an immensely valuable sounding board.

My colleagues at the clinical and applied machine learning group. To Jen, Davis, Joel, Adrian, Amy, Guha, Yun, Divya, Katie Lewis, Katie Matton, Hallee, Jose, Ani, Emily, and Marianne. Thank you for sharing my joys, and tears; thank you for encouraging me, and for pushing back on my questionable research ideas. I have learned how to be a better researcher, and a better mentor thanks to Emily Mu, Advaith Anand, and Meghana Kameneni.

It is true that I did not start this thesis myself, but rather joined the ongoing efforts

of my mother, who started preparing for this work over two decades ago. From the Berenstain Bears all the way to causal inference, she has happily encouraged me to learn and grow as human being and a scholar. This work, like everything else in my life, would have not been possible without her unwavering support. I am grateful for my dad for always being eager and ready to help; regardless of what I need help with. My sister is my biggest cheerleader, I am acutely aware of how proud I make her. The motivation behind a lot of my hard work is live up to and maintain that pride.

I am thankful for the many conversations I've had with the members of the Longwood Christian Community that challenged me to ask questions that matter. Most importantly, I am thankful to God, who has blessed me with every skill and talent that I have; and given me a phenomenal set of mentors, colleagues, family, and friends. I am thankful that He instructed me and taught me in every way I should go, and that He counseled me with His eye upon me (Psalm 32:8). To Him be the glory, now and forever.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Seeking guidance to help with decision making is as old as the story of humanity itself. This need for guidance has been manifested in different ways across different time eras, civilizations and belief systems. Cleromancy, defined as "the casting of lots, in which an outcome is determined by means that normally would be considered random, such as the rolling of dice, but are sometimes believed to reveal the will of God, or other universal forces and entities" (Wikipedia contributors, 2021), is one way to seek such guidance. For example, ancient Romans resorted to casting *sortes* or lots. Ancient Judiac traditions relied on consulting the *Urim and Thummim*. In West-African cultures, e.g., in Yoruba and Yoruba-inspired religions, decision-makers resort to another type of cleromancy called *Ifá divination*. More recently, practitioners in the realms of healthcare, economics, education and energy are turning to machine learning-based cleromancy to provide guidance in key decision.

At its core, Machine Learning (ML) is a set of tools that encodes associations observed in a training data to make predictions about unseen, or test data. Unfortunately, relying on solely learned associations to make decisions might lead to suboptimal or even catastrophic decisions. For example, we might conclude that a particular treatment is associated with death if we observe that patients who receive a particular

treatment are more likely to die. However, looking beyond associations, and incorporating notions of causality, we would first adjust for the patients' underlying sickness level before deciding whether the treatment is truly dangerous. For the purpose of reliable decision making, it is hard to overestimate the importance of developing ML tools that encode causal relationships rather than mere correlations. This thesis will study the exchange of ideas between causal inference and predictive machine learning to build efficient and robust models that are reliable and useful for decision-making.

1.1 Machine learning for causality

Estimating the causal effect of an intervention or a treatment using observational data is an oft-studied problem especially in the statistics literature (Rubin, 2006; Pearl, 2000). For decades, statisticians developed methods primarily focused on average treatment effects, and meticulously studied the asymptotic consistency of these methods (Abadie and Imbens, 2006; Cochran and Rubin, 1973). Interest in estimating the conditional average treatment effect (CATE) remained a secondary focus. The availability of large data, and the advent of ML tools shifted the focus to CATE estimation, and provided new tools to deliver on the promise of personalized, targeted interventions. The theoretical lens moved from asymptotic analysis to analyses of generalization error in finite samples (Shalit et al., 2017; Johansson et al., 2016).

This thesis contributes to the continued evolution of causal inference methods by shifting the focus to the development of efficient and theoretically principled tools designed to mitigate different types of data limitations. At the core of our contributions is the idea that in order to make good decisions, the decision maker does not always need the most accurate possible estimate of the CATE; bounds on causal estimates might be both sufficient, and easier to estimate from limited data. We present one specific case in chapter 2: estimating reliable bounds on the causal effects of treatment options. We show that we can achieve better finite sample efficiency by estimating

bounds on causal effects rather than directly estimating the causal effect. Leveraging our theoretical insight, we propose a kernel-based method for efficient estimation of bounds on the potential outcomes under different treatment choices.

1.2 Causality for Machine learning

ML methods, and especially deep neural networks (DNNs) are often brittle: they exhibit a lack of robustness under minor perturbations in the input data, and they are unable to circumvent issues that arise because of biased training data (Beery et al., 2018; Ilyas et al., 2019; Azulay and Weiss, 2018; Geirhos et al., 2018). These failure modes may seem different, but the root cause can be traced back to flaws in the ML algorithms. Specifically, ML algorithms typically pick the model that achieves the best training and validation errors. When the candidate function class consists of all possible DNNs, multiple models might have a low (or even zero) training and validation error. Some of these performant models may exhibit robustness under distribution shift, while others may not. We refer to this as "underspecification": the training and validation error no longer uniquely specify or identify the optimal model.

In this thesis, we adapt ideas from causality to address the issue of underspecification, with the ultimate goal of building robust models that learn meaningful patterns. We do so by developing methods that require that models conform to some existing causal knowledge in addition to having a low training and validation error. We present two cases for "causally-motivated" predictive models.

First, in chapter 3, we develop causally-motivated regularization schemes using auxiliary labels to discourage shortcut learning (Geirhos et al., 2020). Shortcut learning occurs when a predictor relies on input features that are easy to discover and are predictive of the outcome in the training data, but do not remain predictive when the distribution of inputs changes. For example, using the background (sand/grass) to predict the main object (camel/cow) in an image. We assume that we have access to

auxiliary labels (e.g., the background labels), and require that the predictor's output is independent of the auxiliary labels. We show that causally-movtivated regularization leads to predictors that are robust to distribution shift. We also show that even in absence of distribution shift, causally-motivated regularization leads to more sample-efficient models.

Second, in chapter 4, we address distribution shift that arises because of the presence of asymptomatic carriers: individuals who carry a disease but do not display symptoms, silently spreading the pathogen. At training time, sypmtomatic carriers are over-represented since they are more likely to be tested for the disease. At deployment time, we wish to compute an accurate prediction for symptomatic, and asymptomatic carriers as well as healthy individuals. This creates a distribution shift. Here we leverage our knowledge of existing dependencies (rather than independencies) to build models that predict the onset of a contagious infection. In order to get infected, one must be exposed to the pathogen through their network of contacts. This implies that an individual's infection state depends on their contacts' infection states. In situations where the contacts' infection states are unobserved (e.g., untested asymptomatic carriers), knowing that this dependency exists allows us to impute the missing infection states, and ultimately train a model that is better able to identify both symptomatic and asymptomatic infections.

1.3 Contributions

The core contributions of this thesis include theoretical results analyzing the sample-efficiency of causal models, and the generalizability of prediction models under distribution shift. Guided by our theoretical analysis, we set forth novel techniques for efficient estimation of causal models, and for building predictive models that are robust to distribution shifts. In addition, we present an application of causally-inspired prediction models in the context of infectious disease modeling. Specifically,

Specifically, in **chapter 2** we present a study of how ML methods can be used to make causal inference more efficient, including:

1. **Theoretical analysis supporting bound estimation when it is sufficient for decision-making.** We analyze the finite sample properties of estimation of bounds on the potential outcomes (defined as the outcomes under different treatment decisions). Our analysis highlights a novel trade-off between confidence that the bounds contain the true potential outcomes and tightness of the bounds.

2. **An algorithm to efficiently estimate bounds on potential outcomes.** Guided by our theory, we develop a non-parametric, kernel-based model to estimate the optimal (e.g., tightest) bounds that contain the true potential outcome with high probability. We show that our suggested approach outperforms existing bound-estimation methods, and analyze why that happens.

In chapters 3, and 4 we study methods that utilize ideas from causality to make machine learning more robust. Specifically, we present

3. **Theoretical analysis showing that causally-motivated regularization is statistically-efficient.** We analyze the finite sample properties of models that are explicitly regularized to conform with *a priori* causal knowledge during training time. We compare their generalization error bounds to those of models that rely on the classical L2-norm penalty, showing when and why our approach is more efficient.

4. **An algorithm to train robust prediction models using causally-motivated regularization.** Based on our theoretical findings, we present a weighting scheme, a training objective and a two-step cross-validation algorithm that is both statistically efficient and robust to distribution shift. We compare the empirical performance of our approach to existing baselines, and conduct an

ablation study to highlight the importance of each component in our combined approach. We show that our approach yields predictors that are more robust to distribution shift compared to baselines. Even in the absence of distribution shift, we show that our approach gives more efficient predictors.

5. **An approach for contagious infection prediction.** In order to contract a contagious infection, an individual must be exposed to the pathogen through their network of contacts. We present an algorithm that leverages this dependency between individuals to predict the probability of an untested individual carrying the disease. In addition, we identify two properties of observed data that can be exploited to mitigate the effects of distribution shift caused by incomplete testing. We empirically evaluate the effectiveness of our method on both simulated and real data. We show that predictions from our model can be used to inform efficient testing and isolation policies. Using Electronic Health Record (EHR) data from a large hospital, we show that our model outperforms baselines on the task of predicting a healthcare associated infection.

We conclude in chapter 5 with a discussion of and future extensions to the work presented in this thesis. We end with a road map for future research at the intersection of machine learning and causality.

Chapter 2

Estimation of Bounds on Potential Outcomes For Decision Making

In this chapter, we use techniques developed for ML to study the generalization and efficiency of causal models. We leverage our theoretical analysis to develop a sample-efficient model that estimates intervals or bounds on causal effects.

In many practical situations, a decision maker wishes to intervene or assign a treatment to ensure that an outcome of interest falls within a safe range. One example, which we use throughout this chapter, is when a physician considers whether or not to prescribe anticoagulants to mitigate the risk of stroke, as measured by the International Normalized Ratio (INR). The INR reflects the time it takes for blood to clot. For previous stroke patients, a healthy INR is 2–3. Values lower than 2 signal elevated risk of an Ischemic stroke, and higher than 3 signal elevated risk of a Hemorrhagic stroke. To make an informed decision, the physician needs to know if the potential outcomes under treatment and non-treatment fall within 2–3. We highlight two characteristics of this scenario that guide our approach. First, without making any additional assumptions, learning that the difference between the potential outcomes, i.e., the Conditional Average Treatment Effect (CATE) is 1.5, does not immediately

imply an optimal treatment decision; it could be that the patient's INR decreases from 4 to 2.5 or from 5.5 to 4. Specifically, we do not assume that the outcome remains unchanged if no treatment is administered. This is the case when the patient has a health condition that would lead to deteriorating outcomes in the abscene of interventions. In this case, information about the potential outcomes themselves under different treatment considerations is needed. This motivates us to study estimators of the potential outcomes, rather than estimators of the CATE. Second, knowing the potential outcomes' exact value is not necessary. It is sufficient to know that the patient's INR is somewhere between 2.2 and 2.8 if treated. For example, knowing that it is 2.5 might not provide additional useful insight. This motivates us to study estimators of intervals, or bounds on the potential outcomes. Together these two characteristics motivate studying the task of estimating reliable covariate-conditional bounds on potential outcomes using observational data.

Most existing methods for estimating causal effects and potential outcomes attempt to fit the expected outcomes as functions of observed covariates, typically relying on variants of Empirical Risk Minimization (ERM) strategies (Hill, 2011; Shalit et al., 2017; Alaa and van der Schaar, 2018, 2017). Some of these methods produce prediction intervals centered around the estimated expected response (outcome) surface, which can be used to bound the potential outcome from above and below. These intervals have approximately valid coverage for large samples, provided that the mean estimate is sufficiently unbiased. However, achieving this is not always feasible in small samples, leading to high false coverage rates (FCRs), defined as the rate at which outcomes are observed outside of the given prediction interval.

Instead of attempting to directly fit the potential outcomes, which may be complex and hard to estimate from small samples, we propose to fit simpler functions that bound the outcomes from above and below. Within this simpler function class, we identify estimates of the potential outcomes that maximize a utility (objective) function specified by the decision maker. Figure 2-1 shows the intuition behind our approach. For example, if the decision maker wants to ensure that the uncertainty in

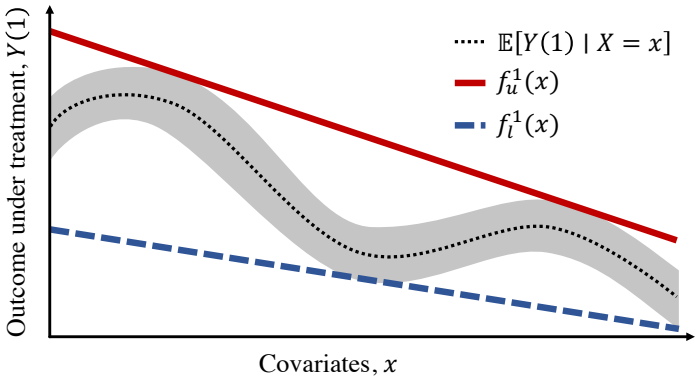

Figure 2-1: Illustration of the intuition behind our theoretical findings. The true potential outcome (black/gray) belongs to a complex class. But the upper (red) and lower (blue) bounds that correctly cover it belong to a simple linear function space.

the potential outcome estimates is small on average, they could require that the average interval width (upper bound - lower bound) is small. Alternatively, if they wish to ensure that no patient sub-population has excessively uncertain estimates (i.e., wide intervals) they could require that the maximum interval width is minimized.

We make the following main contributions:

1. We give results on the generalization properties of learned bounds on potential outcomes and the conditions under which estimation of such bounds yields better sample complexity than fitting the expected outcomes using standard risk minimization methods. Our analysis highlights a trade-off between reliability (i.e., the probability that the bounds correctly cover the data) and the complexity of the learning task.

2. We design an algorithm that finds the optimal bound estimates that maximize a given utility or objective function while providing reliable bounds. We explore different objective functions, analyzing the differences between the resulting bounds, and prove equivalence to quantile regression in a special case.

3. We evaluate our algorithm on both a semi-synthetic clinical dataset and a well-known causality benchmark. We show how our algorithm can guide treatment decisions, and that it achieves a better trade-off between bound violations and

utility than baseline algorithms.

2.1 Related work

Research into methods for estimating conditional causal effects has focused primarily on estimating the expected potential outcomes or conditional average treatment effect (CATE) as functions of observed covariates (Dorie et al., 2019). For example, (Alaa and van der Schaar, 2018) showed that the CATE estimation problem is as hard as modelling the more complex of the two potential outcomes in the minimax sense. Similarly, (Nie and Wager, 2017) show asymptotic bounds that rely on the complexity of the underlying function class of the CATE. More generally, recent work in CATE estimation has focused on the learning challenges associated with the difference between the treated and control populations, and on improving finite sample efficiency by sharing data between treatment groups (Johansson et al., 2016; Shalit et al., 2017; Alaa and van der Schaar, 2017; Hill, 2011). In contrast, we aim to improve sample efficiency by providing bounds on the causal estimands.

Other work focuses on estimating lower or upper bounds of Average Treatment Effect (ATE), to account for the possibility of unobserved confounding (Balke and Pearl, 1997; Bareinboim and Pearl, 2012; Pearl, 2009; Cai et al., 2008). Recently, this type of analysis was extended to include bounds on CATE, but again in the presence of hidden confounding (Kallus et al., 2019). This line of work falls under sensitivity analysis (Rosenbaum, 2014), which is distinct from our work in that we aim to find bounds on the potential outcomes when estimating CATE is not possible because of limited data, even if there is no hidden confounding.

Another related line of work is the problem of conditional quantile treatment effect estimation (Koenker and Bassett Jr, 1978; Chernozhukov and Hansen, 2005). Like our method, quantile methods give can give approximate bounds on the potential outcomes. The distinction is that the main objective of our method is not to estimate

the specific quantile of a treatment effect, but rather to provide the simplest functions that bound the outcomes such that an objective function given by the decision maker is optimized. However, as we prove later, quantile estimation is a special case of our setting for a particular objective function.

Recently, new work extended conformal intervals (Lei et al., 2018) to causal settings similar to ours (Lei and Candès, 2020). Our work is distinct from the work presented in (Lei and Candès, 2020) in three ways (1) we provide theoretical guarantees for the *finite* sample rather than asymptotic regime, (2) our theoretical analysis highlights a fundamental trade-off between the statistical complexity of the learning problem and the confidence with which the learned interval truly covers the potential outcomes. Finally, (3) our approach allows for a more general definition of interval optimality; we not assume that tightness of the bounds is the only important metric to be optimized, but it allows the decision maker to define their own desiderata for optimality (e.g., fairness).

Our work is related to offline policy learning (e.g., Swaminathan and Joachims (2015a,b)). The main difference between this work and ours is that we wish to obtain bounds for the potential outcomes, not an optimal policy. This allows the decision maker to consider the estimated effect of the treatment against a backdrop of additional information that may not be recorded in the observational data.

2.2 Problem setup

We consider learning of bounds on potential outcomes from finite-sample observational data, adopting the notation of the Neyman-Rubin potential outcomes framework (Rubin, 2005). For each unit i (e.g., patient), we observe a set of features $X_i \in \mathcal{X}$, with $\mathcal{X}$ a bounded subset of $\mathbb{R}^d$, an action (also known as treatment or intervention) $T_i \in \{0, 1\}$ and an outcome $Y_i \in \mathbb{R}$. We observe these variables through samples $(x_1, t_1, y_1), ..., (x_n, t_n, y_n) \overset{i.i.d.}{\sim} p(X, T, Y)$ and denote by $n_t = \sum_{i=1}^{n} \mathbb{1}\{t_i = t\}$ the num-

ber of observed samples for treatment group $t \in \{0, 1\}$, and let $p_t(X) = p(X \mid T = t)$. The observed outcome is one of the two *potential outcomes*, $Y(0)$ and $Y(1)$, under control ($T = 0$) and treatment ($T = 1$), respectively. We use $\|a\|_p$ to denote the $p-$norm of a vector a. When the subscript is omitted, we refer to the 2-norm.

We seek to learn high-probability bounds on both potential outcomes, $Y(0)$ and $Y(1)$, conditioned on the set of observed features X. Since only one outcome is observed, the other is not identifiable without strong assumptions. To that end, we assume that the features X are sufficient to deconfound estimates of $Y(0), Y(1)$:

Assumption 2.2.1. *The features X, treatment T and potential outcomes $Y(0), Y(1)$ satisfy for some $\epsilon > 0$*

1. *Strong ignorability:* $Y(0), Y(1) \perp\!\!\!\perp T \mid X$
2. *Overlap:* $\forall x, t : p(T = t \mid x) > \epsilon$
3. *Consistency:* $Y = Y(T)$

Under Assumption 2.2.1, $p(Y(t) = y \mid X = x) = p(Y = y \mid T = t, X = x)$ (Imbens and Wooldridge, 2009). This means that the distribution of potential outcomes can be estimated through regression or other standard methods. When treatment and outcomes are confounded, estimates of causal effects exhibit bias. For example, if medication A was prescribed more often to terminally ill patients than the alternative treatment B, we might learn that the life expectancy on treatment A was lower than on B, regardless of its average causal effect. To undo this bias, it is common to use the propensity score $e(x, t) := p(T = t \mid X = x)$ to re-weight the cohort using importance weighting.

Definition 2.2.1. *The importance weighting function w_t for group $t \in \{0, 1\}$ is* $w_t(x) := p(T = t)/e(x, t)$.

We use w_i to denote $w_{t_i}(x_i)$ for a sample $(x_i, t_i) \sim p$. With w_t as in Definition 2.2.1, we have, for an arbitrary function f on $\mathcal{X}$ (e.g., the expected outcome or a prediction

loss), $\mathbb{E}_X[f(X)] = \mathbb{E}_{X|T}[w_t(X)f(X) \mid T = t]$. By Assumption 2.2.1, we have that the importance weights are bounded, meaning that for some $C_t < \infty$ and $t \in \{0, 1\}$:

$$\sup_{x \in \mathcal{X}} w_t(x) = \sup_{x \in \mathcal{X}} \frac{p(T = t)}{e(x, t)} = 2^{D_\infty(p||p_t)} = C_t, \tag{2.1}$$

where $D_k(p||q)$ is the k$^{\text{th}}$-order Rényi divergence, and the second equality follows by applying the Bayes rule, and the definition of the Rényi divergence. It will be convenient to denote $2^{D_k(p||q)}$ by $d_k(p||q)$. Since $2^{D_{k-1}(p||p_t)} < 2^{D_k(p||p_t)}$, we have $d_2(p||p_t) < C_t$.

2.3 Generalization of bounds on potential outcomes

Our goal is to estimate four functions; lower and upper bounds for the potential outcome under treatment, $\boldsymbol{f^1}(x) = \{f_l^1(x), f_u^1(x)\}$, and similarly defined functions for the outcome under control $\boldsymbol{f^0}(x) = \{f_l^0(x), f_u^0(x)\}$. For these estimates to be useful for decision-making, we want to guarantee that for some small $\nu' > 0$, and for $t \in \{0, 1\}$, we have false coverage rate (FCR) bounded by ν',

$$\text{FCR}_{\boldsymbol{f^t}} := \Pr_{X, Y(t)} \left[Y(t) \notin [f_l^t(X), f_u^t(X)] \right] \leq \nu' \ . \tag{2.2}$$

Without loss of generality, we will focus on estimating a lower bound for the outcome under treatment $T = t$, meaning we will focus on finding some $f_l^t(x)$ such that for a small $\nu > 0$, we have that

$$\Pr_{X, Y(t)} [f_l^t(X) \leq Y(t)] \geq 1 - \nu. \tag{2.3}$$

Note that in expressions 2.2 and 2.3 the probabilities are defined over $p(X, Y(t)) \neq p(X, Y \mid T = t)$, because of confounding. However, under Assumption 2.2.1, this probability is identifiable from observed data.

It will be useful to restate our objective in terms of the (signed) residual of a function f, defined next.

Definition 2.3.1. *For an arbitrary function f, the signed residuals for $x, y \in \mathcal{X} \times \mathcal{Y}$:* $\underline{r}_f(x, y) = y - f(x)$.

Expression (2.3) can now be restated as $\Pr[\underline{r}_{f_l^t}(X, Y(t)) \geq 0] \geq 1 - \nu$. To be more cautious, we might wish to leave a "buffer zone" or a margin, and instead demand that $\underline{r}_{f_l^t}(x, y) \geq \gamma$ for some $\gamma > 0$. In this setting, a violation occurs when $\underline{r}_{f_l^t}(x, y) < \gamma$. Larger values of γ would imply higher reliability: we are more confident that we are unlikely to observe a violation of the bounds, i.e., unlikely to overestimate the outcome under treatment t. With that, direct parallels could be drawn between our setup and that of maximum-margin algorithms: we want to ensure that the signed residual is larger than 0 by a margin of γ. The larger γ is, the more confident we are that our lower bound holds. We can now define the unobserved risk that we wish to study:

Definition 2.3.2. *For $f_l^t \in \mathcal{F}$, $\gamma > 0$, we define the risk of overestimation over the full unknown distribution:*

$$\underline{R}_{f_l^t}(\gamma) = \mathbb{E}_{X,Y(t)} \left[\mathbb{1}\{\underline{r}_{f_l^t}(X, Y(t)) < \gamma\} \right].$$

To account for confounding caused by biased (non-randomized) treatment assignment, we consider a re-weighted risk:

$$\underline{R}^w_{f_l^t}(\gamma) = \mathbb{E}_{X,Y|T} \left[w(x) \mathbb{1}\{\underline{r}_{f_l^t}(X, Y) < \gamma\} \mid T = t \right]$$

Under Assumption 2.2.1, $\underline{R}_{f_l^t}(\gamma) = \underline{R}^{w_1}_{f_l^t}(\gamma)$. Since our notions of confidence are closely related to the margin, γ, it will be useful to reason about the magnitude of margin violations, which is defined next.

Definition 2.3.3. *For $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, known $\boldsymbol{w}_t$, $f_l^t \in \mathcal{F}$,*

and $\gamma > 0$, we define the average weighted magnitude of training set violations as

$$\underline{D}^{\boldsymbol{w}_t}(z, f_l^t, \gamma) = \sum_{x,y \in z} w_t(x) \max\{0, \gamma - \underline{r}_{f_l^t}(x, y)\}$$

In the remainder of this section, we give bounds on expected margin violation as a function of $\underline{D}^{\boldsymbol{w}_t}$. We restrict our analyses to sturdy function classes, with with range $= [a, b]$. Formally, the definition of sturdy function classes is as follows:

Definition 2.3.4. *[Restated from Shawe-Taylor and Williamson (1999)] We say that a function class $\mathcal{F}$ is sturdy if it maps X of size n to a compact subset of $\mathbb{R}^n$ for any $n \in \mathbb{N}$.*

We rely on the covering number as a measure of complexity of the analyzed function classes.

Definition 2.3.5. *Let (X, l_∞) be a pseudo-metric space defined with respect to the l_∞ norm, and let A be a subset of X and $\epsilon > 0$. A set $U \subseteq X$ is an ϵ-cover for A if for every $a \in A$, there exists $u \in U$ such that $||a - u||_{l_\infty} \leq \epsilon$. The ϵ-covering number of A, $\mathcal{N}(\epsilon, A, d)$ is the minimal cardinality of the ϵ-cover for A.*

We use fat-shattering dimensions to study how fast the complexity of a function class can grow with the sample size.

Definition 2.3.6. *[Restated from Bartlett and Shawe-Taylor (1999)] For $\gamma \in [0, \infty]$, and $\mathcal{F} \in \mathbb{R}$, we say that a set of points $\{x_i\}_{i=1}^n$ is $\gamma-$shattered by $\mathcal{F}$ if there exists $\{s_i\}_{i=1}^n \in \mathbb{R}$ such that for all binary vectors $\{\sigma_i\}_{i=1}^n$, there is a function $f \in \mathcal{F}$ satisfying:*

$$f(x_i) \geq s_i + \gamma \qquad if\ \sigma_i = 1$$
$$f(x_i) \leq s_i - \gamma \qquad otherwise$$

The fat-shattering dimension can be thought of as a function from the positive reals to the set of positive integers which maps γ to the largest $\gamma-$shattered set or ∞.

2.3.1 Generalization of reliable estimators

We start by studying the risk of overestimation for re-weighted estimators. To make our main finding simpler to follow, we focus on the class of linear functions in a kernel defined feature space. Theorem A1 in the supplement gives the analogous bounds for more general function spaces. We start by stating the theorem and then outline the key takeaways from the theorem.

Theorem 2.3.1. *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in expression (2.1). For $f_l^t \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_t}(z, f_t^l, \gamma) = D > 0$. With a probability $1 - \delta$ over the draw of random samples, we have that:*

$$\underline{R}_{f_l^t}(\gamma) \leq \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}} \tag{2.4}$$

where, for $t \in \{0, 1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2n_t) + \frac{D}{\gamma} \log \frac{\exp(n_t + {}^{D}/_{\gamma} - 1)}{{}^{D}/_{\gamma}} \right\rceil .$$

Recall that d_2, and C_t are defined below equation 2.1. The proof of the theorem is outlined in the appendix. **Remarks:**

1. Theorem 2.3.1 states that the expected rate of overestimation is bounded by terms that are at most linear in k_t—the sum of the log covering number of $\mathcal{F}$ *as defined by the margin* γ, and the ratio of the violations on the training data to γ. The fact that the covering number is controlled by the margin parameter γ shows that the complexity of this learning task relies on how certain we wish

to be that the lower bound is not overestimated; more certainty requires a larger γ which implies a smaller log covering number. This approach departs from previous literature which instead shows that the sample complexity of risk minimization relies on the covering number of a class containing the true function (Alaa and van der Schaar, 2018). In applications where it is sufficient to have reliable *bounds* on the potential outcomes to make good decisions, this difference can be crucial—especially if the outcomes are difficult to estimate accurately using small samples. Note that the covering number can be bounded by the fat-shattering dimension at a scale proportional to γ.

2. Both terms in k_t decrease as γ increases, which means that the risk of overestimation decreases as γ increases. This property is important because it implies that we can control the risk of overestimation by requiring a large margin. To see that, note that larger γ shrinks the space of viable functions, which decreases the γ-covering number. The second term includes the ratio of the sum of violations on the training set, D, which decreases as γ increases, to γ. Hence the second term also decreases as γ increases.

The following corollary builds on theorem 2.3.1 to get a bound on the generalization error for bounds on the CATE. The risk of overestimation for the CATE can be stated as a simple extension of theorem 2.3.1. We define the CATE as $\tau(x) = Y(x,1) - Y(x,0)$, where $Y(x,t)$ is the potential outcome under treatment $T = t$, for patient with characteristics $X = x$. We use $\tilde{\tau}_l(x)$ to denote $f_l^1(x) - f_u^0(x)$, where f_l^1, f_u^0 are some estimates of the lower bound for the outcome under treatment and the upper bound of the outcome under non-treatment respectively. In addition, we define:

$$\overline{r}_f(x,y) = f(x) - y,$$

and for $z_t = \{x_i, y_i\}_{i:t_i=t}$, define

$$\overline{D}^{\boldsymbol{w}_t}(z, f_u^t, \gamma) = \sum_{x,y \in z} w_t(x) \min\{0, \gamma - \overline{r}_{f_u^t}(x,y)\}$$

Corollary 2.3.1. *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z_t = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X,Y)$, and C_t be as defined in expression (2.1). For $f_l^1, f_u^0 \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_1}(z_1, f_l^1, \gamma) = D_1 > 0$, and $\overline{D}^{\boldsymbol{w}_0}(z_0, f_u^0, \gamma) = D_0 > 0$ Define $\tilde{\tau}_l := f_l^1 - f_u^0$. With probability $1-\delta$ over random samples, we have that:*

$$\underline{R}_{\hat{\tau}_l}(\gamma) \leq \sum_t \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}. \tag{2.5}$$

where, for $t \in \{0,1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2n_t) + \log \mathcal{N}(\gamma/2, L^{D_t}(\mathcal{X}), 2n_t) \right\rceil .$$

Recall that d_2, and C_t are defined below equation 2.1.

2.3.2 Generalization of reliable, informative estimators

Theorem 2.3.1 establishes that the probability of overestimation decreases as we increase the margin γ. However, arbitrarily large values of γ could result in excessively "cautious" estimates with low risk of overestimation, at the expense of being too loose to be useful in guiding decisions. In this work, we consider bounds to be informative or have high utility if they imply low uncertainty in the value of the true potential outcomes. We restrict ourselves to definitions of uncertainty that rely on the interval width (IW) of bounds $\boldsymbol{f} := (f_u, f_l)$

$$\mathrm{IW}_{\boldsymbol{f}}(x) := f_u(x) - f_l(x) \ . \tag{2.6}$$

Smaller $\mathrm{IW}_{\boldsymbol{f}}(x)$ implies that bounds are tighter, which implies less uncertainty in the value of the potential outcomes. Intuitively, for f_u and f_l to give small $\mathrm{IW}_{\boldsymbol{f}}$, they need to be *close* to each other. We define these "close" functions and the informative classes to which they belong as follows:

Definition 2.3.7. *Let $p \geq 1$, and $\mathcal{X} := \{x : ||x|| \leq r\}$. We say that two classes of bounded linear functionals $\mathcal{F}_l, \mathcal{F}_u$ are* informative *if $\mathcal{F}_l \subseteq \{\mathcal{X} \ni x \mapsto \langle f_l, x\rangle, ||f_l|| \leq A\}$ and $\mathcal{F}_u \subseteq \{\mathcal{X} \ni x \mapsto \langle f_u, x\rangle, \forall f_l \in \mathcal{F}_l; ||f_u - f_l|| < B, \forall x \in \mathcal{X} : f_l(x) \leq f_u(x)\}$.*

In words, $\mathcal{F}_l$ is the set of functions with norm $\leq A$, while $\mathcal{F}_u$ is the set of functions that are within B distance from each $f_l \in \mathcal{F}_l$. In addition, we specify that $f_l(x) \leq f_u(x)$ for every $x \in \mathcal{X}$.

The next theorem extends theorem 2.3.1 to these informative function classes, allowing us to study the risk of overestimation for tight intervals. To improve readability, log terms which do not affect the interpretation of the statement have been suppressed. The full statement is presented in Theorem A1 in the appendix.

Theorem 2.3.2. *Let $\mathcal{F}_l^t$, $\mathcal{F}_u^t$, A, B, and r be as defined in definition 2.3.7. Let z, and D be as defined in theorem 2.3.1,and C_t be as defined in expression* (2.1)*. For $f_l^t \in \mathcal{F}_l^t$, $f_u^t \in \mathcal{F}_u^t$ and any $\gamma > 0$, with a probability $1 - \delta$ over the draw of random samples, the bound* (2.4) *in Theorem 2.3.1 applies with*

$$k_t \approx \left\lceil \left(\frac{r(A+B)}{\gamma} \right)^2 + \frac{D}{\gamma} \log \frac{e(n_t + {}^{D}/_{\gamma} - 1)}{{}^{D}/_{\gamma}} \right\rceil ,$$

for $t \in \{0, 1\}$

Theorem 2.3.2 gives us an idea of how to learn informative bounds that reliably cover the potential outcomes. It suggests that one way to reduce generalization error is to minimize A, the norm of f_l^t, B the distance between f_l^t and f_u^t, and D, the sum of violations on the training data.

2.4 Learning reliable, informative bounds

We present the Bounded Potential outcomes algorithm (BP) for learning informative bounds on potential outcomes under the constraint that they are violated with low

probability. The algorithm is flexible in that it can maximize different utilities or notions of informativeness that the decision maker might have. For brevity, we focus on utility as defined by small interval width. BP leverages our theoretical findings by explicitly constraining the violations on the training data, and minimizing some loss function, ℓ, of the interval widths.

The appropriate loss function will vary between applications. We consider optimizing three loss functions of IW over $p(x)$: $\ell^{(1)}$ represents the desire to achieve a tight prediction bound on average, captured in the mean absolute interval width. $\ell^{(2)}$ penalizes the mean squared interval width, placing a higher penalty on points with very wide bounds. The third $\ell^{(\infty)}$ minimizes the worst (widest) interval by penalizing the maximum interval width.

We consider learning under the following conditions. Let $\phi : \mathcal{X} \to \mathbb{R}$ be the feature map corresponding to a reproducing kernel $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. For treatments $t \in \{0, 1\}$ and bounds $b \in \{l, u\}$ (lower/upper), let $f_b^t(x_i) := \langle \theta_b^t, \phi(x_i) \rangle + \rho_b^t$. In this setting, all three losses $(\ell^{(1)}, \ell^{(2)}, \ell^{(\infty)})$ are convex in θ. Let sample weights w_{t_i} be defined as in Definition 2.2.1, and define $\widetilde{w}_{t_i} := w_{t_i} / \sum_{j:t_j=t_i} w_{t_j}$. Finally, let $\Lambda(f)$ denote a term that measures complexity of f, e.g., the squared norm of parameters.

We describe two versions of BP: BP-D, a decoupled version where the bounds for the treated and control groups are fitted separately, and BP-C, a coupled version where the two are fitted simultaneously.

2.4.1 BP-D: decoupled treatment groups

First, we consider estimating bounds f_u, f_l on a single potential outcome $Y(t)$, independently of others. We minimize the weighted loss $\ell_{\widetilde{w}}^{(p)}(\boldsymbol{f})$ and require that the bounds be violated only with small probability over $p(x)$. We let the loss $\ell_{\widetilde{w}}^{(p)}(\boldsymbol{f})$ be defined by either the mean absolute interval width, $\ell_{\widetilde{w}}^{(1)}(\boldsymbol{f}) \sum_{i:t_i=t} \widetilde{w}_{t_i} |\mathrm{IW}_{\boldsymbol{f}}(x_i)|$, the mean squared interval width, $\ell_{\widetilde{w}}^{(2)}(\boldsymbol{f}) = \sum_{i:t_i=t} \widetilde{w}_{t_i} (\mathrm{IW}_{\boldsymbol{f}}(x_i))^2$, or the maximum

interval width, $\ell_{\widetilde{w}}^{(\infty)}(\boldsymbol{f}) = \sup_{i:t_i=t}(\text{IW}_{\boldsymbol{f}}(x_i))$.

$$
\begin{aligned}
\underset{\boldsymbol{f}=\{f_u,f_l\}}{\text{minimize}} \quad & \ell_{\widetilde{w}}^{(p)}(\boldsymbol{f}) + \alpha\Lambda(\boldsymbol{f}) \\
\text{subject to} \quad & \sum_{i:t_i=t} \widetilde{w}_{t_i} \max(y_i - f_u(x_i), 0) \leq \beta_u \\
& \sum_{i:t_i=t} \widetilde{w}_{t_i} \max(f_l(x_i) - y_i, 0) \leq \beta_l \\
& f_l(x_i) \leq f_u(x_i) \ , \forall i : t_i = t \ .
\end{aligned}
\tag{2.7}
$$

Note that the constraints are defined with respect to the magnitude of the violations, which does not immediately translate into a specific FCR. We address this issue in section 2.4.3. Problem (2.7) can be solved separately for the two treatment groups, as is done in two-learners or the treatment variable could be added in as a feature and the two treatment groups can be jointly trained, as is done in single-learners (Künzel et al., 2019). Next, we highlight some important characteristics of this estimator.

1. BP-D minimizes the lower bound in Theorem 2.3.2. Note that BP-D is specified over the set of linear functions with kernel defined feature spaces. With Λ defined as the 2-norm of the vector θ, and because of the last constraint ($f_l \leq f_u$), with high probability the functions returned by BP-D fall within the set of functions defined in definition 2.3.7, and hence theorem 2.3.2 is applicable here. Recall that theorem 2.3.2 states that for the estimated functions to be optimal, A, B, and D need to be minimized while γ needs to be maximized. Problem (2.7) directly minimizes the A, B (for $p = 1, 2, \infty$ depending on ℓ) and D. As for γ: suppose we fix the bias to be $\tilde{\rho}_b^t$, then $\gamma_b^t = \tilde{\rho}_b^t - \rho_b^t$, where the latter is the bias returned by solving problem (2.7). Because problem (2.7) minimizes ρ_b^t, it maximizes γ_b^t for a fixed $\tilde{\rho}_b^t$. Ideally, we would not fix $\tilde{\rho}_b^t$ in advance, but let it be decided by the data. We address this issue in section 2.4.3.

2. BP-D with $\ell^{(1)}$-loss is equivalent to quantile regression. When mini-

mizing the mean absolute interval width, our problem reuces to a quantile regression with non-crossing constraints (Takeuchi et al., 2006) of quantiles q and $1-q$ for for some choice of $q \in (0, .5)$.

Theorem 2.4.1. *Assume that* (2.7) *is strictly convex and has a strictly feasible solution. Then, for any fixed quantile* $q \in (0.5, 1)$*, there are parameters* $\beta_u, \beta_l \geq 0$ *such that the minimizers* f_u^*, f_l^* *of* (2.7) *with absolute loss and the minimizers of the quantile loss for quantiles* $(q, 1-q)$*, with non-crossing constraints, are equal.*

A proof is given in the appendix.

BP-D allows us to learn reliable and informative bounds but it does not make use of the "unlabeled" data from the opposite treatment group. This is addressed next.

2.4.2 BP-C: coupled treatment groups

In the coupled problem, we make use of samples from the counterfactual treatment group in two ways. First, we apply constraints that ensure that the lower and upper bounds do not cross for counterfactual outcomes. Second, the loss functions are defined with respect to the full marginal distribution of subjects (including counterfactual treatment assignments). We define the coupled version of the mean absolute loss $\ell^{(1)} = \sum_{i=1}^{n} \sum_{t=0}^{1} \widetilde{w}_{t_i} |\mathrm{IW}_{\boldsymbol{f}^t}(x_i)|$, mean squared interval width, $\ell^{(2)} = \sum_{i=1}^{n} \sum_{t=0}^{1} \widetilde{w}_{t_i} \mathrm{IW}_{\boldsymbol{f}^t}(x_i)^2$, and maximum interval width, $\ell^{(\infty)} = \sup_{i=1}^{n} \sum_{t=0}^{1} \mathrm{IW}_{\boldsymbol{f}^t}(x_i)$. The coupled problem becomes:

$$
\begin{aligned}
\underset{\{\boldsymbol{f}^t = \{f_u^t, f_l^t\}\}}{\text{minimize}} \quad & \ell_{\widetilde{w}}^{(p)}(\boldsymbol{f}^0, \boldsymbol{f}^1) + \alpha \cdot (\Lambda(\boldsymbol{f}^0) + \Lambda(\boldsymbol{f}^1)) \\
\text{subject to} \quad & \sum_{i:t_i=t} \widetilde{w}_{t_i} \max(y_i - f_u^t(x_i), 0) \leq \beta_u, \forall t \\
& \sum_{i:t_i=t} \widetilde{w}_{t_i} \max(f_l^t(x_i) - y_i, 0) \leq \beta_l, \forall t \\
& f_l^t(x_i) \leq f_u^t(x_i) \ , \forall t, i : t_i = t \ . \qquad (2.8)
\end{aligned}
$$

Given the overlap assumption (stated in Assumption 2.2.1), this encourages the counterfactual outcome intervals to be small even if the corresponding treatment assignment is not observed. By coupling the two objectives, we allow information to be shared between the treated and non-treated populations in a semi-supervised way. We caution, however, that in the absence of overlap, the coupled loss might be overly optimistic in regions of non-overlap, returning intervals that do not cover the true data. With f_l, f_u linear in the representation ϕ and $\Lambda(f)$ defined as the L2 norm of the function weights, expressions (2.7) and (2.8) are both convex programs that can be readily solved by a general solver. Our code is available at <`github.com/mymakar/bpo.git`>.

2.4.3 Cross-Validating BP

BP constrains the magnitude of the violations rather than the FCR directly. This allows the algorithm to directly utilize the theory and makes the optimization problem easier. The disadvantage is that the magnitude of violations does not directly translate into a specific FCR. We address this issue by designing a cross-validation algorithm that picks the hyperparameters of the model to achieve a required FCR, ν.

BP-C/D requires a regularization parameter, α, a level of tolerance to violations, $\beta_{u,l}$, and σ, which controls the kernel (e.g., the length scale for Gaussian kernels or the polynomial degree for polynomial kernels). Suppose that we solve problem (2.7) or (2.8) and get some estimate for the bias $\tilde{\rho}_b^t$, we specify an additional parameter $\gamma > 0$, and take the final estimate $\rho_l^t := \tilde{\rho}_l^t - \gamma$ and $\rho_u^t := \tilde{\rho}_u^t + \gamma$. This allows us to set γ based on the data rather than specify it *a priori*.

The algorithm takes as an input the training data, ν, ℓ, the required loss to minimize, and M, the set of hyperparameters to consider. We then split the data into training and validation. For each set of parameters indexed by $m = [1, \ldots, M]$, we use the training set to solve problem (2.7) or (2.8). We estimate $\hat{\nu}_m$ and $\widehat{\ell}_m$, the

FCR and loss corresponding to m on the held-out set. We discard of all the hyperparameters with a corresponding $\hat{\nu}_m > \nu$, and define $M' = \{m : \hat{\nu}_m \leq \nu\}$. We set the optimal hyperparameters $m^* := \min_{m \in M'} \widehat{\ell}_m$. The procedure is summarized in Algorithm 2.4.1. Let Ω denote a set of candidate hyperparameters. Suppose we have M possible hyperparameters, cross-validating BP proceeds as follows:

Algorithm 2.4.1: BP K fold cross-validation for M sets of hyperparameters, and required FCR = ν

Input: $\mathcal{D} = \{x_i, t_i, y_i, w_i\}$, p, ν, $\{\Omega\}^M$
Output: Ω^*
for $m = 1$ ***to*** M **do**
  for $k = 1$ ***to*** K **do**
    Split $\mathcal{D}$ into $\mathcal{D}^k_{\text{train}}$, $\mathcal{D}^k_{\text{validate}}$
    Use $\mathcal{D}^k_{\text{train}}$ to solve problem (2.7) or (2.8)
    Estimate $\hat{\nu}^{(k,m)}$, and $||\widehat{\text{IW}}||_p^{(k,m)}$ on $\mathcal{D}^k_{\text{validate}}$, using the weights w_i
  end
  Compute the average metrics over the K folds; $\hat{\nu}^{(m)} = K^{-1} \sum_k \hat{\nu}^{(k,m)}$, and $||\widehat{\text{IW}}||_p^{(m)} = K^{-1} \sum_k ||\widehat{\text{IW}}||_p^{(k,m)}$
end
Define $M' = \{m : \hat{\nu}^{(m)} \leq \nu\}$
Set $\Omega^* := \min_{m \in M'} ||\widehat{\text{IW}}||_p^{(m)}$

2.5 Experiments

We compare our model to other interval estimation methods. First is classical confidence-interval based approaches. We use **XX-CCI** to refer to this approach, where XX will be replaced by the name of the base model (e.g., if it is a Gaussian Process, we use GP-CCI). Though popular, classical confidence intervals are known to have poor coverage in finite samples (Sargent et al., 1992; Lei et al., 2018). Conformal intervals, the second interval estimation method we compare against, were introduced as an alternative with better finite sample coverage (Lei et al., 2018). Conformal intervals are estimated by splitting the training data into two parts. The first part is used to train the outcome model, where parameters are picked via the

usual cross-validation techniques. We estimate the residuals on the second subset of the training data. If the required FCR is q, we take the $1 - q^{th}$ quantile of the residuals to be a "shifting" parameter (akin to γ in our setting). The conformal intervals for a test sample are taken to be the estimated outcome $\pm$ the shifting parameter. We use **XX-CI** to refer to this approach. Finally, we introduce γ-intervals, which we refer to as **XX-γ**. Similar to conformal intervals, we split the data into two, fitting the best model on the first half and then picking the smallest shifting parameter γ that achieves the required FCR on the second half. We use **BP-V-Lp** to refer to our models, where V refers to the D (decoupled) or C (coupled) version and Lp refers to the norm of the loss (1, 2, or ∞). Recall that the 1-norm is similar to quantile regressions (QR) (by theorem 2.4.1).

We evaluate the performance of our models and the baselines on a held-out test set with respect to two criteria: the achieved FCR, as defined in equation (2.2) and the utility as measured by the mean IW and the max IW, as defined in equation (2.6). Additional cross-validation details for our model and the baselines are included in the appendix.

We analyze the performance of BP as compared to baselines in multiple settings. We highlight the following settings where we expect BP to outperform baselines.

1. **Residual distribution assumptions.** Most baselines make restrictive assumptions about the distributions of the residuals. When such assumptions break, the resulting intervals are no longer tight or do not correctly cover the outcomes. We briefly outline such assumptions:

   (a) **Symmetry**. This assumption states that in order to get a 5% FCR, we need to ensure that the lower and upper bounds are violated by at most 2.5% each. In some cases, the tightest bounds would be achieved by non-symmetrical bounds, e.g., the lower bound is violated by 1% whereas the upper bound is violated by 4%. Violations to the symmetry assumption occur, for example, when the model is misspecified, which leads to biased

estimates. In that case, tight bounds should reflect the direction of bias: if the estimates are biased downwards (meaning lower than the true value), it is more important that the upper bounds are not violated, whereas violations to the lower bound are more permissible (since the estimate itself is a lower value than the true outcome). We empirically analyze this setting in section 2.5.1.1.

(b) **Well-behaved residual distribution**: This assumption states that the residuals concentrate around a single, central value. Such an assumption is also violated when there is model misspecification, or if the outcome noise is heteroskedastic. We empirically analyze this setting in section 2.5.1.2.

2. **Small samples.** Most baselines estimate the conditional expectation $\mathbb{E}[Y(t) \mid x]$ as accurately as possible, then add, and subtract some value from that estimate to get the bounds. When the observed data is not enough to accurately estimate $\mathbb{E}[Y(t) \mid x]$, these bound estimates are not reliable. We empirically analyze this setting in section 2.5.2.

When there is enough data to accurately estimate $\mathbb{E}[Y(t) \mid x]$, we expect BP to perform as well as the baselines. We present this experiment setup in section 2.5.3.

2.5.1 IST data

We begin with a simple illustrative example that highlights the strengths of BP vis-a-vis baselines and the properties of different utility functions in a practical setting. We aim to answer the following: (1) How do different losses reflecting different notions of utility affect the estimates? (2) How does the coupled objective make use of counterfactual data?

We study the task of a physician deciding whether or not to prescribe Heparin, an anticoagulant, to reduce the risk of Ischemic and Hemorrhagic strokes. Patients with

an elevated risk of forming blood clots can reduce their risk of an Ischemic stroke by taking Heparin. However, some patients experience excessive bleeding if placed on Heparin increasing their risk of a Hemorrhagic stroke. In this setting, to make an informed decision, the physician only needs to know if the INR under treatment falls within the healthy range of 2–3 as described in the introduction. The exact value of INR provides little additional insight.

We use data from a randomized control trial measuring the effects of Heparin (International Stroke Trial Collaborative Group, 1997). We restrict our analysis to the patients who received Heparin (treatment, $n_1 = 4530$) or no anticoagulant (control, $n_0 = 4534$). To introduce confounding, we drop 70% of the older (age > 70), untreated population thus ensuring a strong correlation between age and receving the treatment. Note that the original distribution of age in the trial is skewed, with a mean of 71.8 and a skewness of -0.79, which means that young patients are underrepresented. Figure 2-2 in shows the distribution of ages for the treated and control groups in the training set.

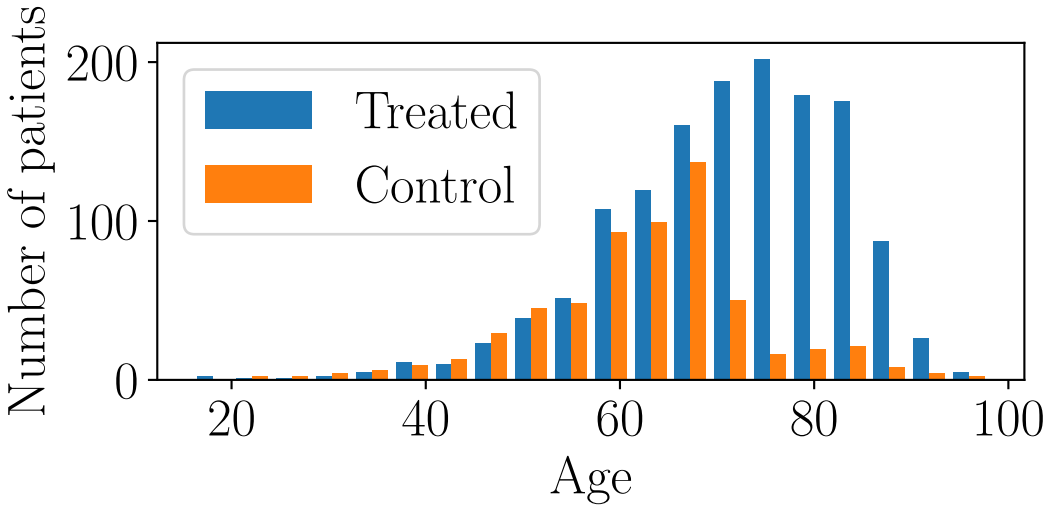

Figure 2-2: Distribution of data in the IST experiment

Because INR was not measured in the data, we simulate $\mathbb{E}[Y_i(1) \mid age_i]$, and $\mathbb{E}[Y_i(0) \mid age_i]$ according to two different scenarios described later. For both scenarios, we fit a

kernel regression with a linear kernel (**KR**) for the baselines. We repeat our simulation 20 times and report averages. In each simulation, we randomly sample 3000 patients for training and validation and 3000 held out for testing. Following Chernozhukov et al. (2016), we use half the training data to estimate the nuisance parameter, that is the propensity scores, and the other half to fit the potential outcomes. For propensity scores, we fit a logistic regression. We pick the regularization parameter for the propensity score model and all the response surface models via 3-fold cross-validation as described in the appendix. For all experiments, we set the required FCR to be ≤ 0.01, i.e., $\leq 1\%$.

All models in the IST experiments presented in sections 2.5.1.1, and 2.5.1.2 rely on inverse propensity score weighting (i.e., weighting by w_i) during training. All hyperparemeters for all models are picked based on weighted performance metrics evaluated on the held out validation set; for both kernel regression models, we pick the regularization penalty based on the weighted loss computed on the held out validation set. We pick the value of γ for the KR-γ model based on the weighted FCR. For our models, we pick the hyperparameters based on the weighted FCR, and the weighted interval widths as described in algorithm 2.4.1.

2.5.1.1 Model misspecification

Here, we study a scenario where the symmetry assumption is violated because of model misspecification.

We simulate the INR under treatment according to $\mathbb{E}[Y_i(1) \mid age_i] = S(-5, age'_i) + 2.5 + \varepsilon$, where $S(a, x)$ denotes the sigmoid function with coefficient a, age' is the age rescaled between -10, 10 and $\varepsilon \sim$Gaussian(0, 0.1). This setup ensures that the majority of the population falls within the normal range if treated, while the few patients younger than 60 have high INR if treated. Similarly, the outcome under control is determined by $\mathbb{E}[Y_i(0) \mid age_i] = S(-5, age'_i - 4) + 1.5 + \varepsilon$. This reflects the setting where patients older than 70 (who are under-represented in the untreated

population) would have too low an INR if not placed on Heparin.

We assume that we are restricted to linear models for interpretability reasons. In this setting the models are inherently misspecified, which means that the residuals violate the symmetry and well-behavedness assumptions.

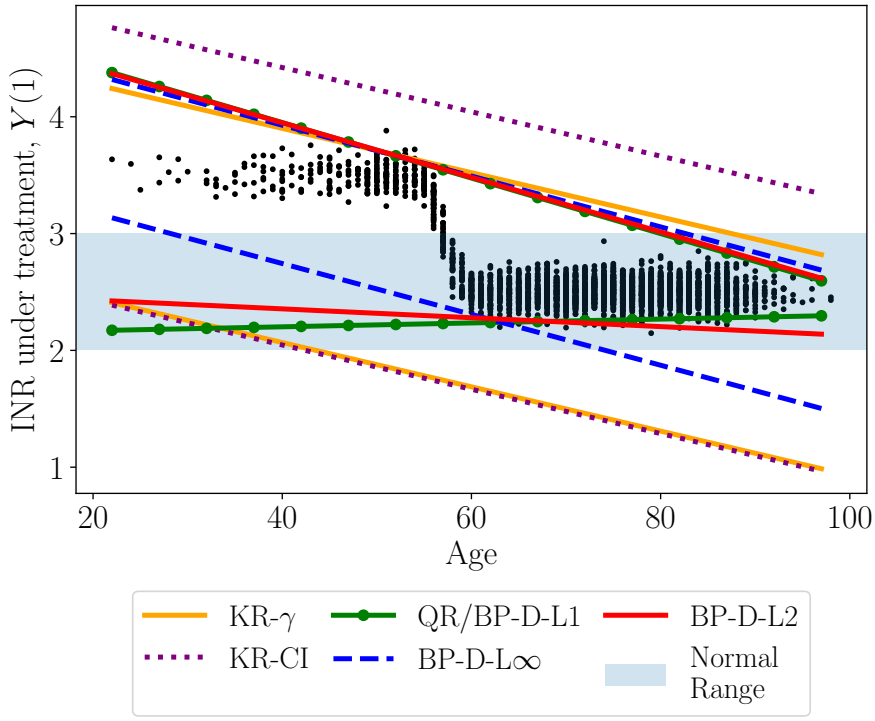

Figure 2-3: Comparing different loss functions: Plot shows results from a single simulation. Black dots show potential outcomes on the test set, lines show fitted values, and shaded region shows healthy range. We see that BP-D-L∞ is a "fair" objective, ensuring that the younger (≤ 60) population has intervals as tight as those for the older population. QR (equivalent to BP-D-L1) ensures intervals are tight for older population but returns wider intervals for the younger population. BP-D-L2 gives an estimate "in-between" the two objectives, penalizing large intervals more aggressively than QR/BP-D-L1. Baselines (KR-CI/KR-γ) return bounds that are loose for both populations.

Table 2.1 (top), and Figure 2-3 show the results. Most notably, KR-CI and KR-γ return loose estimates compared to BP/QR. This is because KR-CI assumes symmetry of the residuals, returning overly loose upper bounds. KR-γ implicitly assumes non-fat tailedness by shifting the estimates by the same constant for all individuals. More generally, the baselines fail because they aim to first estimate the outcome as best as possible, and then estimate the intervals post-training. Ultimately, the

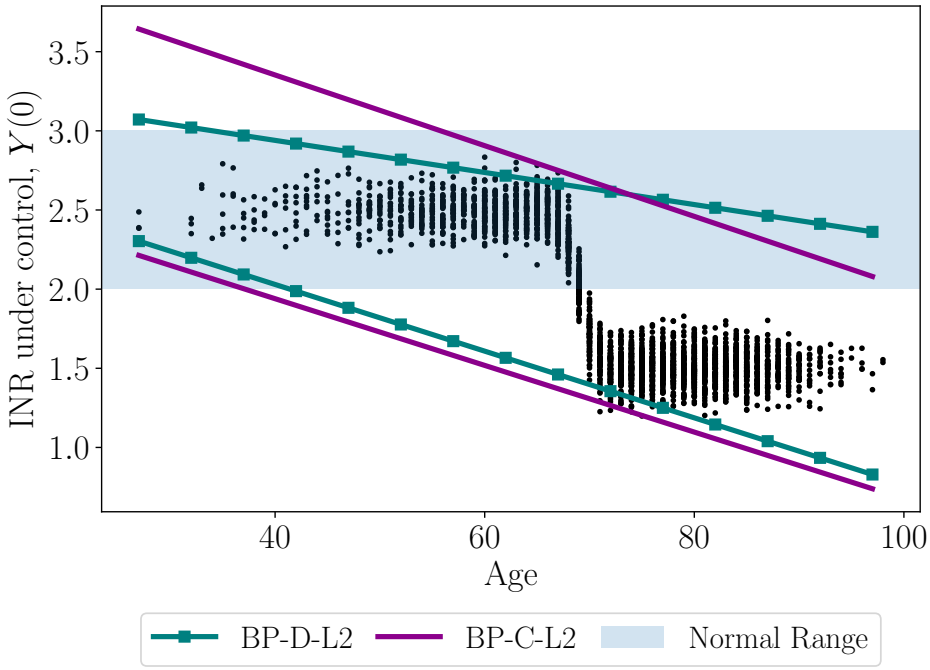

Figure 2-4: Decoupled and coupled versions. Plot shows results from a single simulation. Black dots show potential outcomes on the test set, lines show fitted values, and shaded region shows healthy range. We see that penalizing the counterfactual interval widths enables the coupled objective, BP-C-L2, to return a tighter fit for $Y(0)$ in the area where few untreated examples exist in the training data (age$>$ 70).

model is picked based on what reduces the mean squared error, not what reduces over/under-estimation.

In addition, table 2.1 (top) shows that BP-D-L∞ achieves the smallest maximum IW. BP-D-L2 and QR (equivalent to BP-D-L1) achieve the smallest mean IW, with the former achieving a smaller max IW. Figure 2-3 explains why. BP-D-L∞ achieves the smallest max IW since it penalizes large intervals in the younger population at the cost of fitting a wider interval for age $\geq$ 60. Such an objective is most appropriate when notions of fairness might be at play, such as if a physician wants to ensure that younger patients are never given abnormally large intervals compared to the older group. QR/ BP-D-L1 achieves a tight mean IW for the older population but does less well on the younger population. Such an objective is appropriate when we want estimates that are as tight as possible on average, even if that entails computing wide estimates for small subpopulations. BL-D-L2 falls between the two extremes of

Table 2.1: IST results. Table shows results averaged over 20 simulations, confirming conclusions from figures 2-3 and 2-4.

Model	FCR	Mean IW	Max IW
$Y(1)$ results			
BP-D-L2	0.007 (0.0036)	1.04 (0.05)	2.15 (0.19)
BP-D-L$_\infty$	0.007 (0.0037)	1.16 (0.06)	1.16 (0.06)
QR/BP-D-L1	0.007 (0.0043)	1.07 (0.09)	2.25 (0.26)
KR$-\gamma$	0.004 (0.0081)	1.96 (0.09)	1.96 (0.09)
KR-CI	0.0 (0.0)	2.41 (0.07)	2.41 (0.07)
$Y(0)$ results			
BP-C-L2	0.007 (0.0059)	1.35 (0.17)	1.62 (0.26)
BP-D-L2	0.005 (0.0051)	1.37 (0.13)	1.72 (0.2)

BP-D-L∞ and BP-D-L1/QR; its mean IW is slightly higher than that of BP-D-L1 (for the younger population) and lower than that of BP-D-L∞, its max IW is lower than that of BP-D-L1 but higher than that of BP-D-L∞. This is because the L2 loss penalizes large IWs more aggressively than L1.

Recall that here the physician's objective is to correct the patients' INR level without overshooting, hence ensuring that the patients' INR falls within the healthy range. A physician who prescribes Heparin only when they are certain that a patient's INR would fall in the normal range (i.e., both upper and lower bounds fall in the normal range) would not prescribe heparin to anyone if they rely on KR-γ, KR-CI, or BP-D-L∞ estimates. The latter has the advantage of providing tighter bounds for the younger patient group, whereas the former three also fails on that task.

Table 2.1 (bottom) shows that the decoupled version achieves a smaller mean and max IW compared to the coupled version, though the difference is not statistically significantly different. Figure 2-4 gives insight into the difference between the two versions. The coupled objective returns tighter intervals for the majority of the population, that is patients with age > 70, who are under-represented in the control group. This happens because the coupled objective has an incentive to minimize the interval width for older, untreated patients since a wider counterfactual interval for

the older treated patients is penalized, whereas the decoupled objective is unaware of these patients.

2.5.1.2 Heteroskedasticity

Here, we study a scenario where the well-behavedness assumption is violated because of heteroskedasticity.

We use the IST data, and for brevity focus only on the outcome under treatment, $Y(1)$. Specifically, we generate the outcome under treatment as $Y(1) = x^2 + \varepsilon$, where x is the age rescaled to fall between -2, 2, and ε_i is drawn from a Gaussian distribution with mean 0 and standard deviation = 0.1 if $x \leq 0$, and from a Gaussian distribution with mean 0 and standard deviation = $0.1 + x$ otherwise. We set the required FCR to be ≤ 0.01. Since our main aim is to analyze how the different models perform when when heteroskedasticity occurs, we focus only on tightness of bounds as an objective.

Table 2.2 shows the results averaged over 20 simulations. It shows that of all the models that achieve the required FCR, BP-D-L2 achieves the tightest intervals. Figure 2-5 shows why: neither BP-D-L2 and QR (equivalent to BP-D-L1) make assumptions about well-behavedness of the residual distribution. They give adaptive intervals, which are tight when the heteroskedastic noise is low, and loose when it is high. In addition, BP-D-L2 has a better performance than BP-D-L1 because it places a higher penalty on very large intervals. In this specific example, because the majority of individuals have age greater than 60, and hence fall in the area where the dispersion is high, the L2 penalty ends up achieving tighter intervals than the L1 loss.

2.5.2 Small ACIC data

Here, we study a setting where the data available for training might not be enough to accurately estimate $\mathbb{E}[Y(t) \mid x]$. We evaluate our approach in a challenging, high-

Table 2.2: IST heteroskedasticity results. Table shows results averaged over 20 simulations 2-5.

Model	FCR	Mean IW	Max IW
BP-D-L2	0.007 (0.005)	5.55 (0.56)	10.68 (2.35)
QR/BP-D-L1	0.006 (0.0031)	6.49 (0.96)	11.63 (2.37)
KR-γ	0.065 (0.0086)	3.98 (0.06)	3.98 (0.06)
KR-CI	0.007 (0.0052)	6.94 (0.69)	6.94 (0.69)

dimensional task: semi-simulated data from the Atlantic Causal Inference Conference Competition (Dorie et al., 2017). In this task, 58 variables were extracted from the Collaborative Perinatal Project, a study on pregnant women and their children. The treatment assignment and the response surfaces were simulated. We focus on the simulation with limited overlap and high heterogeneity where the treatment response surface is polynomial and the response surface is exponential (setting number 12). We sample 200 data points for the training/validation of the main models, and 1000 for our test set. We sample 1000 data point for training/validation of the propensity score models. Propensity scores are estimated using 3 fold cross-validation.

To fit the potential outcomes, we use an RBF kernel for our BP/QR models. We also use an RBF kernel for the kernel regression models. We only present KR-CI, excluding KR-γ since it performs comparably to KR-CI. In addition, we include single-learners (Künzel et al., 2019) with Gaussian processes as the base-estimators (GP), and Bayesian Additive Regression Trees (BART; Hill (2011)). For the latter 2 models, we compute the classical confidence intervals (GP-CCI, and BART-CCI), and a variant of the γ-intervals (GP-γ, and BART-γ). Here, γ is used as a scaling rather than a shifting parameter; for an estimated outcome $\hat{y}$, and estimated standard deviation $\hat{\kappa}$, the lower/upper bounds are estimated as: $\hat{y} \pm \gamma \cdot \hat{\kappa}$, and the optimal γ is picked based on cross-validation as described previously.

All models except BART rely on inverse propensity score weighting (i.e., weighting by w_i) during training, and validation[1]. Similar to the IST experiments, we use the

[1]We note that training BART without inverse propensity score weighting is consistent with previous implementations of BART in causal settings e.g., in Hill (2011); Shalit et al. (2017)

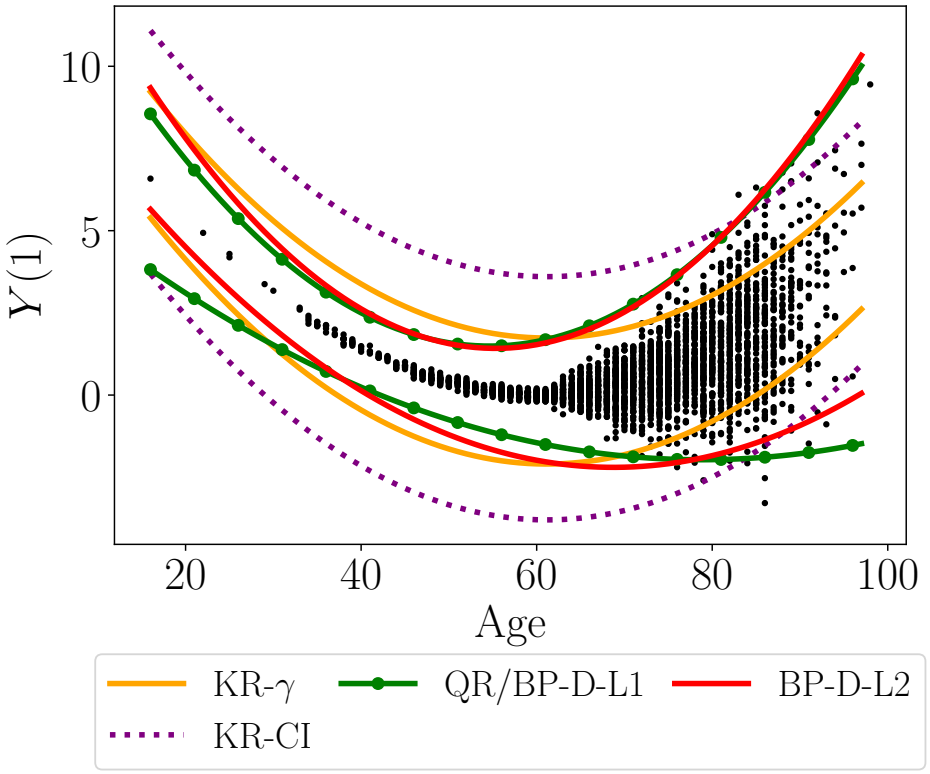

Figure 2-5: IST heteroskedasticity results. Plot shows results from a single simulation. Black dots show potential outcomes on the test set, lines show fitted values. The plot show that BP-D-L2 and QR (equivalent to BP-D-L1) are the only ones that are able to fit *adaptive* intervals (wider where there is high heteroskedasticity). BP-D-L2 achieves the tightest intervals on average.

weighted FCR to pick the value of γ for BART-γ, GP-γ, and KR-γ. Because BART-CCI does not utilize the propensity scores at any part of the training and validation pipeline, we use all 1200 samples to train the BART-CCI models.

We focus on getting the tightest bounds, so we only present results from BP-C-L2. We measure the performance of the models at required FCR = {0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2 }.

In this setting, the small sample size makes it hard to get an accurate estimate of the potential outcomes, which may belong to a complex function class. This can be thought of as a "forced" model misspecification since the limited data does not afford us the ability to fit the true function, and limits us to simpler function classes. This is once again, a setting where we expect our models to outperform baselines that make

strong assumptions about the residuals.

Figure 2-6 shows the mean achieved FCR on the $x-$ axis, and the mean IW on the $y-$axis for our model and baselines averaged over 20 simulations. First, we see that the mean IWs for all the models decrease as the achieved FCR increases. This confirms our theoretical findings of a trade-off between confidence that the bounds cover the potential outcomes and complexity of the function class; lower required FCRs (i.e., higher confidence that the bounds cover the true date) are associated with simpler function classes, which sacrifices accuracy, leading to higher mean IW. Second, we see that our models achieve interval widths that are tighter than all other kernel-based methods, and comparable to BART at every value of achieved FCR. Note, however, figure 2-7 shows that our models achieve smaller violation compared to BART. This implies that our models are better able to exploit the trade-off between confidence and complexity.

Results from GP-CCI, and BART-CCI are excluded from the plots, and presented in the section A.5.2 in the appendix since they achieve very large violations. Unlike the IST setting, in this high dimensional setting, diagnosing why classical confidence interval methods fail to correctly cover the potential outcomes is difficult. Comparing the results from the CCI models to their γ counterparts point to the notion that CCI models might be underestimating the conditional variance. Another conjecture to explain the poor coverage by CCI methods is that because of model misspecification, or bias due to imperfect estimation in finite samples the true potential outcomes lie outside of the estimated confidence interval as illustrated in figure 2-8. We note that these findings conform with previous studies which show that CCI methods tend to have poor coverage in finite samples (Sargent et al., 1992; Lei et al., 2018).

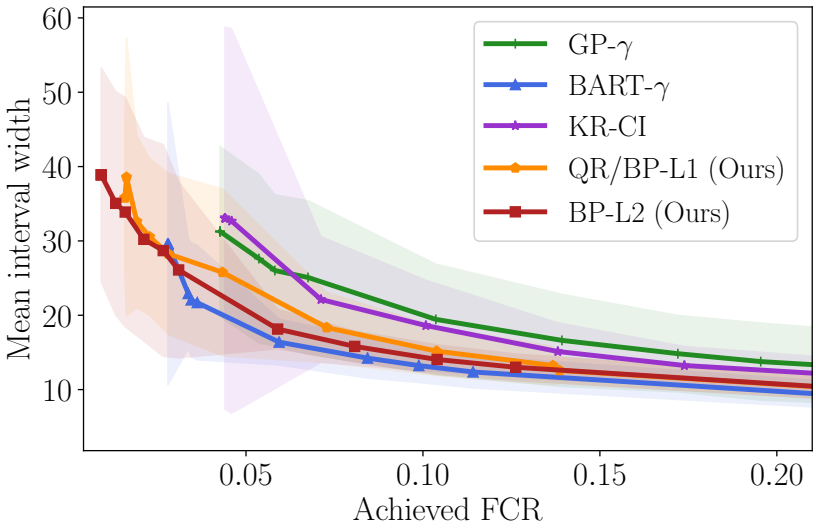

Figure 2-6: Small ACIC data results: comparing tightness of estimated intervals: Plot shows the mean interval width for different values of the achieved FCR on a held-out test set, averaged over 20 simulations. Our approach (BP) achieves a mean interval width comparable to the best performing model (BART), and better than other kernel-based methods.

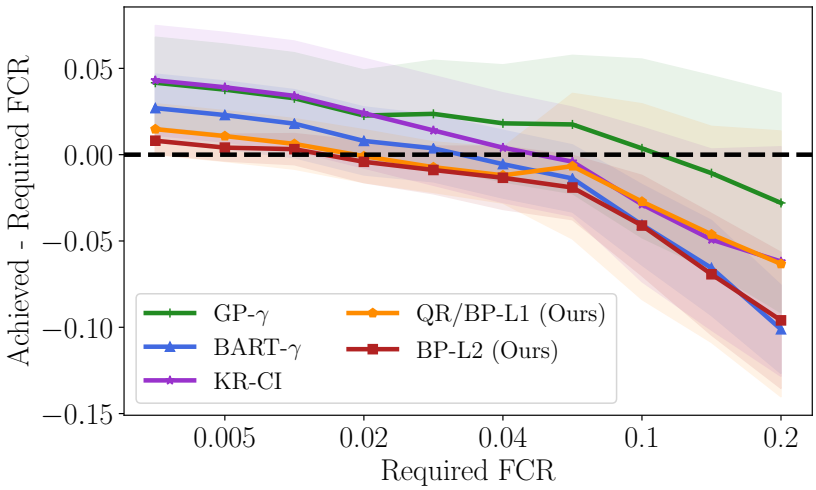

Figure 2-7: Small ACIC data results: comparing violation to the required FCR. Plot 2-7 shows the violation of the required FCR (= achieved - required) at different values of required FCR, averaged over 20 simulations. Models above the dotted black line are in violation of the required FCR. Our approach (BP) achieves lower violation of the required FCR.

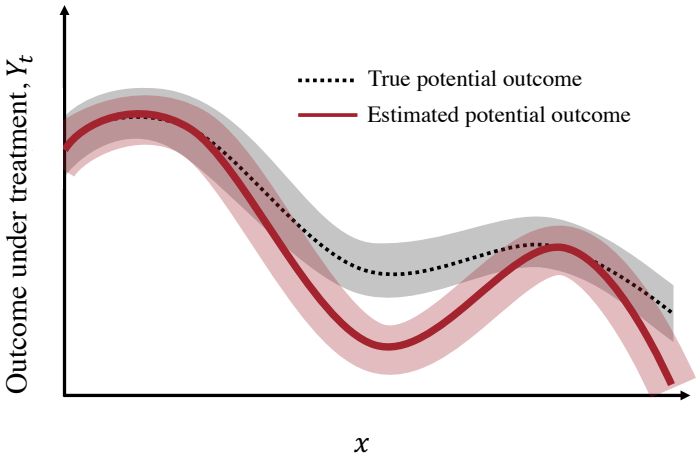

Figure 2-8: Illustration of a possible failure mode for CCI methods. Dotted grey line shows the true potential outcome, grey cloud shows Gaussian noise. Red line shows the estimated potential outcome which is imperfect due to finite sample error or model misspecification. Red cloud shows the estimated confidence interval. The true potential outcome falls outside of the estimated confidence interval for some subpopulations.

2.5.3 Large ACIC data

We repeat the same analyses presented in section 2.5.2 but here we consider a larger sample size than that presented in the previous section. Instead of sampling $n = 200$, we sample 1000 data points for training and validation of the main models[2]. In this setting, all models are better able to fit the true outcomes since the larger sample size affords us the ability to fit more complex models. Figures 2-9, and 2-10 show the results. Once again we see that our models outperform all kernel based methods. Here we see that BART-γ achieves a tighter interval width than our model for the same level of FCR violation. This highlights the strength of tree based models in that they fit highly adaptive "kernels". We note that the typical BART intervals, computed using the classical confidence intervals achieved high violations of the required FCR.

Similar to the small sample setting, results from GP-CCI, and BART-CCI are ex-

[2]Similar to the small ACIC data analysis, BART-CCI uses all 2000 data points for training since it does not rely on inverse propensity score weighting at any point in the training, and validation pipelines

cluded from the plots, and presented in the section A.5.2 in the appendix since they achieve very large violations.

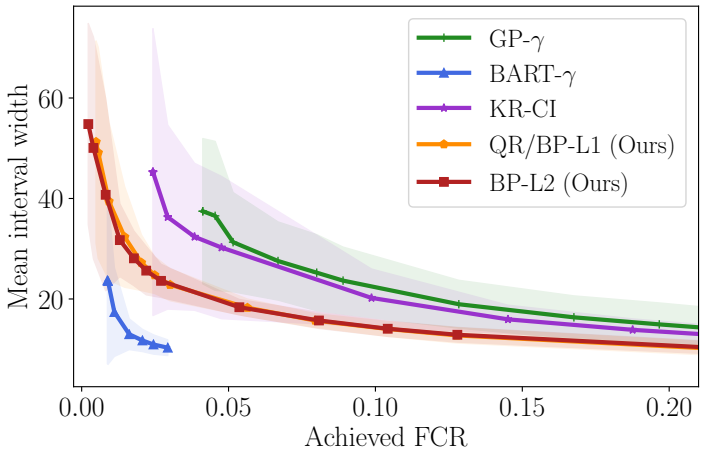

Figure 2-9: Large ACIC data results: comparing tightness of estimated intervals: Plot shows the mean interval width for different values of the achieved FCR on a held-out test set, averaged over 20 simulations. Our approach (BP) outperforms all kernel-based methods in terms of mean interval width. BART with γ-intervals, a tree based method returns tighter interval widths compared to our approach, at a comparable achieved FCR (see plot 2-10).

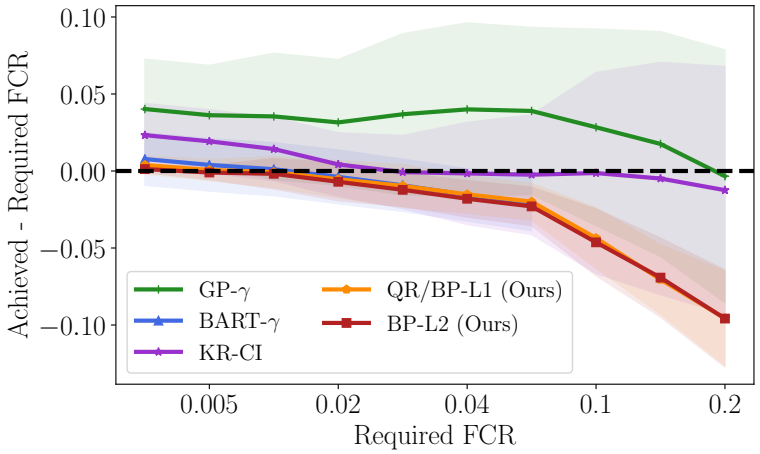

Figure 2-10: Large ACIC data results: comparing violation to the required FCR. Plot shows the violation of the required FCR (= achieved - required) at different values of required FCR, averaged over 20 simulations. Models above the dotted black line are in violation of the required FCR. Our approach (BP) achieves lower violation of the required FCR.

2.6 Summary

In this chapter, we establish that the sample complexity of learning bounds on potential outcomes depends on how confident we wish to be that the bounds cover the true potential outcomes. For applications where it is sufficient to have reliable bounds on the potential outcomes, and the outcomes are complex functions, our findings indicate how to simplify the learning problem. Based on these findings, we introduced an algorithm that maximizes an objective, specified by the user, subject to constraints that guarantee validity of the bounds with high probability. Using semi-synthetic data, we showed that our method outperforms baselines, estimating tight prediction intervals without violating a required level of false coverage rate.

We note that our approach would be favorable even in the context of a randomized control trial, where the treatment assignment is not biased. The core discovery that we present here is not predominantly addressing bias issues which arise in observational data. Rather, we emphasize that one of the main contributions presented in this chapter is exploring a novel trade-off between utility of a learned estimator and its statistical complexity, which depends on its credibility.

Chapter 3

Causally-motivated Shortcut Removal Using Auxiliary Labels

In this chapter, we use ideas from causal inference to develop efficient and robust predictive models. We show that our causally-motivated regularization scheme leads to predictors with low generalization error within, and outside of the training distribution.

3.1 Background

Despite their immense success, predictors constructed from deep neural networks (DNNs) have been shown to lack robustness under distribution shift (Beery et al., 2018; Ilyas et al., 2019; Azulay and Weiss, 2018; Geirhos et al., 2018), especially naturally occurring distribution shifts (Taori et al., 2020). One particular mechanism for this brittleness is *shortcut learning* (Geirhos et al., 2020). Shortcut learning occurs when a predictor relies on input features that are easy to represent (i.e., shortcuts) and predictive of the outcome in the training data, but do not remain predictive when the distribution of inputs changes. For example, a DNN trained for image

classification could exploit correlations between the foreground and background in the training distribution, and use a representation of the background as a shortcut to predict the foreground (Beery et al., 2018; Sagawa et al., 2019). This can occur even if the foreground object alone is sufficient to achieve optimal predictive performance (Nagarajan et al., 2020; Sagawa et al., 2020a).

Throughout this chapter, we use the example of classifying the foreground object as land or water bird. The two classes are visually distinct but the majority of the former often appear on land backgrounds, and the latter on water backgrounds. DNNs that exploit shortcuts could achieve strong performance on unseen instances from the training distribution, but would fail if the foreground object and background were correlated differently in the test distribution (e.g., if water birds appeared on land backgrounds).

In this chapter, we consider the problem of learning a performant predictor whose risk is invariant to interventions that change the correlations between irrelevant factors and the main label. Ideally, such a predictor would rely exclusively on input features that are invariant to irrelevant factors. However, identifying such invariant input features in the standard supervised learning setup is difficult, for the same reason that shortcut learning is successful: in learning setups where there are many distinct ways to construct predictors that perform well on held-out data (i.e., when the learning problem is *underspecified* (D'Amour et al., 2020)), the influence of correlated factors is difficult to disentangle without additional supervision (Locatello et al., 2019).

For this reason, we focus on a modified setting where we are also given an auxiliary label that gives information about the irrelevant factor. Such labels often appear in the form of metadata associated with training data—for example, labels of the background—but are often not available at test time. In this setting, we propose an approach that exploits this auxiliary label to construct a predictor whose risk is approximately invariant across a well-defined family of test-distributions. Our method makes use of two tools from causal inference in combination: (1) weighting

the training data to mimic an idealized population, and (2) enforcing an independence implied by the causal Directed Acyclic Graph (DAG) in that idealized population. While each of these approaches has been applied separately, we show here through both theoretical arguments and empirical analysis that these methods are particularly effective when applied together.

Our methodological contributions can be summarized as follows:

1. We suggest an approach that relies on auxiliary labels to discourage shortcut learning. We specify a set of distribution shifts across which a robust model is risk-invariant.

2. We give a theoretical justification of our approach, highlighting that in some scenarios it yields models that have a lower generalization error than typical regularization schemes. We also show that our approach is robust to a set of distribution shifts.

3. We empirically validate our theoretical findings using a semi-simulated benchmark, showing our approach has favorable in- and out-of-distribution generalization properties.

4. We compare against baselines that ablate each part of our approach to show that their combination yields more performant, stable training.

The remainder of the chapter is organized as follows. In section 3.2, we formally introduce our objective. We also discuss important properties of the unconfounded distribution, where the main label and the auxiliary label are independent. In section 3.3, we present our main approach, and briefly state the main claims that guide the design of our approach. We revisit these claims in section 3.4 with a greater detail, giving theoretical justification for each. In section 3.5 we present our empirical analysis. We conclude the chapter with a summary in section 3.7.

3.2 Preliminaries

3.2.1 Setup

We consider a supervised learning setup where the task is to construct a predictor $f(\mathbf{X})$ parameterized by weights $\mathbf{w}$ that predicts a label Y (e.g., foreground object) from an input $\mathbf{X}$ (e.g., image). In addition, we have an auxiliary label V (e.g., background label) available at training time that labels a factor of variation along which we hope the model will exhibit some invariance. Throughout, we will use capital letters to denote variables, and small letters to denote their value. Our training data consist of tuples $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$ drawn from a source training distribution P_s. We restrict our focus to the case where Y and V are binary and f is a classifier. Specifically, we will consider functions f of the form $f = h(\phi(\mathbf{x}))$, where ϕ is a representation mapping and h is the final classifier.

We assume that P_s has a generative structure shown in Figure 3-1, in which the inputs X are generated by the labels (Y, V). We assume that the labels Y and V are correlated, but not causally related; that is, changing the value of V does not imply a change in the value of Y, and vice versa. Such correlation often arises through the influence of an unobserved third variable such as the environment from which the data is collected. We represent this in Figure 3-1 with the dashed bidirectional arrow.

In addition, we assume that there is a sufficient statistic $\mathbf{X}^*$ such that Y only affects $\mathbf{X}$ through $\mathbf{X}^*$, for which the sufficient reduction $\mathbf{X}^* = e(\mathbf{X})$ is unknown. For this reason, we denote $\mathbf{X}^*$ as unobserved in Figure 3-1. The fact that $\mathbf{X}^*$ is a function of $\mathbf{X}$ only implies that $\mathbf{X}^*$ is invertible, i.e. for all $\mathbf{X}$, $\mathbf{X}^*$ can be exactly recovered from $\mathbf{X}$. We state that formally in the following assumption

Assumption 3.2.1. *(Invertability) There exists some function e such that* $\mathbf{X}^* = e(\mathbf{X})$ *for all* $\mathbf{X}$.

3.2.2 Risk Invariance

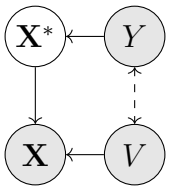

Figure 3-1: DAG depicting the setting we consider in this chapter. The main label Y and auxiliary label V generate observed input $\mathbf{X}$, but Y only affects $\mathbf{X}$ through the sufficient statistic $\mathbf{X}^*$.

We define the generalization risk of a function f on a distribution P as $R_P = \mathbb{E}_{X,Y\sim P}[\ell(f(X), Y)]$, where ℓ the logistic loss.

We focus on obtaining an accurate *risk invariant* predictor, with the property that the risk is invariant across a family of target distributions P_t that can be obtained from P_s by interventions on the causal model in Figure 3-1. Specifically, we consider interventions on the confounding relationship between Y and V that keep the marginal distribution of Y constant. Each distribution in this family can be obtained by replacing the source conditional distribution $P_s(V \mid Y)$ with a target conditional distribution $P_t(V \mid Y)$:

$$\mathcal{P} := \{P_s(\mathbf{X} \mid \mathbf{X}^*, V)P_s(\mathbf{X}^* \mid Y)P_s(Y)P_t(V \mid Y)\}. \tag{3.1}$$

This family allows the dependence between Y and V to change arbitrarily.

We make the following overlap assumption on the source distribution:

Assumption 3.2.2. *(Overlap)* $P_s(V)P_s(Y) \ll P_s(V, Y)$

Given the family $\mathcal{P}$, we define the set of risk invariant predictors to be all predictors that have the same risk for all $P_t \in \mathcal{P}$,

$$\mathcal{F}_{\text{rinv}} = \{f : R_{P_t}(f) = R_{P_t'}(f) \quad \forall P_t, P_t' \in \mathcal{P}\}.$$

An optimal risk-invariant predictor f_{rinv} has the property

$$f_{\text{rinv}} \in \arg\min_{f \in \mathcal{F}_{\text{rinv}}} R_{P_t}(f) \text{ for any } P_t \in \mathcal{P}.$$

Risk invariance is an appealing property because guarantees about the performance of the predictor f_{rinv} derived under one distribution can be adapted to other distributions in $\mathcal{P}$.

3.2.3 The Unconfounded Distribution P°

Within the family of distributions $\mathcal{P}$, we pay special attention to the *unconfounded distribution* $P^\circ \in \mathcal{P}$ where $P^\circ(V \mid Y) := P_s(V)$. Under P°, $Y \perp\!\!\!\perp V$ and the dashed bidirectional arrow in Figure 3-1 can be dropped. Both our methodological approach and theoretical analysis revolve around mapping the problem of learning a risk invariant predictor under P_s to the problem of learning an optimal predictor under P°.

P° has two useful properties that are revealed by the DAG in Figure 3-1: (1) under the unconfounded distribution P°, the optimal predictor (with some abuse of notation) would take the form $f(\mathbf{X}^*)$, and (2) for any predictor of the form $f(\mathbf{X}^*)$, the joint distribution $P(f(\mathbf{X}^*), Y)$ (and thus the risk) is invariant across the family $\mathcal{P}$. Together, these imply that the optimal risk-invariant predictor $f_{\text{rinv}}(\mathbf{X}^*)$ is the optimal predictor under P°. We state this formally in the following proposition.

Proposition 3.2.1. *Under P°, the Bayes optimal predictor is (i) only a function of $\mathbf{X}^*$, and (ii) an optimal risk-invariant predictor f_{rinv} with respect to $\mathcal{P}$.*

Proof is in the appendix. This motivates our approach to design an objective that enables efficient estimation of the optimal predictor under P°, even when the training data $\mathcal{D}$ are drawn from a different distribution, P_s.

3.3 Approach

Here, we describe our approach to learning an optimal risk-invariant predictor $f_{\text{rinv}}(\mathbf{X})$ from training data $\mathcal{D} \sim P_s$. Our approach relies on the following claims, which follow from the causal structure of the problem:

1. When the source distribution is the ideal distribution P°, enforcing an independence implied by the causal DAG leads to efficient estimation (proposition 3.4.3). It also shrinks the gap between the error on the target distribution P_t and P° (proposition 3.4.6)

2. From any source distribution P_s, we can efficiently learn the optimal predictor under P° by enforcing the independence implied by the DAG, and applying appropriate weights to examples (proposition 3.4.5).

We will formally discuss these claims in the next sections. Based on these claims, we construct our strategy in two steps. We begin by designing a regularizer for efficiently training a predictor f in the unconfounded setting where $\mathcal{D} \sim P^\circ$. We then show how this can be generalized to training distributions $\mathcal{D} \sim P_s$ using importance weighting.

Regularization under P° We design our regularizer to leverage the auxiliary label V, using two facts that hold under P°: (1) $V \perp\!\!\!\perp \mathbf{X}^*$, and (2) the optimal predictor is a function of only $\mathbf{X}^*$ (proposition 3.2.1). Based on these facts, we specify a regularizer for f that encourages $f(\mathbf{X}) \perp\!\!\!\perp V$. We do this by penalizing the distributional discrepancy between conditional distributions of the representation $P^\circ(\phi(\mathbf{X}) \mid V = 0)$ and $P^\circ(\phi(\mathbf{X}) \mid V = 1)$ that would be identical under independence. Although any number of estimable distributional discrepancy metrics could be used, here we choose to use the Maximum Mean Discrepancy (MMD), defined as follows:

Definition 3.3.1. *Let Z, and Z', be two arbitrary variables with $Z, Z' \in \mathcal{Z}$, and their*

corresponding distributions P_Z and $P_{Z'}$. And let Ω be a class of functions $\omega : \mathcal{Z} \to \mathbb{R}$,

$$\mathrm{MMD}(\Omega, P_Z, P_{Z'}) = \sup_{\omega \in \Omega} \big(\mathbb{E}_{P_Z} \omega(Z) - \mathbb{E}_{P_{Z'}} \omega(Z') \big).$$

When Ω is set to be a general reproducing kernel Hilbert space (RKHS), the MMD defines a metric on probability distributions, and is equal to zero if and only if $P_Z = P_{Z'}$. Throughout, we will assume that our predictor f and our loss ℓ are contained in Ω, and in practice choose Ω to be the RKHS induced by the radial basis function (RBF) kernel. We will use the shorthand $\mathrm{MMD}(P_Z, P_{Z'})$ to denote $\mathrm{MMD}(\Omega, P_Z, P_{Z'})$. See Gretton et al. (2012) for a review of MMD and its empirical estimators.

Weighting to Recover P°. When the training data is drawn from some $P_s \neq P^\circ$, we weight the data to obtain empirical risk and MMD expressions that are unbiased estimates of the expressions we would obtain if $\mathcal{D} \sim P^\circ$, and proceed as before. In particular, we define weights

$$u(y, v) = \frac{P_s(Y = y) P_s(V = v)}{P_s(Y = y, V = v)}, \tag{3.2}$$

such that for each example, $u_i := u(y_i, v_i)$. For any distribution P_s, these are importance weights that map expectations under P_s to expectations under P°. In the appendix, we show that the reweighted risk is an unbiased estimator of the risk under P°, i.e., that

$$\mathbb{E}_{Ps} \left[\hat{R}^{\mathbf{u}}_{Ps}(f) \right] = R_\circ(f),$$

where $\hat{R}^{\mathbf{u}}_{Ps}(f) = \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$, and $R_\circ(f) = \mathbb{E}_{\mathbf{X}, Y \sim P^\circ} \left[\ell(f(\mathbf{X}), Y) \right]$.

Method Putting the different components of our approach together gives us a final objective to optimize: let $\boldsymbol{\phi}_v$ denote $\{\phi(\mathbf{x}_i)\}_{i : v_i = v}$, and $\boldsymbol{\phi}^{\mathbf{u}}_v$ denote its re-weighted analogue, and let u_i be as in equation 3.2. For $\mathcal{D} \sim P_s$, and some $\alpha > 0$, the main

objective to minimize is:

$$h^*, \phi^* = \underset{h,\phi}{\operatorname{argmin}} \sum_i u_i \ell(h(\phi(\mathbf{x}_i), y_i) \ + \alpha \cdot \widehat{\mathrm{MMD}}^2(P_{\boldsymbol{\phi}_0^{\mathbf{u}}}, P_{\boldsymbol{\phi}_1^{\mathbf{u}}}). \tag{3.3}$$

To estimate $\widehat{\mathrm{MMD}}^2$, we use a weighted version of the U-statistic estimator presented in Gretton et al. (2012). Specifically, we compute:

$$\widehat{\mathrm{MMD}}^2 = \sum_{i,j:v_i,v_j=0} u_i u_j k_\gamma(\phi_i, \phi_j) + \sum_{i,j:v_i,v_j=1} u_i u_j k_\gamma(\phi_i, \phi_j) - 2 \sum_{i,j:v_i=0,v_j=1} u_i u_j k_\gamma(\phi_i, \phi_j),$$

where $k_\gamma(x, x')$ is the radial basis function, with bandwidth γ.

Cross-validation The objective function in (3.3) depends on two hyperparameters. The first is the cost of the MMD penalty α, and the second is γ, the kernel bandwidth necessary to compute the MMD term. Unlike the usual regularization schemes, the MMD-regularization term also depends on the distribution of the data, and is subject to errors caused by finite samples. In other words, it is possible to overfit this objective such that the MMD on the training data is 0 but it remains large on a validation set. For this reason, we follow a two-step cross-validation procedure. In the first step, we calculate the weighted MMD on each of the K validation folds. We then exclude all models that achieve a weighted MMD that is statistically significantly different from zero. This gives us a subset of the function candidates that encode the desired invariances. In the second step, we pick the best performing model out of this subset of candidate functions.

3.4 Theory

Our goal is to estimate the generalization error of our estimator presented in section 3.3. Meaning, we wish to bound the difference between the error on any target distribution $P_t \in \mathcal{P}$, and the empirical error on the source distribution P_s. This in

turn means that we want to control $R_{Pt}(f) - \hat{R}^{\mathbf{u}}_{Ps}(f)$. A key observation is that this difference can be decomposed as follows:

$$R_{Pt}(f) - \hat{R}^{\mathbf{u}}_{Ps}(f) = \underbrace{R_{Pt}(f) - R^{\circ}(f)}_{\text{Structural risk gap}} + \underbrace{R^{\circ}(f) - \hat{R}^{\mathbf{u}}_{Ps}(f)}_{\text{Finite-sample gap}}, \tag{3.4}$$

where $\hat{R}^{\mathbf{u}}_{Ps}(f)$ is the weighted empirical error $= \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$.

This decomposition is a summary of our strategy: we decompose the difference $R_{Pt}(f) - \hat{R}^{\mathbf{u}}_{Ps}(f)$ as the difference between the error on the target distribution and the error on P°, added to the difference between the empirical error on the source distribution and the error on P°. In the remainder of this section, we first bound the finite-sample gap in section 3.4.1. Second, we bound the structural gap in section 3.4.2. All proofs are in the appendix.

3.4.1 Bounding the finite-sample gap

For the purpose of studying the finite sample gap, we focus on the special case of linear models (i.e., one-layer fully dense neural networks) to establish key insights about the properties of causally-motivated regularization. Specifically, we consider the special case where ϕ is a linear mapping, i.e., $\phi(\mathbf{x}) = \mathbf{w}^{\top}\mathbf{x}$, and h is the sigmoid, i.e., $h(x) = \sigma(x) = 1/1 + \exp(-x)$. Extensions of our theoretical analysis to more complex neural networks are possible (e.g., through approaches studied in Golowich et al. (2018)).

We will compare the efficiency of the MMD-regularization approach to the more commonly used L_2 regularization approach. The two regularization schemes characterize

two different function spaces:

$$\mathcal{F}_{L_2} = \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A\}, \tag{3.5}$$

and

$$\mathcal{F}_{L_2,\text{MMD}} = \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A,\ \text{MMD}(P^\circ_{\phi_0}, P^\circ_{\phi_1}) \leq \tau\}, \tag{3.6}$$

Before delving into the comparison between those two function classes, we highlight a core finding that provides intuition about the advantages of the MMD penalty.

3.4.1.1 An intuition for the MMD penalty

In the special case where ϕ is a linear mapping, the MMD constraint has direct implications for the weights $\mathbf{w}$. Here, the constraint restricts the projection of $\mathbf{w}$ onto the dimension that distinguishes the conditional means $\boldsymbol{\mu}_0 := \mathbb{E}_{\mathbf{x}\sim P^\circ}[\mathbf{x}_i \mid v_i = 0]$ and $\boldsymbol{\mu}_1 := \mathbb{E}_{\mathbf{x}\sim P^\circ}[\mathbf{x}_i \mid v_i = 1]$. To make this precise we denote the difference between the mean vectors as $\Delta := \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. Δ is the average change in $\mathbf{x}$ caused by different values of V. Define the projection matrix $\Pi := \Delta(\Delta^\top\Delta)^{-1}\Delta^\top = \|\Delta\|_2^{-2}\Delta\Delta^\top$, which projects any vector onto Δ. We then define $\mathbf{w}_\perp := \Pi\mathbf{w}$ as the projection of $\mathbf{w}$ onto the mean distinguishing direction, which can be thought of as the "bad" or "irrelevant" dimension.

We can directly relate $\|\mathbf{w}_\perp\|$ to the MMD penalty, in the following proposition.

Proposition 3.4.1. *Let $f(\mathbf{x}) = \sigma(\phi(\mathbf{x})) = \sigma(\mathbf{w}^\top\mathbf{x})$ be a function contained in $\mathcal{F}_{L_2,\text{MMD}}$. Then, $\|\mathbf{w}_\perp\| \leq \frac{\tau}{\|\Delta\|}$.*

Intuitively, proposition 3.4.1 says that the MMD penalty limits the effect of the irrelevant components of $\mathbf{w}$ proportionally to τ. In the image classification example, this means that the parts of $\mathbf{w}$ that can distinguish between land backgrounds and water background is limited. In a simple linear setting, this means the MMD penalty

identifies which features in the input space are associated with a change in V, and penalizes their weights.

3.4.1.2 Finite-sample bound when $P_s = P^\circ$

We are now ready to compare the efficiency of the MMD-regularization approach to the more commonly used L_2 regularization approach. We start first by analyzing the finite-sample bound in the special case where $P_s = P^\circ$, and consider the more general case later.

We compare $\mathcal{F}_{L_2}$, and $\mathcal{F}_{L_2,\text{MMD}}$ in terms of the Rademacher complexity:

Definition 3.4.1. *Let $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^n$ denote a vector of independent random variables drawn from the Rademacher distribution, i.e., uniform on $\{-1, 1\}$. For a function family $\mathcal{F}$, and $\mathcal{D} \sim P$, the Rademacher complexity for a sample of size n is defined as: $\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(\mathbf{x}_i)\right]$.*

For a bounded function $f \in \mathcal{F}$, a loss function that is L-Lipschitz, and a training data of size n, with probability $1 - \delta$, the following holds (Mohri et al., 2018):

$$R(f) \leq \hat{R}(f) + L \cdot \mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}, \tag{3.7}$$

where $\hat{R}(f)$ is the empirical error $= \frac{1}{n}\sum_i \ell(f(\mathbf{x}_i), y_i)$

Proposition 3.4.2 states that even in the absence of distribution shift, when $P_s = P_t = P^\circ$, explicitly penalizing MMD is advantageous because it reduces the hypothesis space without introducing bias. Since $\mathcal{F}_{L_2,\text{MMD}} \subseteq \mathcal{F}_{L_2}$, we expect the MMD penalty to reduce the hypothesis space. However the key thing to note here is that this reduction does not introduce bias. We formally show that in the following proposition.

Proposition 3.4.2. *For $\mathcal{D} \sim P^\circ$, and for any for any $\mathcal{F}_{L_2}$ such that $f_{rinv} \in \mathcal{F}_{L_2}$, there exists a $\mathcal{F}_{\text{MMD},L_2} \subseteq \mathcal{F}_{L_2}$ such that $f_{rinv} \in \mathcal{F}_{\text{MMD},L_2}$. And the smallest $\mathcal{F}_{\text{MMD},L_2}$*

such that $f_{rinv} \in \mathcal{F}_{\mathrm{MMD},L_2}$ *has* $\mathrm{MMD} = 0$.

To examine how much smaller $\mathcal{F}_{\mathrm{MMD},L_2}$ is compared to $\mathcal{F}_{L_2}$, we derive comparable bounds on the Rademacher complexity of the two function classes by splitting f in terms of how the observed features $\mathbf{x}$ align with the mean difference vector Δ. Here, we let $\mathbf{x}_\perp := \Pi\mathbf{x}$ be the component of $\mathbf{x}$ that is parallel to the mean discrepancy i.e., parallel to the "irrelevant component," and hence perpendicular to the relevant components. We let $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$ be the orthogonal component to the irrelevant components (i.e., the "relevant component").

Proposition 3.4.3. *Let* $\mathbf{x}_\perp := \Pi\mathbf{x}$, $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$. *For training data* $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, $\sup_{\mathbf{x}_\perp} \|\mathbf{x}_\perp\|_2 \leq B_\perp$, $\sup_{\mathbf{x}_\parallel} \|\mathbf{x}_\parallel\|_2 \leq B_\parallel$, *then*

$$\mathfrak{R}(\mathcal{F}_{L_2}) \leq \frac{A\sqrt{B_\parallel^2 + B_\perp^2}}{\sqrt{n}},$$

and

$$\mathfrak{R}(\mathcal{F}_{\mathrm{MMD},L_2}) \leq \frac{A \cdot B_\parallel + \tau\frac{B_\perp}{\|\Delta\|}}{\sqrt{n}}.$$

The proof shown in the appendix applies the standard Rademacher complexity bound for the L_2 class, and obtains a looser bound for $\mathcal{F}_{\mathrm{MMD},L_2}$ by separately bounding the worst-case terms involving $\mathbf{x}_\perp$ and $\mathbf{x}_\parallel$. Comparing these bounds is instructive. In particular, the upper bound on $\mathfrak{R}(\mathcal{F}_{\mathrm{MMD},L_2})$ is smaller than that of $\mathfrak{R}(\mathcal{F}_{L_2})$ whenever τ satisfies:

$$0 \leq \tau < A\left[\sqrt{B_\parallel^2 + B_\perp^2} - B_\parallel\right]\frac{\|\Delta\|}{B_\perp}. \tag{3.8}$$

The key part of this expression is the ratio $\|\Delta\|/B_\perp$, which can be understood as a characterization of how much of the variation in $\mathbf{x}_\perp$ comes from the mean-shift in $\mathbf{x}$ induced by v. In particular, when the variation caused by mean shift is large, we expect even weak MMD regularization to yield better generalization than L_2 regularization alone, and we expect this effect to be even stronger when variation in

the mean shift direction is large relative to variation in orthogonal directions. This occurs in cases where V controls features that are highly salient in the input $\mathbf{x}$. For example, in object recognition using images, if V denotes the background type, we expect $\|\Delta\|/B_{\perp}$ to be large if the background features are very different between $V = 0$ and $V = 1$, but relatively consistent within values of V. Further, if the background accounts for the majority of pixels in each image, we expect $B_{\perp} \gg B_{\|}$, resulting in an even stronger regularizing effect from the MMD penalty.

Having derived a bound on the Rademacher complexity, we can extend the results from proposition 3.4.3 to get full generalization error bound by plugging in the Rademacher bounds into expression (3.7). Similar to Lyle et al. (2020) and Chen et al. (2020), among others, we focus on comparing the Redmacher complexities of the two functions, and conjecture that the MMD-regularization does not significantly inflate the empirical (training) error $\hat{R}_{P^{\circ}}(f_{\mathrm{MMD},L_2})$ relative to $\hat{R}_{P^{\circ}}(f_{L_2})$. This conjecture is not unreasonable since the majority of existing DNN architectures are flexible enough to achieve zero or near zero training error.

With that, and whenever τ satisfies inequality 3.8, the MMD regularization scheme will have a favorable (i.e., lower) generalization error compared to that of L_2 regularization.

3.4.1.3 Finite-sample bound when $P_s \neq P^{\circ}$

When the data is sampled from any distribution other than P°, proposition 3.2.1 does not hold, and the population risk minimizer does not correspond to the optimal invariant risk predictor f_{rinv}. In addition, we are not guaranteed that there exists some value of τ such that $\mathcal{F}_{L_2,\mathrm{MMD}} \subseteq \mathcal{F}_{L_2}$. Recall that the smallest $\mathcal{F}_{L_2,\mathrm{MMD}}$ has $\tau = 0$ (see expression 3.8), the following proposition shows that when sampling from a biased distribution, the smallest τ' that does not introduce bias is greater than 0.

Proposition 3.4.4. *Let* $\mathcal{F}'_{L_2,\mathrm{MMD}} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^{\top}\mathbf{x}), \|\mathbf{w}\|_2 \leq A, \ \mathrm{MMD}(P_{\phi_0}, P_{\phi_1}) \leq$

$\tau'\}$ *be the smallest function class that contains* f_{rinv}*. Then* $\tau' = c \cdot A$ *for some* $c > 0$*, and the corresponding generalization error on* P° *is*

$$R^\circ(f) \leq \hat{R}_P^{\mathbf{u}}(f) + L \cdot \frac{A \cdot B_\parallel + c \cdot A \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

This means that when $\mathcal{D} \sim P_s \neq P^\circ$, the MMD-regularized family that contains f_{rinv} might be larger than the L_2 family. To address this issue and recover the results from the previous section, we will rely on reweighting the training data to mimic the independencies in the optimal distribution P°.

Note that assumption 3.2.2 implies overlap of the source distribution P_s with the ideal distribution P° since $P_s(V)P_s(Y) = P^\circ(V)P^\circ(Y)$. As a consequence of the overlap assumption, we have that:

$$\sup u(y, v) = \sup \frac{P^\circ(Y \mid V)}{P_s(Y \mid V)} = 2^{\Xi_\infty(P^\circ||P_s)} = C_{Ps} < \infty \tag{3.9}$$

where $\Xi_k(p||q)$ is the k^{th}-order Rényi divergence, and the second equality follows by applying the Bayes rule, and the definition of the Rényi divergence. It will be convenient to denote $2^{\Xi(p||q)}$ by $\Lambda_k(p||q)$. Since $2^{\Xi k-1(P^\circ||P_s)} < 2^{\Xi(P^\circ||P_s)}$, we have $\Lambda_2(P^\circ||P_s) < C_{Ps}$. Following similar work (e.g., Makar et al. (2020)), we will assume that the weights $\mathbf{u}$ are known, or can be perfectly estimated from the data. In other words, we do not consider estimation error that might arise because of poor estimation of $\mathbf{u}$. Work by Foster and Syrgkanis (2019) has shown that under mild assumptions, the error due to estimation of $\mathbf{u}$ from finite samples only results in a fourth order dependence in the final classifier, and hence does not greatly affect our derived generalization bounds.

To get the generalization error of weighted estimators, we apply results from Cortes et al. (2010b) in the following proposition.

Proposition 3.4.5. *For a training dataset* $\mathcal{D} \sim P_s$*, a corresponding* C_{Ps} *as defined*

in equation 3.9, $\mathbf{u}$ *as defined in equation 3.2,* $\varepsilon > 0$*, and for universal constants* $c', c'' > 0$*, , with probability* $1 - \delta$*:*

$$R^{\circ}(f) \leq \hat{R}^{\mathbf{u}}_{P}(f) + \frac{2C_{Ps}(\kappa(\mathcal{F}_{\mathrm{MMD},L_2}) + \log\frac{1}{\delta})}{2n} + \sqrt{\frac{\Lambda(P^{\circ}||P_s)\cdot(\kappa(\mathcal{F}) + \log\frac{1}{\delta})}{n}},$$

where

$$\kappa(\mathcal{F}_{\mathrm{MMD},L_2}) = c''\left(\frac{c'\sqrt{\log(n)}\cdot\left(A\cdot B_{\|} + \tau\frac{B_{\perp}}{\|\Delta\|}\right)}{\varepsilon}\right)^2$$

Comparing the result from proposition 3.4.4 to that of proposition 3.4.5 does not give a clear winning strategy: it is possible to get better generalization without reweighting if τ' is small enough, and it is possible to get better generalization under reweighting if C_{Ps} and $\Lambda(P^{\circ}||P_s)$ are small enough. However, without reweighting, it is crucial for τ' to be large enough so that $f_{\mathrm{rinv}} \in \mathcal{F}'_{\mathrm{MMD},L_2}$ but small enough so that the hypothesis space is small. As we show in the empirical analysis section, typical cross validation methods are prone to select values for τ' that are larger than necessary, leading to a less robust estimator. This makes the reweighting strategy more practical, and hence more appealing.

3.4.2 Bounding the structural risk gap

In the previous section, we showed that the MMD penalty leads to efficient estimators, and bounds the second term in equation 3.4. In the following proposition, we show that the MMD penalty also bounds the first term in equation 3.4, the structural risk gap.

For this proposition, we require that y is exactly recoverable from $\mathbf{x}$. This assumption allows us to ensure that the MMD between the different V classes for *each* value of y remains bounded. Such a strong assumption can be avoided for variants of the MMD

penalty which explicitly minimize the MMD for each value of y. We return to this in the discussion section (section 3.7).

Proposition 3.4.6. *For training data $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, and a corresponding learned $f = h(\phi(\mathbf{x}))$ with risk $R^\circ(f)$, suppose that y is $\phi-$representable, i.e., that there exists $g(\phi(\mathbf{x})) = y$, and that $g(\phi)\ell(\phi) \in \Omega$. For some β that depends on P_t, such that $-2 < \beta < 2$, and $\beta = 0$ if $P_t = P^\circ$, then*

$$R_{Pt}(f) \leq R^\circ(f) + \beta \cdot \tau.$$

Proposition 3.4.6 provides another motivation for using the MMD penalty. The MMD penalty directly encourages small values of τ; this regularizes the solution toward a predictor that has similar risks on P_t, and P°. This in turn means that the first term in equation 3.4 is small, leading to low generalization error of our proposed weighted estimator.

3.5 Experiments

We empirically analyze the performance of causally motivated regularization in two settings. In the first setting, we consider training data that is sampled from the optimal distribution P°. This setting helps us study the implications of proposition 3.4.3, which suggests that even under uncorrelated sampling, we can see improvement in finite sample efficiency when we use the MMD penalty. In the second setting, the training data is sampled from some P_s, where the auxiliary label and the main label are correlated, i.e., $V \not\perp\!\!\!\perp Y$.

To control the correlation between Y, and V we follow a procedure similar to that presented in Sagawa et al. (2019). Specifically, we construct a dataset that combines images of water birds (Gulls) and land birds (Warblers) extracted from the Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al., 2011) with water and land

background extracted from the Places dataset (Zhou et al., 2017). Figure 3-2 shows examples of the generated images.

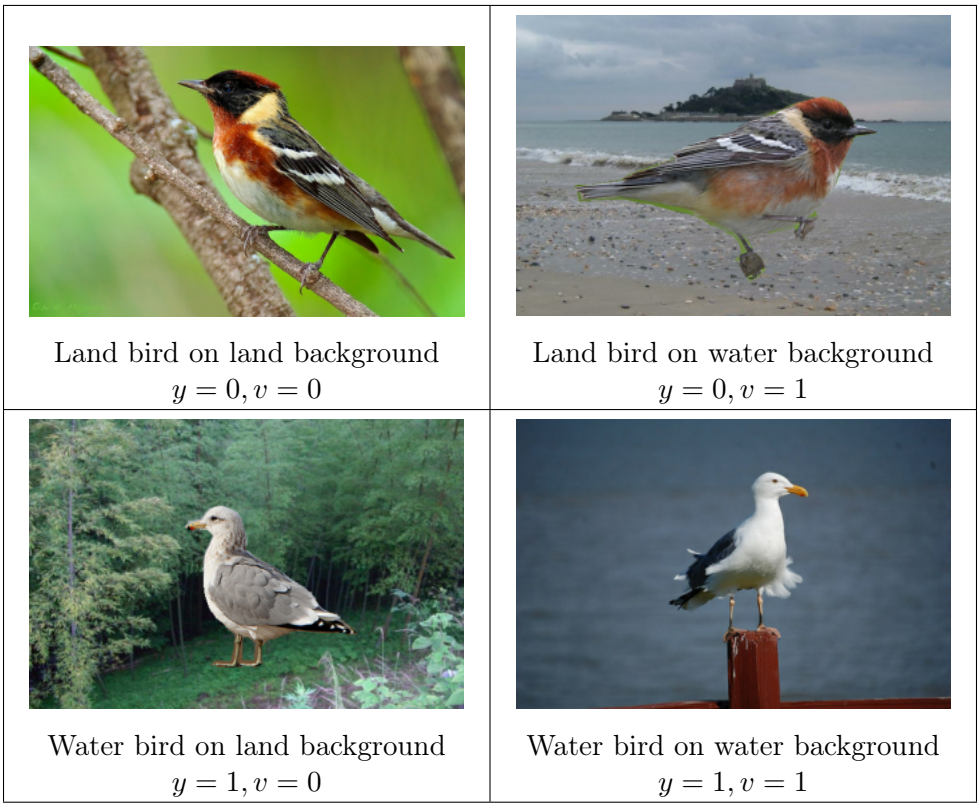

Figure 3-2: Examples of the generated images of water, and land birds on water, and land backgrounds

We found that the original background images frequently contain landscapes that are difficult to distinguish (e.g., water backgrounds with very small water bodies that mostly reflect the surrounding trees). Instead, we pick 300 "clean" images for each of the land and water backgrounds. Using those clean images, we generate 10,000 land backgrounds, and 9,000 water backgrounds by applying random transformations (rotation, zoom, darkening/brightening) to the selected images. In the appendix, we present the results on the original backgrounds.

For the first setting, we generate the training data from the optimal distribution P°, with $P^{\circ}(Y|V = 1) = P^{\circ}(Y|V = 0) = 0.5$. In the second setting, we generate the data such that $P(Y = 1|V = 1) = P(Y = 0|V = 0) = 0.9$, representing a scenario where the majority of water birds are on water backgrounds and the majority of land birds

are on land backgrounds. While our theory assumes a no-noise setting, to test our model on a realistic setting, we introduce noise by randomly flipping 1% of each of the labels. We generate a number of held-out test sets, each one corresponding to a different probability of observing a waterbird with a water background, and similarly with land birds.

We present results from the following variants of our approach:

1. **wMMD-reg-T**: this model corresponds to minimizing equation 3.3. It includes *w*eighting by the weights $\mathbf{u}$, penalizing the MMD, followed by the two-step cross validation process described in the implementation section.

2. **wMMD-reg-C**: similar to wMMD-reg-T, this model minimizes equation 3.3 but does the classical cross-validation process, where it simply picks the model that has the best performance on the held out validation set.

3. **MMD-reg-T**: this model minimizes a variant of equation 3.3 that excludes weighting by $\mathbf{u}$. The optimal hyperparameters are picked using the two-step cross-validation algorithm, taking the $\mathbf{u}-$weighted estimates of the MMD and prediction performance into account.

4. **MMD-reg-uT** is similar to MMD-reg-T, however it uses unweighted validation metric estimates during the two-step cross-validation procedure.

5. **MMD-reg-C**: this model minimizes a variant of equation 3.3 that excludes weighting by $\mathbf{u}$, and does the classical cross-validation process.

In addition, we present results from the following baselines.

1. **L2-reg**: this is the standard DNN trained to minimize the empirical risk. We introduce regularization by penalizing the L2-norm of the weights, picking the value of the penalty from 0.0 (no regularization) or 0.0001, which is the value typically used for this setting (Sagawa et al., 2019; He et al., 2016).

2. **wL2-reg**: similar to L2-reg but also incorporates weighting using u_i as defined in 3.2.

3. **Rand-Aug-C**: a baseline that attempts to create a robust estimator by augmenting the data at training time using random flips and random rotations. Cross-validation is done in the typical way.

For the uncorrelated setting, we only present the unweighted variants of all the models, since the weights are roughly constant across data points in that setting.

We present the results from 20 simulations. We keep the architecture fixed across all models. Specifically we use ResNet-50 (He et al., 2016), pretrained on ImageNet, and fine tuned for our specific task. All models are implemented in TensorFlow (Abadi et al., 2015).

Results: Sampling from the optimal distribution. Figure 3-3 shows the results from the first setting, where the training data is sampled from the optimal, uncorrelated distribution P°, with $P^{\circ}(Y|V=1) = P^{\circ}(Y|V=0) = 0.5$. The $x-$axis shows $P(Y=1|V=1) = P(Y=0|V=0)$ at test time, while the $y-$axis shows the corresponding mean AUROC, averaged over 20 simulations. The vertical dashed line shows the conditional probability at training time. We see that both variants of our proposed approach, with classical and two-step cross-validation outperform the L_2-regularized model and the random augmentation model within the training distribution (i.e., at the dashed line) and also when there is distribution shift. This conforms with proposition 3.4.3. Even when the data are sampled from the optimal distribution, using a causally-motivated regularization scheme leads to more efficient models, which translates into better performance in finite samples.

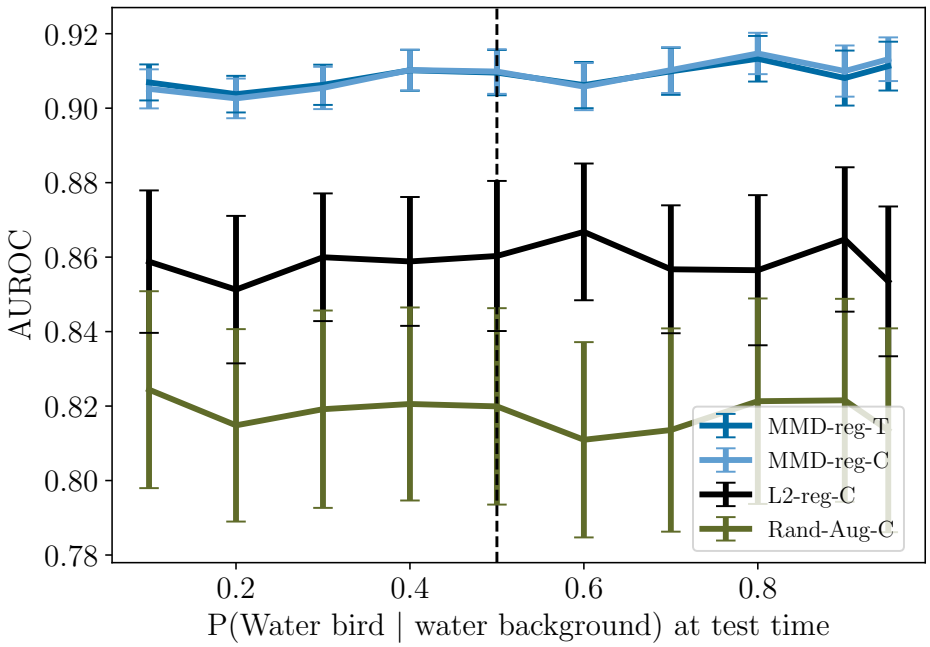

Figure 3-3: Training data sampled from P°, with $P^\circ(Y|V=1) = P^\circ(Y|V=0) = 0.5$. x-axis shows $P(Y|V)$ at test time under different shifted distributions. y-axis shows AUROC on test data. Vertical dashed line shows training data. MMD-regularized models outperform baselines within, and outside the training distribution.

Results: Sampling from a correlated distribution. Figure 3-4 shows the results from the second setting, where the training data is sampled from a correlated distribution with $P(Y=1|V=1) = P^\circ(Y=0|V=0) = 0.9$. The x, and y axes are similar to figure 3-3. Here we see that our main suggested approach (wMMD-reg-T) outperforms other models especially at high divergence from the training distribution. Out of all the non-MMD regularized baselines, the weighted L2-regularized model performs best. This suggests that minimizing the empirical risk on the $\mathbf{u}-$reweighted distribution contributes to model robustness.

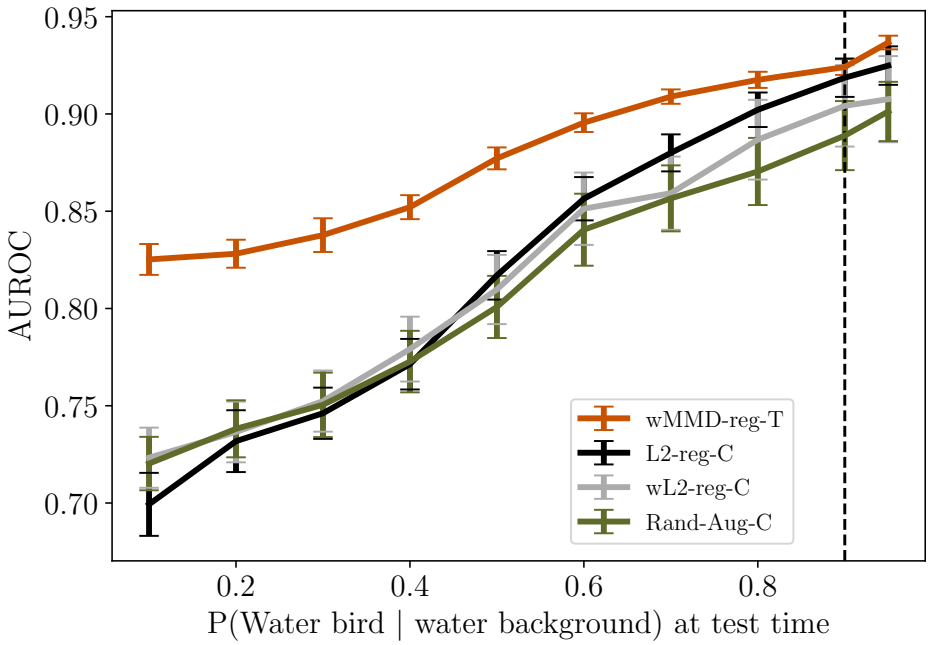

Figure 3-4: Training data sampled from P, with $P(Y = 1|V = 1) = P^{\circ}(Y = 0|V = 0) = 0.9$. Vertical dashed line shows training data. x, y axes similar to figure 3-3. MMD-regularized models outperform baselines showing better robustness against distribution shifts at test time.

Figure 3-5 shows an ablation study where we remove different components of our model to see how each contributes to improved performance. The largest increase in performance is attributable to weighting by $\mathbf{u}$ at training time, since the two weighted variants outperform the two unweighted variants. Within those two groups, the two-step approach with weighted validation metrics outperforms the others, especially in terms of robustness to distribution shifts. This shows that when training models using the MMD-penalty, it is important to take into consideration that the MMD-penalty (unlike L2-norm regularization) also depends on the training data, and is prone to overfitting. The results also show that it is possible to improve the performance of models that are unweighted at training time by using our two step cross validation approach with weighted validation metrics, since MMD-reg-T slightly outperforms MMD-reg-C. Recall that MMD-reg-uT strictly enforces the MMD-penalty without addressing the fact that the training distribution has been sampled from a correlated distribution. We see that while it gives a robust model, that model has relatively poor

performance. This conforms with our findings stated in prop 3.4.4, which implies that there will be a bias-robustness trade-off if the correlated sampling is not corrected.

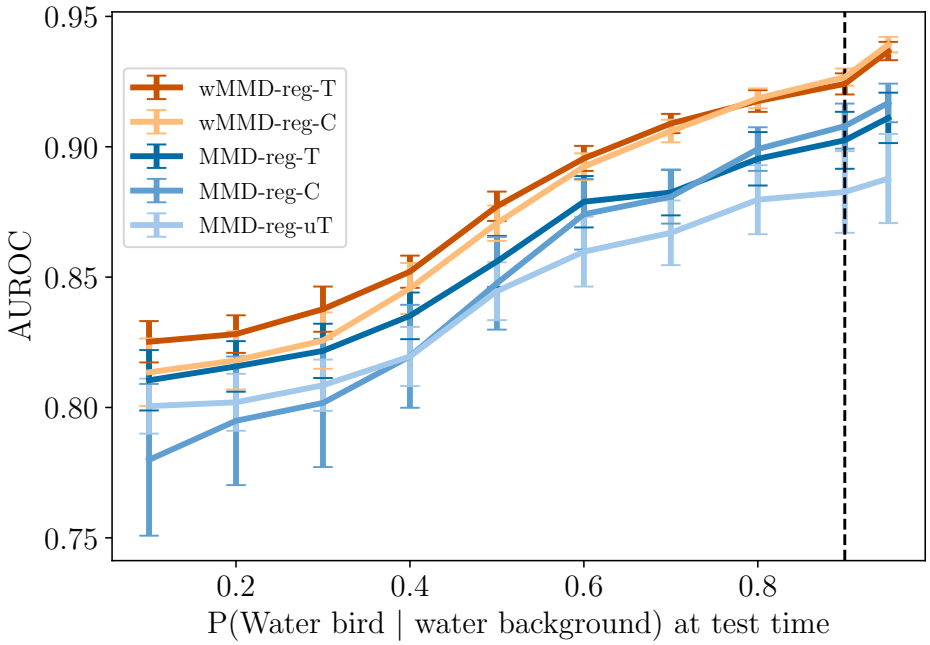

Figure 3-5: Training data sampled from P, with $P(Y = 1|V = 1) = P^{\circ}(Y = 0|V = 0) = 0.9$. x, y axes similar to fig 3-3. An ablation study to show how different components of our suggested approach (wMMD-reg-T) contribute to improved performance.

3.6 Connections to existing work

The work we presented here unifies several threads that have appeared in the ML literature.

Shortcut learning. The majority of work addressing shortcut learning relies on data augmentation: the practitioner defines a set of transformations (e.g., rotation, translation, cropping) that should not affect the main label, and adds the augmented examples to induce invariance to these transformations (Hendrycks et al., 2020; Yin et al., 2019; Lyle et al., 2020; Lopes et al., 2019; Cubuk et al., 2018). One disadvantage of this approach is that it assumes the set of transformations is known *a priori*. If this set of transformations is misspecified, the desired robustness might not be

achieved, as evidenced by the empirical performance of the random augmentation baseline presented in the experiments section. Our approach is different in that we do not claim to know these transformations. Instead, our approach leads to invariant models by leveraging the auxiliary labels to inform the relevant transformations the main label should be independent to.

Shortcut learning can be viewed as a consequence of model underspecification (D'Amour et al., 2020). Many of the issues related to underspecification have appeared in the ML literature within the context of overparameterization. For example, in Sagawa et al. (2020b) authors observe that overparameterization exacerbates the reliance on spurious correlations. Their suggested approach is somewhat similar to the reweighting baseline (W-DNN) presented in the experiments section, which is outperformed by our model.

Invariant representations. Our work sheds light on properties of invariant representations, which are extensively studied in the fairness literature (Madras et al., 2018), causality literature (Shalit et al., 2017; Johansson et al., 2016), and domain shift literature (Tzeng et al., 2014; Long et al., 2015). Specifically, our analysis (proposition 3.4.1) highlights *how* invariant representations regularize "redundant" dimensions. We also address one key question that has been discussed extensively recently: the question of whether there is a trade-off between invariance and accuracy, or stated differently: whether invariance leads to biased estimation (Zhang et al., 2019; Calders et al., 2009; Johansson et al., 2019; Dutta et al., 2020; Zhao and Gordon, 2019). Propositions 3.4.2 shows that if the training data is sampled from the optimal distribution, then the MMD penalty does not lead to bias (i.e., there is no tradeoff). However, proposition 3.4.4 shows that naively implementing the MMD penalty when the data is sampled from a biased distribution might lead to issues of bias. Our analysis suggests that if coupled with the correct sample reweighting the MMD penalty does not lead to bias.

More generally, while proposition 3.4.6 bears some similarity to statements presented

in the domain adaptation literature (Long et al., 2015; Ben-David et al., 2007, 2010), our work is distinct in that we do not aim to generalize to a specific target domain. Instead, we aim to build models that generalize across a *family* of target domains. One consequence of this distinction is that, unlike unsupervised domain adaptation, we do not require access to examples from a target domain.

Causally-motivated invariance. Our work is similar to Anchor regression (Rothenhäusler et al., 2018) in that we also view the question of invariance through the lens of causality. Our work is distinct in that we do not assume linear relationships between $\mathbf{X}, Y$, and the "anchor" variable V, and we are not limited to linear models. Arjovsky et al. (2019) propose an invariant risk minimization (IRM) approach that is inspired by ideas from causality. Unlike our approach, IRM does not explicitly penalize dependence on the redundant dimensions, and instead relies on the idea that the invariant risk minimizer should achieve the lowest error across datasets sampled from different target distributions P_t. As others (e.g.,Guo et al. (2021); Rosenfeld et al. (2020)) have noted, when the family of functions is as flexible as DNNs, and without further assumptions on the training distribution, it is possible to find a predictor that achieves the objective of IRM but is not robust. Guo et al. (2021) have attempted to address the limitations of IRM using an MMD penalty, however, they do not correct the estimates of the MMD for biased sampling and hence have the same limitations as the unweighted MMD penalized models presented in the empirical section.

Similar to our work, (Wang et al., 2018) suggest a regularization scheme that discourages a learned model from relying on irrelevant features. The work presented in this chapter is distinct in that we do not require expert knowledge of the relevant features, or factors. Janizek et al. (2020) utilize the causal DAG to induce robustness to distribution shifts. However, they rely on adversarial training to induce independence, which has been shown to be unstable. In addition, they enforce the independence on the source distribution, without utilizing a reweighting scheme which may introduce bias (see proposistion 3.4.4).

3.7 Discussion

We presented an approach to using auxiliary labels to build models that are invariant to distribution shifts defined by interventions on factors that should affect the auxiliary label but not the target label. Our analysis highlights important theoretical properties of the MMD penalty, which is often used in the fairness, causality, and domain adaptation literatures. Guided by our theoretical insight, we suggested a causally-motivated regularization scheme that combines reweighting and the MMD penalty to train robust, and accurate models. Using a well-known robustness benchmark, we show that our approach empirically outperforms others.

One of the core findings of this work is fact that mapping the observed distribution onto an "ideal" unconfounded distribution is advantageous since it allows building robust classifiers. In this chapter we focused on one such distribution, $P^\circ = P_s(\mathbf{X}|\mathbf{X}^*, V)P_s(\mathbf{X}^*|Y)P_s(Y)P_s(V)$ but our results (e.g., proposition 3.2.1) extend to a wider family of ideal distributions where the marginal distribution over V is allowed to vary arbitrarily. Formally, this distribution is defined as

$$\mathcal{P}^\circ = \{P_s(\mathbf{X}|\mathbf{X}^*, V)P_s(\mathbf{X}^*|Y)P_s(Y)Q(V)\},$$

for $0 < Q(V) < 1$. We chose to fix $Q(V) = P_s(V)$ since it is the closest possible distribution to the source distribution, i.e., it has the smallest divergence $\Lambda(P^\circ \| P_s)$. Having a lower divergence is favorable, since lower $\Lambda(P^\circ \| P_s)$ leads to tighter generalization error (see proposition 3.4.5).

Limitations and extensions. Some of the results in this chapter require a more stringent assumption, that is y is exactly recoverable from $\mathbf{x}$. We note that this assumption can be relaxed for a slightly different version of the MMD penalty. Specifically, let $\text{MMD}_y := \text{MMD}(P(\phi(\mathbf{x}_i) \mid V = 0, Y = y), P(\phi(\mathbf{x}_i) \mid V = 1, Y = y))$ the assumption of perfect recoverability can be avoided for the conditional version of our

penalty, defined as

$$\text{MMD}' = \sum_y \text{MMD}_y.$$

However, such a penalty "slices" the data into four (rather than two) inevitably smaller subgroups. The data available to estimate each of MMD_0, and MMD_1 is smaller leading to less accurate estimates. This is especially problematic when the training process relies on stochastic gradient descent where small batches are used to estimate MMD'.

For this reason we opted for the version of the MMD penalty which aggregates the two y groups in estimating the MMD and relied on the assumption of recoverability of y. Further research that "switches" between the two MMD penalties in the abundance of data would strengthen our approach.

Another possible limitation to our approach is that it requires *a priori* knowledge of the shortcut that might be exploited by the model. However, if the shortcut is unknown, practitioners can use interpretability methods to understand the main factors that the model relies on. If the interpretability analysis reveals that the model is relying on a shortcut, and if an auxiliary label that corresponds to that shortcut is available, our proposed approach can then be used. For example, in image classification, saliency maps can reveal the learned factors that influence the prediction the most (Simonyan et al., 2013). In contexts other than image classification, other interpretability tools such as Shapley values are more appropriate (Lundberg and Lee, 2017; Wang et al., 2021).

In addition, we make an assumption similar to the overlap assumption in causality. In other words, we assume that the correlation between the main and auxiliary labels cannot be perfect at training time. Extensions of this work that relax this assumption would require making stronger assumptions about the properties of f, but they might make the approach more applicable.

Finally, as presented, our approach allows for only one binary auxiliary label. Extensions that consider non-binary auxiliary labels, as well as multiple auxiliary labels would lead to models that are even more robust.

Chapter 4

Exploiting structured data for learning contagious diseases under incomplete testing

In the previous chapter we investigated how penalizing predictive models to encode *independencies* implied by the causal graph can lead to efficient and robust models. In this chapter, we focus on predictive models incorporating known *dependencies*. We study such models in the context of infectious disease prediction.

Preemptively identifying individuals at a high risk of contracting a contagious infection is important for guiding treatment decisions to mitigate symptoms, and preventing further spread of the contagion. In this chapter, we study how to build individual-level predictive models for contagious infections while explicitly addressing the challenges inherent to contagious diseases.

Building accurate infection prediction models is hindered by two main factors. First, contagious infections defy the usual *iid* assumption central to most machine learning methods. This is because an individual's infection state is not independent of other individuals' infection states. Previous work has often relied on expert knowledge to

construct exposure proxies (Wiens et al., 2012; Oh et al., 2018). It is then assumed that conditional on the exposure proxy and individual characteristics, individual outcomes are independent of one another. Such an assumption may be violated if the exposure proxy is noisy or misspecified leading to inaccurate predictions.

Second, the observed data is biased. The primary clinical purpose of testing for a disease is to provide guidance for treatment decisions for the individual being tested. Therefore, there is a strong bias in who is tested–people for whom knowing whether they have the disease will affect treatment (e.g., symptomatic individuals) are far more likely to be tested than other members of the population. But for many infectious diseases, only a fraction of those individuals carrying the pathogen experience noticeable symptoms. We use the term "incomplete testing" to describe the scenario where only a small, biased subset of infected individuals get tested. Incomplete testing makes learning accurate models difficult since the collected labels are missing not at random leading to biased, inconsistent estimates. At deployment time, we wish to apply the model to the full population of uninfected, and symptomatic individuals as well as asymptomatic carriers. Since uninfected individuals, and asymptomatic carriers are often under-represented at training time relative to deployment time, there is a distribution shift.

In this chapter, we leverage the non-independence of outcomes, and *a priori* knowledge of transmission patterns to construct robust predictors. Specifically, we use the knowledge that infections are caused by exposure to the pathogen through contacts to impute missing infection labels. Our proposed approach uses the fact that an individual's infection state provides useful information about their contacts' true infection states. This information is used to generate pseudo-labels for untested individuals, mitigating issues caused by incomplete testing. The key idea behind our approach is that highly structured patterns of contagion transmission can serve as a complementary signal to identify even untested carriers. The stronger that signal is, the less impact that incomplete testing will have. Our contributions can be summarized as follows:

1. We identify two properties of the collected data that can be exploited to mitigate the effects of incomplete testing.

2. We present an algorithm that leverages that insight to predict the probability of an untested individual carrying the disease.

3. We empirically evaluate the effectiveness of our method on both simulated data, and real data for a common and serious contagious disease. We show that predictions from our model can be used to inform efficient testing and isolation policies. Using Electronic Health Records (EHR) from a large hospital, we show that our model outperforms baselines on the task of predicting a healthcare associated infection.

4.1 Related work

Infectious disease modeling. Modeling the transmission of infectious diseases has been extensively studied in the epidemiology literature using SIS/SIR models and several other variants (Kermack and McKendrick, 1927). These epidemiological models focus on the *aggregate* levels of infections in a community. In contrast, we focus on predicting individual level infections. In the machine learning literature, previous work has has often relied on expert knowledge to construct exposure proxies (Wiens et al., 2012; Oh et al., 2018). They assume that conditional on the exposure proxy and individual characteristics, individual outcomes are independent of one another. Similar to our approach, Fan et al. (2016) and Makar et al. (2018) take into account structured data, namely contact networks to compute infection estimates. We differ from these approaches in that (1) we do not make parametric assumptions about the joint distribution of the observed or latent variables; instead we use nonparametric models (neural networks) to model the infection states, (2) we do not assume that all infections will become symptomatic as is done in Fan et al. (2016), and (3) unlike the approach taken by Makar et al. (2018), we model time evolving sequences of infections

taking into account the exposure states of potential asymptomatic carriers.

Semi-supervised learning. Our proposed approach relies on transductive reasoning to generate labels for untested individuals. In that, it is related to semi-supervised learning methods, such as pseudo-labeling (Lee, 2003), and self-training (Robinson et al., 2020). However, in traditional pseudo-labeling, the transductive power comes from the fact that points similar to each other in the input space have similar outputs. Here, the rich structure in the data allows for more: we can construct pseudo-labels for untested individuals not just by relying on their similarity to other labeled instances, but also by observing their observed contacts' infection states. Our empirical results, and analysis are similar in spirit to concepts presented in the semi-supervised literature, specifically the cluster assumption (Seeger, 2000; Rigollet, 2007), which we discuss later.

Graph Neural Networks. Our proposed approach incorporates knowledge of the contact network. In that it is similar to Graph Neural Networks (GNNs), which utilize relational data to generate prediction estimates (Zhou et al., 2018). GNNs fall into two categories. The first relies on transductive reasoning and cannot generalize to new communities (e.g., Kipf and Welling (2017)). The second relies on inductive reasoning, which can be used to generate estimates for previously unseen graphs (e.g., Hamilton et al. (2017)). Our work is similar to the latter category with an important distinction: our approach leverages unlabeled data giving more accurate, and robust estimates.

Our work can be viewed as combining the strengths of semi-supervised learning, and GNNs to address limited testing. We augment the strengths of those two approaches with ideas from domain shift and causal inference, such as importance weighting (Cortes et al., 2010b) to address biased testing.

4.2 Problem setting

Setup. Let $y^t \in \{0, 1\}$ denote an individual's true infection state at time t, with $y^t = 0$ if an individual is not infected and 1 if they are. We use $\overline{\mathbf{x}}^t \in \mathcal{X}^t$ to denote a vector of the individual's features at time t, and define J_i^t to be the set of indices of i's contacts at time t. We assume that contact network is fully observed, i.e., that the contact indices are known. We note that the assumption of fully observed networks is less likely to be violated in the context of hospital associated infections, where the majority of patients' interactions and contacts are routinely recorded. Our results on real data show that even with incomplete networks, our approach outperforms others.

Let $e_i^t \in \mathbb{R}_{\geq 0}$ denote i's exposure state at time t, with $e_i^t = \sum_{j \in J_i^t} y_j^t$. The exposure state is fully observed only when all of i's contacts have been tested, but otherwise either partially observed or unobserved. Define $\mathbf{x}^t = \overline{\mathbf{x}}^t || e^t$, where $||$ is the concatenation operator, i.e., $\mathbf{x}^t \in \mathcal{X}^t \times \mathbb{R}_{\geq 0}$. Let $o^t \in \{0, 1\}$ denote the observation state, with $o^t = 1$ if an individual's label is observed, i.e., if the individual has been tested for the infection. We use the super-script $:t$ to denote variables from time $t = 0$ up to and including t, e.g., $\mathbf{x}^{:t} = [\mathbf{x}^0, ..., \mathbf{x}^s, ..., \mathbf{x}^t]$.

Throughout, we use capital letters to denote variables, and small letters to denote their values. We use $P(\boldsymbol{X}^t, O^t, Y^{t+1})$ to denote the unknown distribution over the full joint. Under biased testing, we have that $P(\boldsymbol{X}^t | O^t = 1) \neq P(\boldsymbol{X}^t | O^t = 0) \neq P(\boldsymbol{X}^t)$. We assume that $0 < P(O^t = o | \boldsymbol{X}^t = \mathbf{x}) < 1$, for all $\mathbf{x} \in \mathcal{X}$, and $o \in \{0, 1\}$. This is the same as the overlap assumption in the causality literature. In addition, we assume that i's outcome is conditionally independent of i's contacts given $\mathbf{x}_i$ (which is itself a function of the contacts' outcomes). We consider the case where we have access to (1) a labeled (i.e., tested) set of individuals $\mathcal{D}_1 = \{\mathcal{D}_1^t\}_{t=0}^T = \{(\mathbf{x}_i^t, y_i^t), \dots (\mathbf{x}_{n_1^t}^t, y_{n_1^t}^t)\} \sim P(\boldsymbol{X}^t, Y^{t+1} | O^t = 1)$, and (2) an unlabeled (untested) set of individuals $\mathcal{D}_0 = \{\mathcal{D}_0^t\}_{t=0}^T = \{\mathbf{x}_i^t, \dots, \mathbf{x}_{n_0^t}^t\} \sim P(\boldsymbol{X}^t | O^t = 0)$, such that for each $i \in \mathcal{D}_0 \cup \mathcal{D}_1$, and each $t \in [0, T]$, we have that $J_i^t \in \mathcal{D}_0 \cup \mathcal{D}_1$. We use $\mathcal{U}^t$ to denote the set of indices of untested individuals at time t.

Notation	Meaning
y_i^t	i's infection state at time t
$\overline{\mathbf{x}}_i^t$	i's features at time t
e_i^t	i's (partially) observed exposure state at time t
$\mathbf{x}_i^t$	The concatenation of $\overline{\mathbf{x}}_i^t$, and e_i^t
$\mathbf{x}^{:t}$	The collection of an individual's features, and exposure states from time $t = 0$ till $t = t$, i.e., $\mathbf{x}^{:t} - [\mathbf{x}^0, ..., \mathbf{x}^s, ..., \mathbf{x}^t]$
J_i^t	The set of indices of i's contacts at time t
o_i^t	Observation state for the infection label. $o_i^t = 1$ if i's infection state is observed at time t (i.e., if i was tested for the infection at time t), and 0 otherwise
$\mathcal{D}_1$	Data $(\mathbf{x}, y$ tuples) for tested individuals
$\mathcal{D}_0$	Data $(\mathbf{x})$ for untested individuals
$w^t(\mathbf{x}_i^t)$	Probability that an individual with characteristics $\mathbf{x}_i^t$ gets tested
$\mathcal{U}^t$	the set of indices of untested individuals at time t.
$\mathcal{A}_i^t$	The set of ancestors of i at time t whose outcomes are unobserved i.e., $\mathcal{A}_i^t = J^t(i) \cap \mathcal{U}^t$

Table 4.1: Summary of notation used in chapter 4

Learning objective. We want to learn $f : \mathbf{x}^{:T} \to y^{T+1}$. To focus the discussion on the novel component of our approach, we first consider a setting in which we predict the outcomes for a single time step: making predictions for $t = 2$, using data from $t = 0, 1$, dropping the time superscript when it can be inferred from the context. We present the full model predicting infection sequences over time in section 4.4. Let ℓ be the logistic loss. Our goal is to find $f \in \mathcal{F}$, where $\mathcal{F}$ is some hypothesis space such that the risk of incorrectly classifying the infection state $R(f) = \mathbb{E}_{\boldsymbol{X},Y}[\ell(f(\boldsymbol{X}^t), Y^{t+1})]$ is minimized. We briefly consider a scenario where we have oracle access to the infection states of the untested population, but we return to the more realistic, non-oracle scenario later. Note that having access to the untested population's infection states implies that exposure states are also fully observed (by definition of the exposure states). Under the conditional independence assumption, we can view the risk as a sum of independent losses. Define the inverse probability of being tested as $w^t(\boldsymbol{X}) = {}^{P(O^t = o)}/_{P(O^t = o|\boldsymbol{X}^t)}$, following Robins (1998), and Robins

et al. (2000). Because of the overlap assumption, under biased testing we have that:

$$R(f) = R^{w^t}(f) = \mathbb{E}[w^t(\boldsymbol{X})\ell(f(\boldsymbol{X}), Y)], \tag{4.1}$$

(Cortes et al., 2010b). $R^{w^t}(f)$ cannot be directly computed since the expectation is defined with respect to the unobserved distribution. However, by Cortes et al. (2008) the following reweighted empirical loss is an unbiased estimator of $R^{w^t}(f)$:

$$\varepsilon(f) = \sum_{i \in \mathcal{D}_0^t \cup \mathcal{D}_1^t} w_i^t \ell(f(\mathbf{x}_i^t), y_i^{t+1}),$$

where $w_i^t = {}^{p(O^t = o_i^t)}/_{g(o_i^t|\mathbf{x}_i^t)}$, $p(O^t = o_i^t)$ is the empirical estimate of $P(O^t = o)$, and $g(o_i^t|\mathbf{x}_i^t)$ is the estimated probability of getting tested conditioned on individual characteristics. Without oracle access to untested individuals' infection states, we cannot directly minimize $\varepsilon(f)$ for $i \in \mathcal{D}_0^t$. In addition, without access untested individuals' infection states, the samples $\mathbf{x}^t \sim P(\boldsymbol{X}^t|O^t = 1)$ are incomplete. This is because $\mathbf{x}_i^t$ includes e_i^t, which is a function of $y_j^t : j \in J_i^t$. We only fully observe e_i^t, and hence $\mathbf{x}_i^t$ for individuals whose contacts have all been tested. To address this, we define Q as the set of all possible distributions over y_i^t for $i \in \mathcal{D}_0^t$. Our risk is now defined with respect to both Q, and f.

Let $\hat{y}_i \sim Q$, $\hat{e}_i^t = \sum_{j \in J^t(i)} \mathbb{1}\{j : o_j^t = 1\} \cdot y_j^t + \mathbb{1}\{j : o_j^t = 0\} \cdot \hat{y}_{i,j}^t$, $\hat{\mathbf{x}}_i = \overline{\mathbf{x}}_i^t || \hat{e}_i^t$, and $\hat{w}_i^t = {}^{p(O = o_i)}/_{g(\hat{\mathbf{x}}_i, o_i)}$, our task is to find Q and f, such that the following empirical risk is minimized:

$$\varepsilon(f, Q) = \sum_{i \in \mathcal{D}_1^t} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), y_i^{t+1}) + \sum_{i \in \mathcal{D}_0^t} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), \hat{y}_i^{t+1}). \tag{4.2}$$

We next consider how to leverage properties of the problem to efficiently minimize $\varepsilon(f, Q)$.

4.3 Exploiting structure as a regularizer

We seek to constrain the candidate sets $\mathcal{F}$ and Q to avoid overfitting. To do so, we exploit both the interdependence among individuals' infection states and the availability of unlabeled data. Recall that the exposure state of an individual is the sum of that individual's contacts' infection states. This means that when we draw $\hat{y}_i^t$ from Q, *we are implicitly drawing the exposure states for i's contacts', by definition of* $\hat{e}_i^t$. This becomes obvious if we decompose $\hat{e}_i^t$ as follows: $\hat{e}_i^t = \sum_{j \in J^t(i)} \mathbb{1}\{j : o_j^t = 1\} \cdot y_j^t + \mathbb{1}\{j : o_j^t = 0\} \cdot \hat{y}_{i,j}^t$, and $\hat{y}_{i,j}^t \sim Q$. This decomposition immediately implies two properties that should hold for "good" Q's. First, Q should assign infection states, $\hat{y}_i^t$, that are consistent with i's contacts' infection states. Consider the case where two individuals i, and j came into contact with each other at $t = t$. Suppose that j tests positive at time $t + 1$. For simplicity, suppose that i, and j have no contacts other than each other. Here Q should assign i a high probability of infection because in order to become infected j must have been exposed to the pathogen through i. Second, note that Q is assigning pseudo-labels for the infection states of untested contacts, this means that Q's imputed labels should be similar to the labels predicted by f. A good regularization method should then explicitly encourage the pseudo-labels to be similar to the estimated labels from f. This intuition is encoded in the main loss in our proposed approach:

$$f^*, Q^* = \min_{f,Q} \frac{1}{n_1^t} \sum_{i:o_i^t=1} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), y_i^{t+1}) + \frac{\lambda}{|J_i^t \cap \mathcal{U}^t|} \sum_{j \in J_i^t \cap \mathcal{U}^t} \hat{w}_j^{t-1} \ell(\mathbb{1}\{f(\mathbf{x}_j^{t-1}) > \tau\}, \hat{y}_{i,j}^t) \tag{4.3}$$

where $|.|$ denotes the set cardinality, $\lambda \geq 0$, and τ are parameters to be picked via cross validation, $\hat{y}_{i,j}^t \sim Q^*$, and $\hat{e}_i^t = \sum_{j \in J^t(i)} \mathbb{1}\{j : o_j^t = 1\} \cdot y_j^t + \mathbb{1}\{j : o_j^t = 0\} \cdot \hat{y}_{i,j}^t$. When $\lambda > 0$, this objective is somewhat similar to pseudo-labeling (Lee, 2003), it would encourage the votes of each of j's contacts to conform with the prediction from f, and implicitly with one another. When $\lambda = 0$, equation 4.3 prioritizes finding good predictions for the labeled data, ignoring possible structure implied by the data.

Note that in the second term in equation 4.3, we have $f(\mathbf{x}_j^{t-1})$, rather than $f(\hat{\mathbf{x}}_j^{t-1})$, meaning we assume no imputed exposure component for contacts at time $t-1$. This is only because we are considering the simple setting where $t-1=0$, i.e., $t-1$ is the beginning of the observation period and no exposure has happened yet. We later consider more complicated settings where the contacts' inputs also include an exposure state.

4.3.1 When does structure work as a regularizer?

We now ask: when do we expect equation 4.3 to yield models superior to those that ignore structure? First, if the imputed $\hat{y}_{\cdot,j}$ concentrates around significantly different values for $j : y_j = 1$, and $j : y_j = 0$, we expect minimizing equation 4.3 to yield better models. We stress that we do not require $\hat{y}$ to be an accurate estimate of the true labels, but only require that there is significant *separation* between the imputed values for untested-infected individuals and untested-uninfected individuals, i.e., they are distinguishable. This distinction means that even noisy and inaccurate estimates of $\hat{y}$ can be sufficient. In practice, high separability should occur, even in settings of low and biased testing, assuming the observed data satisfies a property we refer to as the **potency property**. The potency property can be viewed as an extension of the margin condition in classification (Tsybakov et al., 2004; Audibert et al., 2007). It implies that infections cluster so that infected-untested individuals tend to have many more infected contacts than do uninfected-untested individuals. Such a condition will be satisfied if the infection is sufficiently contagious.

Second, even if the imputed $\hat{y}$ allows high seperability, but $\hat{\mathbf{x}}$ makes it difficult to identify a learnable mapping from $\hat{\mathbf{x}}$ to $\hat{y}$, minimizing equation 4.3 instead of the objective on only the labeled data does not help. Such is the case when untested-healthy and untested-infected individuals "look" the same, meaning they have very similar characteristics and exposure states. This property is often referred to as the cluster assumption in semi-supervised learning literature (Rigollet, 2007; Seeger, 2000). The

cluster assumption states that individual characteristics, and exposure states tend to form near discrete clusters, with homogeneous labels within each cluster. Intuitively, it means that we can learn the correct clustering of individuals that separates infected from uninfected individuals, up to a permutation of the labeling. We refer to this property as the **dissimilarity property**.

The degree to which these two properties are satisfied in the observed data will depend largely upon the infection being studied and the environment in which it is spreading. However, as we show in section 4.5, even when these properties do not hold, our proposed approach performs as well as the best baseline. I.e., even in the worst case scenario, the regularization "does no harm."

4.4 Proposed method

Our proposed model, a Model for Infections under Incomplete Testing (MIINT) leverages labeled and unlabeled data to predict sequences of infections over time. MIINT minimizes a slight variant of equation 4.3, which is modified to predict the spread of infection over time. Let $\mathcal{A}_i^t$, be the set of ancestors of i at time t whose outcomes are unobserved, i.e., $\mathcal{A}_i^t = J^t(i) \cap \mathcal{U}^t$, $\mathcal{A}^{t-1} = \bigcup_{j \in \mathcal{A}_i^t} J^{t-1}(j) \cap \mathcal{U}^{t-1}$, etc. The loss at time t is defined as:

$$\mathcal{L}^t = \frac{1}{n_1^t} \sum_{i \in \mathcal{D}_1} \hat{w}_i^t \ell(f(\hat{\mathbf{x}}_i^t), y_i^{t+1}) + \sum_{s=0}^{t} \frac{\lambda}{|\mathcal{A}_i^t|} \sum_{j \in \mathcal{A}_i^s} \hat{w}_j^s \ell(\mathbb{1}\{f(\hat{\mathbf{x}}_j^s) > \tau\}, \hat{y}_{i,j}^s), \tag{4.4}$$

and the objective is to find f^*, Q^*, such that:

$$f^*, Q^* = \min_{f,Q} \frac{1}{T} \sum_t \mathcal{L}^t.$$

It is possible to consider the family of candidate functions $\mathcal{F}$ to be any family of non-parametric estimators. For our implementation, we take $\mathcal{F}$ to be the space of recurrent neural networks (RNNs). We assume that f does not vary over time (though that is

an assumption that could be relaxed). We propagate the predicted state forward in time, meaning f takes in $\mathbf{x}^t, e^t$ and $\hat{y}^t$ to predict $\hat{y}^{t+1}$. This ensures that exposures at time $< t$ are taken into account when predicting infections at time t. Note that equation 4.4 can be decomposed into the independent sums of individual losses, as well as their ancestors' losses. This means we can use stochastic gradient descent, with gradient updates defined with respect to mini-batches, as is typically done. One limitation is that equation 4.4 as stated would require keeping track of all the ancestors' states since $t = 0$, which can be prohibitive for long observation periods. In practice, one would consider a subset of $\mathcal{A}_i^t$ based upon the properties of the disease being studied.

The algorithm used to train MIINT, similar to pseudo-labeling (Lee, 2003), is an expectation maximization algorithm, where we iterate between computing the expected label for the untested samples (i.e., finding the optimal $\hat{Q}$, and identifying the optimal f that maximize the likelihood of the observed labels under $\hat{Q}$) until convergence. Convergence is achieved when the change in loss defined over the samples with observed labels in a held out validation set is $< \epsilon$ for some small ϵ. For our purposes, we find it sufficient to let Q be a deterministic function rather than an actual distribution. However, our approach is extendable to allow Q to be a distribution, for example using techniques described in Tran et al. (2017).

Finally, recall that we need to estimate $\hat{w}_i^t = {}^{p(O=o_i)}/_{g(\hat{\mathbf{x}}_i, o_i)}$. We follow Chernozhukov et al. (2017) in using an independent sample to estimate g. Importantly, g depends on $\hat{\mathbf{x}}$. So we follow an iterative process: after every epoch of training, we use the most updated f to estimate the unobserved outcomes in the validation set, and hence to get an estimate for $\hat{e}$ and $\hat{\mathbf{x}}$ for the independent weighting sample. We use these imputed values to learn an updated g. The updated g provides estimates for the weights of the training samples of the main prediction model, which are used to reweight the loss function for the next epoch, and so forth.

4.5 Experiments

We evaluate our model on a simulated, and a real data setting. All models presented in this chapter are implemented using Tensorflow (Abadi et al., 2016).

In the simulated setting, unlike the real data setting, we have access to the true infection state, which allows us to evaluate the performance of the model and baselines under different patterns of infection. In both settings, we present results from our model (MIINT) and five baselines:

1. Optimistic Model (**OM**): a model that assumes that all unobserved labels are equal to 0,
2. No Exposure Model (**NEM**): a model that ignores exposure, and attempts to predict infections solely based on the individual characteristics,
3. GraphSAGE (**GNN**): a graph neural network that takes into account the contact network, and observed infection states (Hamilton et al., 2017) but ignores untested individuals,
4. Pseudo-Labeling (**PL**): a semi-supervised learning method that takes into account untested individuals but ignores the graph structure (Lee, 2003),
5. ORacle Model (**ORM**): an unattainable model that has oracle access to the true labels for the whole population.

For all models, we weight the loss from each individual by the inverse of their estimated propensity to be tested, w_i^t, which is estimated using an independent sample following Chernozhukov et al. (2017). For our model, we use the iterative weighting technique outlined in section 4.4. For all these models, we keep the neural network architecture fixed. We use cross-validatation to get the values of λ, and τ. Results from unweighted models and details about cross-validation and network architecture are included in the appendix.

4.5.1 Simulation experiments

The simulation experiments demonstrate how MIINT can be used to inform testing and isolation policies that lead to reduction in infection rates, as well as empirically validate our conjectures regarding the conditions under which MIINT is expected to perform better than other methods.

Setup. We simulate a world in which there are three types of people: symptomatic if exposed (G_0), asymptomatic (but infected) if exposed (G_1), and immune (G_2). If exposed, individuals in group G_0 become infected and symptomatic, hence they are more likely to get tested. If exposed, individuals in group G_1 become infected without displaying symptoms. This group is unlikely to get tested. Finally, G_2, the immune group, is unlikely to get the infection even if exposed. To simulate individuals' characteristics (i.e., $\overline{\mathbf{x}}$), we map the distinct groups to the distinct MNIST digits, 0, 1, and 2. We use MNIST images because (1) they provide a complex input space compared to randomly generated data, and (2) images can be easily classified as similar or dissimilar, which enables us to design experiments where the dissimilarity property can be manipulated, as described later.

Let ν_i denote the pixels of an MNIST image i. For G_0 we randomly sample without replacement $n/3 \cdot T$ elements from the set $\{\nu_i\}_{i:d_i=0}$, where n is the total sample size, and T is the time horizon. For G_1, and G_2 we sample from $\{\nu_i\}_{i:d_i=1}$, and $\{\nu_i\}_{i:d_i=2}$, respectively. Note that the infection states will be different within each group, since infection also depends on the exposure state, and injected noise. We draw the edge sets $\{J^t(i)\}_{i\in n, t\in[0,T]}$ according to a stochastic block model, parameterized by the matrix B, where $B_{k,l}$ is the probability that an individual from G_k forms an edge with an individual from G_l. B is important in simulating different levels of carrier potency. When $B_{1,k}/B_{1,2}$ for $k = \{0, 1\}$ approaches 1, members of the asymptomatic carrier group are equally likely to form an edge with individuals who are not immune (G_0) as with individuals who are immune (G_2). In this setting, a carriers' contacts have a 50-50 percent chance of becoming infected, depending on whether they belong

to the immune group. This is a low carrier potency setting, which is unfavorable for our approach. On the other hand, if ${}^{B_{1,k}}/_{B_{1,2}} = 5$, for example, individuals in G_0, and G_1 are 5 times more likely to form an edge with someone in a susceptible group as compared to forming an edge with an individual in the immune group G_2. In this setting, the carriers' contacts are more likely to be from the susceptible group, hence they will have a higher chance of becoming infected. This a favorable high carrier potency setting.

We mimic the situation where testing started after a significant proportion of the population has been exposed by randomly setting the true exposure state of 20% of the population to be 1 at time $t = 0$. We set the exposure for each individual $e_i^t = \sum_{j \in J_i^t} y_j^t \geq 1$, and the true infection label $y_i^{t+1} = \mathbb{1}\{i \in (G_0, G_1)\} \cdot \mathbb{1}\{e_{i,t} = 1\}$. We introduce noise by randomly flipping the labels of 1% of the population. If an individual tests positive at time $t < T$, their label remains positive until $t = T$. We define $p_{\texttt{obs}}$ to be the proportion of the population tested (their true label is observed). We pick the probability of observing i's label based on i's true infection state, meaning, $p(o_i|y_i = 1) \neq p(o_i|y_i = 0)$. For all the simulations, we set $T = 6$ and we draw 500×6 samples for each of the training, validation, and testing sets. We simulate an independent sample to compute the weights w_i, so we also draw 500×6 samples that are used to train and validate weighting models. For each experiment, we draw 10 different datasets, and report the mean and standard deviation of the performance metric across the 10 draws.

In the first two experiment settings, we empirically validate our conjectures about the two properties that enable our model to outperform others, and explore what happens as these favorable properties are weakened to the point of non-existence.

Sensitivity to the potency property. Here, we fix $p_{\texttt{obs}} = .1$ and ${}^{p(o_i|y_i=1)}/_{p(o_i|y_i=0)} = 5$, and sweep over carrier potency by varying the value of ${}^{B_{1,k}}/_{B_{1,2}}$ from 1 (low potency) to 5 (high potency). Figure 4-1 shows ${}^{B_{1,k}}/_{B_{1,2}}$ on the $x-$axis and the AUROC on the $y-$axis. The plot shows that MIINT outperforms other baselines when there

is high potency, and as potency declines, its performance becomes similar to that of the other baselines. This supports our conjecture that our regularization approach is advantageous when the true infection states for an untested individual is strongly related to their contacts' infection states.

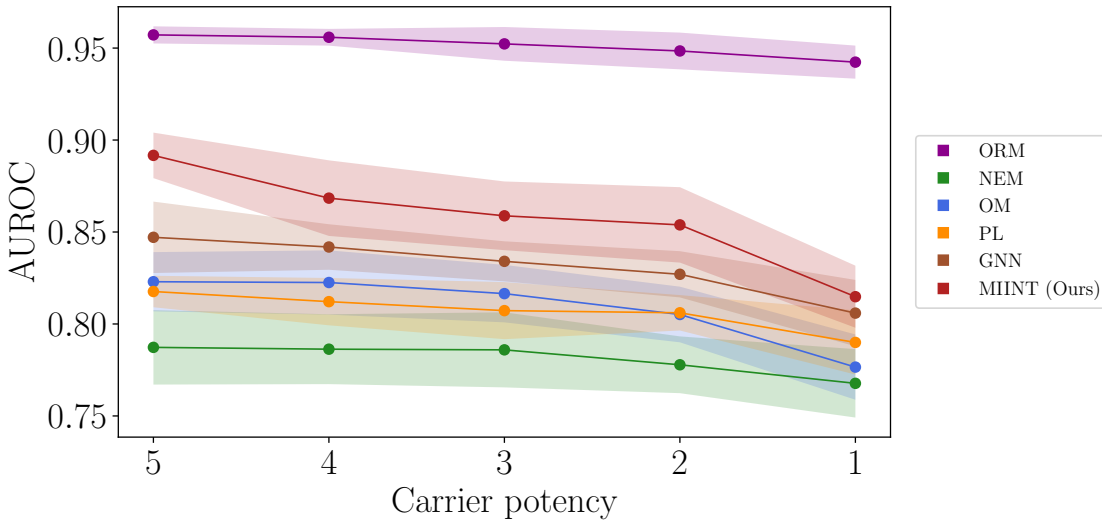

Figure 4-1: Impact of varying levels of carrier potency controlled by $^{B_{1,k}}/_{B_{1,2}}$. Our model outperforms baselines, especially in cases with high potency.

Sensitivity to the dissimilarity property. Here we examine what happens when the cluster assumption breaks down, meaning when untested individuals with similar characteristics have different infection states. We do so by moving the untested, and possibly infected[1] individuals to "look" similar to the immune individuals. Specifically, we sample pairs of images $\{(\nu_i, \nu_j)\}_{i,j:d_i=1,d_j=2}$. We then use VoxelMorph (Balakrishnan et al., 2018), a learning-based framework for deformable, pairwise image registration to learn a function that gives us a deformation field which we then apply it to pairs of images, moving ν_i to look more similar to ν_j. Using VoxelMorph in this way allows us to control the degree of similarity between images.

Figure 4-2 shows a sample image morphing for a pair of images using VoxelMorph.

[1]Individuals in G_1 are only infected if they get exposed.

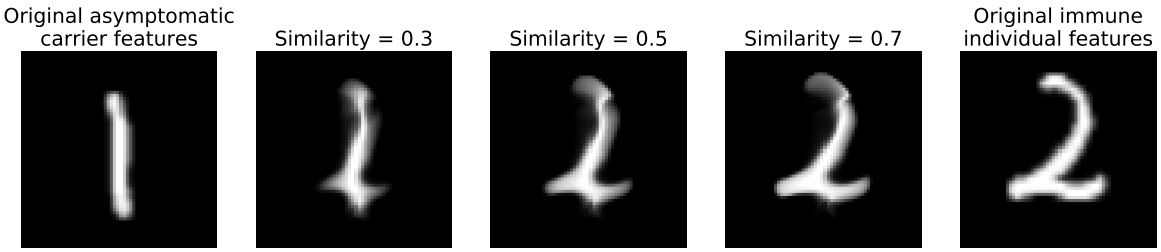

Figure 4-2: Varying similarity between asymptomatic carrier features and immune individual features using VoxelMorph (Balakrishnan et al., 2018)

Figure 4-3 shows the results of this setting. The $x-$axis can be viewed as the degree of similarity between the two untested groups with 0 being dissimilar (i.e., the original images without any deformation) and 1 being very similar (i.e., all images of the digit 1 look almost identical to 2's). The $y-$axis is the average AUROC. We see that all models perform worse as members in G_1 look more and more similar to those in G_2. We also see that MIINT outperforms all baselines when the two groups are dissimilar, and performs as well as the others when the mapping from input space to label becomes more difficult.

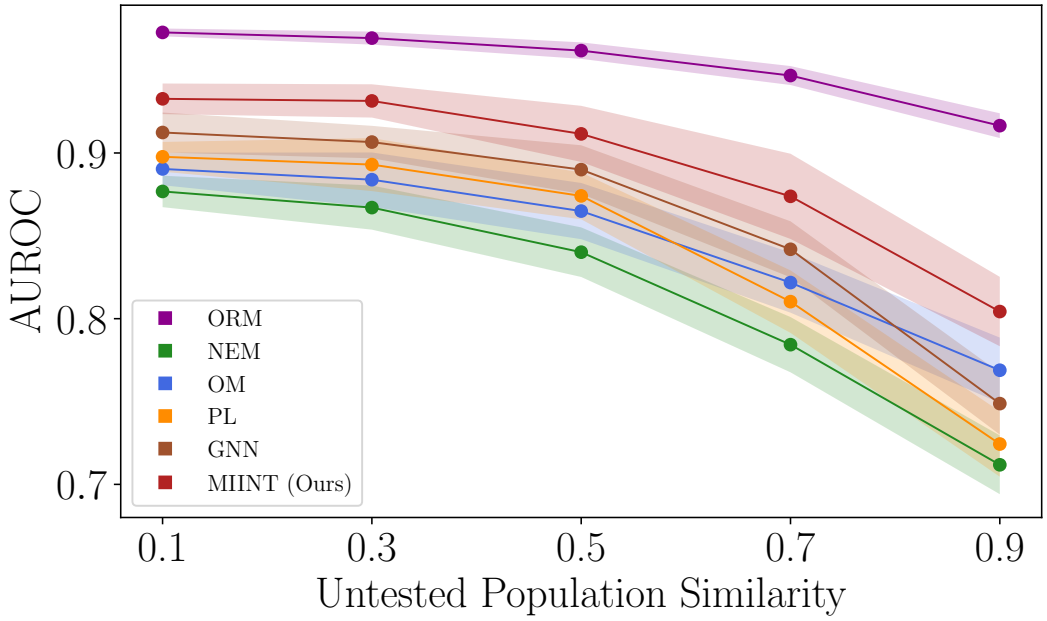

Figure 4-3: Impact of high (=.9) and low (=.1) similarity between the characteristics of the untested-uninfected and untested-infected populations. Our model outperforms baselines when the two populations are dissimilar.

The previous two experiments confirm our conjectures about the properties necessary for MIINT to perform well, and imply that MIINT "does no harm": at worst it performs comparably to alternatives, and at best it can give significantly better performance.

In the next two settings, we investigate the effect of bias and limited testing.

Impact of biased testing. To explore the impact of biased testing under favorable conditions, we create high potency by setting $B_{1,k}/B_{1,2} = 5$ for $k = \{0, 1\}$. We set $p_{\texttt{obs}} = .1$, and sweep over the odds of testing conditional on group membership. Results are shown in figure 4-4, where the $x-$axis shows the odds of testing $(= p(o_i|y_i = 1)/p(o_i|y_i = 0))$, and the $y-$axis shows the AUROC on the held-out test set, averaged over 10 simulations. We see that the weighted version of MIINT outperforms all others. This happens because NEM completely ignores exposure, OM assumes that 90% of the population $(1 - p_{\texttt{obs}})$ has $y_i = 0$ (which affects its estimate of e^t), whereas MIINT tries to impute the labels for those 90% based on their neighbors infection states. Here the difference between OM and NEM is not large because $p_{\texttt{obs}} = .1$, which is very low. This means that the exposure estimate that OM relies on is a poor estimate. Results from the subsequent experiment highlight that.

In addition, we see that PL has very high variance for highly biased testing. This makes sense because PL assigns labels for the untested population by considering similar patients in the tested population. Under highly biased testing, the labeled and unlabeled population are drastically different, making it difficult to generalize to the unlabeled population without leveraging the rich structured data.

Impact of limited testing. The setup for this experiment is similar to the previous one but here we fix the testing odds, $p(o_i|y_i = 1)/p(o_i|y_i = 0) = 5$, and sweep over the level of testing $p_{\texttt{obs}}$. Figure 4-5 shows the results, with $p_{\texttt{obs}}$ on the $x-$axis and the AUROC on the $y-$axis, averaged over 10 simulations. We see that weighted MIINT performs as well as the other models at the two extremes of testing levels, and does better at all other levels of testing. Here we see that OM outperforms NEW when the

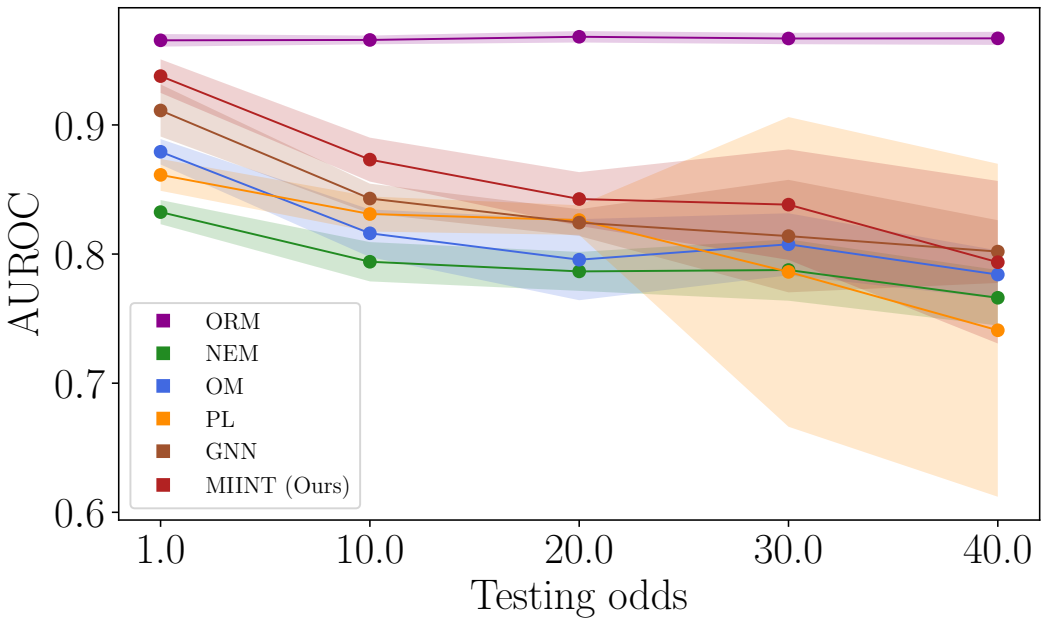

Figure 4-4: Impact of biased testing, $x-$axis shows the odds ratio of testing given characteristics ($= {}^{p(o_i|y_i=1)}/_{p(o_i|y_i=0)}$), 1 implies randomized testing. Our model does better than baselines for most levels of bias, and similar to baselines at extreme bias.

level of testing is sufficiently high, which is expected since OM inherently assumes no unobserved infections. As the testing levels increase, that assumption becomes more correct. The performance of NEM also improves with higher levels of testing since it has access to a cleaner y label, however, it is never able to perform as well as MIINT or OM because it does not take exposure as an input.

In the final set of experiments, we show the utility of having an accurate infection prediction tool in curbing the spread of infections.

Informing testing and isolation policies. We highlight how our model can inform efficient testing and isolation policies. We simulate biased and limited testing by setting ${}^{p(o_i|y_i=1)}/_{p(o_i|y_i=0)} = 5$, and $p_{\texttt{obs}} = .1$ respectively. We set ${}^{B_{1,k}}/_{B_{1,2}} = 5$, making it a high potency setting where MIINT is expected to perform well. We mimic a situation where no isolation interventions are taken at training time. At deployment time, we fix a testing budget of at most $p_{\texttt{test}}\%$ of the total population on each time step. We use the predictions from each model to inform who gets tested by picking the top $p_{\texttt{test}}\%$ with the highest predicted probability of infection.

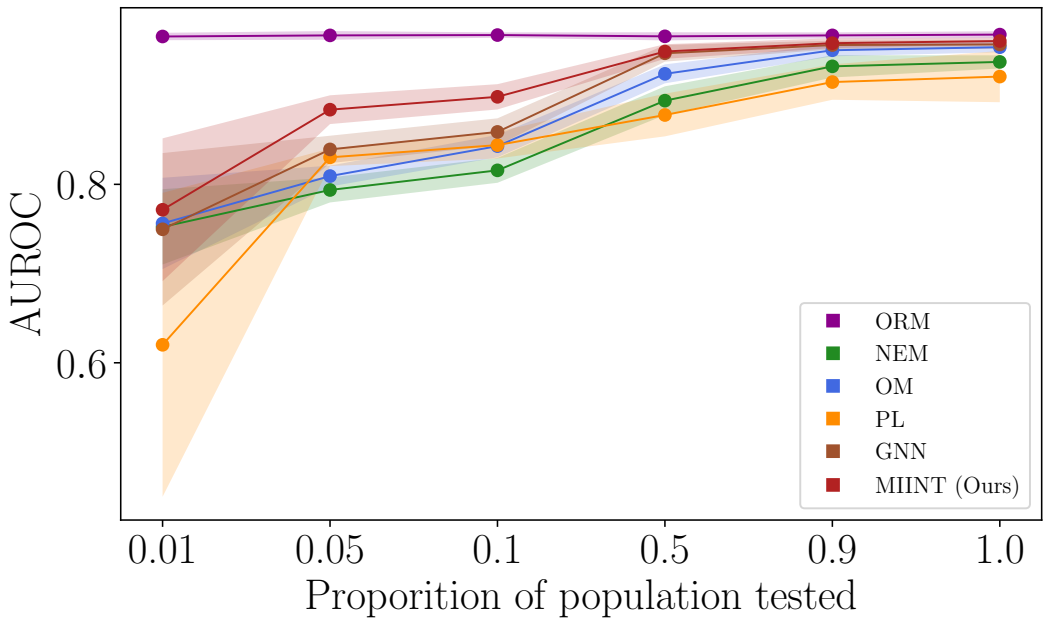

Figure 4-5: Impact of limited testing. Our model does better than baselines at every level of testing. Our model achieves near oracle accuracy at low levels of testing bias, and high proportion tested.

Of those tested, individuals who are truly infected are "isolated" by setting their edges for the subsequent time steps to 0. They are also taken out of the population eligible for further testing. We compute the infection rate, π_M for a model M as $\pi_M = n^{-1} \cdot \sum_i \max_t y_{i,t}$. We define π_0 as the infection rate under a no-action policy, that is if no isolation interventions are taken. Our main metric of interest is the reduction in infection rate relative to the no-action policy $= {}^{\pi_0 - \pi_M}/_{\pi_0}$. Figure 4-6 shows the reduction in infection rate on the $y-$axis for different values of the testing budget $p_{\texttt{test}}\%$ on the $x-$axis. In addition to the main baselines, we also show results from a random testing policy. The results show that for any given testing budget, our model outperforms all feasible baselines leading to uncovering more individuals who should be isolated, thus achieving a higher reduction in infection rates. The results imply that our model is able to achieve near oracle infection control with 70% testing, compared to $\approx 90\%$ for the baselines.

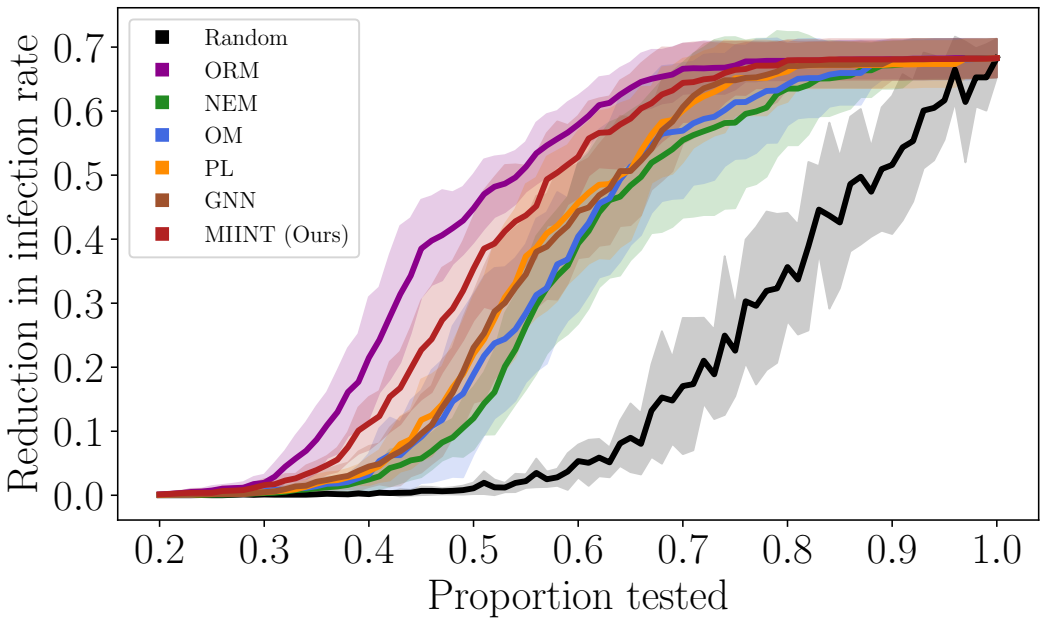

Figure 4-6: Reduction in infection rates relative to a policy that does not isolate infections (no-action policy) as the daily testing budget varies. Our model achieves the highest reductions in policy relative to all realistic (i.e., non-oracle) models.

4.5.2 Real data experiment

Here, our task is to predict the onset of Clostridioides difficile infections (CDI) among patients in a large urban hospital. CDI is a contagious bacterial infection that attacks the gut, and causes over 300,000 infections annually in the US (Magill et al., 2014). As with most contagious infections, asymptomatic carriers of CDI exists and can contribute to the spread of the infection (Riggs et al., 2007).

Setup. Using Electronic Medical Records of a large urban hospital, we extract daily characteristics of patients who were admitted to the hospital between 09/01/2012 and 06/01/2014. We follow similar inclusion criteria as Oh et al. (2018); Makar et al. (2018), outlined in detail in the appendix. We collect all patient characteristics available upon admission (e.g., gender, age, medical history) as well as daily characteristics (e.g., lab tests). We collect contact networks, where an edge exists if two patients are in the same room on the same day or if they came into contact with the same nurse on the same day.

Here, we have partial access to the true infection states, since not all the patients are tested, making accurate evaluation of different models difficult. However, the hospital's testing protocols allow us to construct a proxy "true" label and a proxy "observed" label. In this hospital, whether a patient is diagnosed as CDI positive or not is a result of a one or two-step testing protocol. In the first step, an enzyme immunoassay (EIA) and Glutamate dehydrogenase (GDH) test are conducted. If the results of the two tests are discordant, a second step of testing is conducted. Specifically, the hospital uses a polymerase chain reaction (PCR) assay to act as a tie-breaker. Previous studies comparing the outcomes of the two groups (those who have non-discordant EIA/GDH+ results vs. discordant and PCR+) have shown that the former experiences more severe complications (Origüen et al., 2018; Polage et al., 2015). This implies the EIA/GDH+ label can act a proxy for symptomatic infections, whereas PCR+ might be picking up on patients who are carrying the bacteria but have low toxin levels and therefore mild or no symptoms.

In this experiment, we hide the PCR+ labels during training, presenting them as untested individuals to all models. At test time, however, the true positives are defined to be patients who tested positive with either EIA/GDH or PCR. In addition to the baselines outlined in section 4.5.1, we allow one of the models full access to the EIA/GDH+ and PCR+ labels, and refer to it as a "partial oracle" model (POM) since it has access to the PCR+ labels, but not the full infection states. The latter are unavailable because the majority of patients in the hospital are not tested. We also compare our results to the state-of-the-art prediction model for CDI (Wiens et al., 2012), which is a logistic regression model that takes into account the varying importance of different risk factors over the hospitalization, and relies on medical knowledge to construct exposure proxies. We refer to this model as the Expert driven Logistic Regression (ELR).

We split the data into 5 subsets based on time. The first subset holds 6 months of data and is used to train the main infection prediction models. The second and third subsets contain 5 months of data each, and are used for validation and testing of the

	TPR@ FPR=10%	AUROC
POM	0.49 (0.014)	0.73 (0.003)
NEM	0.33 (0.008)	0.69 (0.006)
NEM–U	0.45 (0.009)	0.7 (0.006)
OM	0.44 (0.008)	0.74 (0.006)
OM–U	0.45 (0.012)	0.7 (0.005)
ELR	0.53 (0.008)	**0.82 (0.006)**
GNN	0.24 (0.005)	0.59 (0.005)
GNN–U	0.22 (0.007)	0.55 (0.007)
PL	0.26 (0.008)	0.62 (0.005)
PL–U	0.58 (0.012)	0.78 (0.006)
MIINT–U	**0.6 (0.007)**	0.81 (0.006)
MIINT	0.51 (0.007)	0.78 (0.007)

Table 4.2: Performance metrics for CDI prediction on the test set.

main prediction model. The last 2 subsets are used for training and validation of the weighting models, and each contain 2 months worth of data. We report the AUROC, the True Positive Rate (TPR) at the threshold which achieves a False Positive Rate (FPR) of 10% on the test set.

Table 4.2 shows the results on the test set. We present the results from all the models which incorporate weighting by w_i^t, and their unweighted counterparts, with the suffix $--U$. Standard deviations are calculated by taking 100 bootstrap replicates of the test set data. For several models, the unweighted model outperforms its weighted counterpart. We see that unweighted MIINT outperforms almost all others on both reported metrics. The one exception is ELR: MIINT and ELR achieve comparable AUROCs but MIINT has a significantly better TPR. MIINT outperforms POM even though the latter has access to better labels. We hypothesize that this is because in addition to accurately estimating the PCR+ patients, MIINT is also capturing truly untested infections, and utilizing these estimates to accurately impute the exposures of the EIA/GDH+ patients as well as the PCR+ patients.

4.6 Summary

We presented MIINT, a model that predicts contagious infections. Unlike other models, MIINT works well even when labels are generated using biased and limited testing. It does so by exploiting the fact that, in practice, data related to contagious diseases are not *i.i.d.* The key idea is that highly structured patterns of contagion transmission can serve as a complementary signal to identify even untested carriers. The stronger that signal is, the less impact that biased and incomplete testing will have.

We identified two properties that determine the extent to which MIINT outperforms other approaches. The first states that the more contagious the infection, the better MIINT performs. The second is the degree to which characteristics of untested and infected individuals and characteristics of the untested and healthy individuals form discrete clusters–an important property in general for semi-supervised learning.

We showed empirically that MIINT can be used to guide testing policies that lead to reduced infection rates, and that even if the two properties outlined above are absent, MIINT still performs well. In an experiment using EHR data, we showed that MIINT outperforms baselines when when used to predict CDI.

In conclusion, we believe this work is a first step down an important path. If predictive models are to play a useful role in limiting the spread of contagious infections, they must take into account the interdependence of outcomes, and the fact that untested individuals are capable of spreading the disease before they have been diagnosed.

Chapter 5

Conclusion

More than ever, society is relying on models developed using observational data to guide various decision-making processes. In this thesis, we present an approach to building robust and sample-efficient models by combining ideas from machine learning and causal inference.

We highlight some of the challenges that render learned models unfit for decision guidance, and how the work presented in this thesis tackles them.

Inaccurate causal inference because of limited data. Estimation of conditional average treatment effect (CATE) is commonly used as the basis for contextual decision making in fields such as healthcare, education, and economics. In many cases, the CATE is defined by a function that is complex, and hard to estimate accurately from finite, or limited data. In chapter 2 we make the observation that it is often sufficient for the decision maker to have estimates of upper and lower bounds on the potential outcomes of decision alternatives. We show that, in such cases, we can improve sample efficiency by estimating simple functions that bound these outcomes instead of estimating their conditional expectations. Our analysis highlights a trade-off between the complexity of the learning task and the confidence with which the learned bounds hold. Guided by these findings, we develop an algorithm for learning

upper and lower bounds on potential outcomes that optimize an objective function defined by the decision maker, subject to the probability that bounds are violated is small. Using a clinical dataset and a well-known causality benchmark, we demonstrate that our algorithm outperforms baselines, providing tighter and more reliable bounds. Our contributions are especially useful when the amount of training data available is limited.

Predictors that fail to generalize beyond the training distribution. Predictors that can reliably guide decision making are often expected to be invariant to inconsequential changes in the data generating process. However, many predictors constructed from deep neural networks (DNNs) lack robustness under distribution shift (Beery et al., 2018; Ilyas et al., 2019; Azulay and Weiss, 2018; Geirhos et al., 2018), including naturally occurring distribution shifts(Taori et al., 2020). In chapter 3, we study a flexible, causally-motivated approach to enforcing such invariance to distribution shifts. Our approach uses auxiliary labels, typically available at training time, to enforce conditional independences between the latent factors that determine these labels. We show both theoretically and empirically that causally-motivated regularization schemes (a) lead to robust estimators that generalize well under distribution shift, and (b) have better finite sample efficiency compared to usual regularization schemes, even in the absence of distribution shifts.

Inaccurate infection prediction caused by asymptomatic carriers By identifying individuals at elevated probability of being infected, ML algorithms can inform decisions on whom to isolate to control the spread of an infectious disease. One of the main challenges of building ML models for accurate infection prediction is the presence of a distribution shift caused by asymptomatic carriers, who are under-represented at training time because of biased testing. In chapter 4, we show that we can build reliable infection prediction models even when the observed data is collected under limited, and biased testing that prioritizes testing symptomatic individuals. Our analysis suggests that when the infection is highly contagious, incomplete testing might be sufficient to achieve good out-of-sample prediction error. Guided by this insight,

we develop an algorithm that predicts infections, and show that it outperforms baselines on simulated data. We apply our model to data from a large hospital to predict Clostridioides difficile infections; a communicable disease with high morbidity that is characterized by asymptomatic (i.e., untested) carriers. Using a proxy instead of the unobserved untested-infected state, we show that our model outperforms benchmarks in predicting infections.

5.1 Future directions in machine learning for causal inference

Minimal causal estimates. Estimating the conditional average treatment effect (CATE) is often hailed as the golden ticket to guiding personalized intervention decisions. However, in cases where the CATE is complex and hard to learn from finite data, insisting on estimating accurate CATE makes causal inference an inefficient and unreliable tool, and hinders its practical adoption. The path for widespread use of causal inference tools starts with the realization that every decision is different, and there should not be a one size fits all approach to building models; in some cases aiming to estimate the CATE might be necessary, but in most cases *minimal* estimates of the causal effect of an intervention might be sufficient.

The work presented in chapter 2 is one example of tailoring the estimation question to fit the specific needs of the decision maker. The guiding principles behind this work can be utilized to design efficient, and useful causal effect estimation methods in many practical settings. For example, ranking the causal effect of different intervention choices on an outcome of interest is often useful when considering multiple treatment choices. In that setting, we can get an accurate ranking without estimating the CATE for each one of the treatment choices. Viewing causal estimation through the lens of minimalism sets us on a path where model-driven evidence-based decision making can become a reality not a distant ambition.

Causal models for infectious diseases. Making intervention decisions in the context of infectious diseases is particularly challenging because one patient's treatment affects more than just that patient's outcomes, it also affects their contacts' outcomes and intervention decisions. Estimating the causal effect of interventions in this context is challenging for the same reasons: the traditional assumptions about the independence of each unit of treatment (i.e., each patient) that enable consistent estimation are violated.

Existing work in the area of causal inference on networked-data has focused on asymptotic consistency of causal estimators (e.g., Ogburn et al. (2017); Sofrygin and van der Laan (2016)) but no finite sample analysis of this set of causal problems exists to date. Establishing the finite sample proprieties of causal estimators in the context of networked data is an important future direction that will enable us to understand how well these estimators generalize to different contact networks, and different diseases with varying levels of contagiousness.

Because of the challenging nature of causal estimation on networked data, aiming to estimate the precise CATE of different interventions is often infeasible, even with abundant data. Developing methods, and theoretical analysis for minimal causal estimates in this context could enable physicians, and key decision makers to design effective interventions that curb the spread of infections.

5.2 Future directions in causal inference for machine learning

Causally-motivated regularization. Limiting model capacity to avoid overfitting is one of the most fundamental core concepts in ML. Historically, we have relied on mechanical regularization to control model capacity, for example limiting the norm of the weights parameterizing a model, choosing a larger kernel bandwidth to ensure

that a model is sufficiently smooth, and so forth. As deep learning method gain more popularity, mechanical regularization schemes are losing favor for at least two reasons. First, they are challenging to explain, and justify: do we need to regularize the weights of all the layers to get a sufficiently complex model? Does regularizing the weights of the final layer suffice? Second, as our empirical analysis in chapter 3 highlights, mechanical regularization might not lead to robustness against distribution shifts: limited capacity models do not always translate to robustness.

Against this backdrop, the appeal of conceptual, or causally-motivated regularization schemes is increasing. Methods that limit the model capacity by ensuring that it obeys known independencies (chapter 3) or dependencies (chapter 4) allow efficient and robust estimation. However, little is known about the theoretical properties of causally-motivated regularization schemes. Rigorous comparisons between different methods of enforcing invariance are lacking. For ML, and specifically deep learning, to deliver on its promise of reliable and efficient predictors we need to have a better theoretical, and empirical understanding of how causally-motivated regularization works.

Auditing predictive models. The impressive infiltration of ML tools into key areas of our lives, from housing and insurance to personalized medicine and criminal justice has brought issues of fairness, and interpretability to the forefront. As ML becomes more widely used, we must address whether the core concepts that a model encodes conform to societal standards, and code of ethics. This in turn means that we need to be able to be able to explain the core paradigms encoded in the inner workings of ML models.

Viewed through a causal lens, the question of looking into the inner workings of a deep neural network, for example, is equivalent to asking "what is the causal graph implied by this neural network?" A promising direction here is adapting methods of causal discovery (Glymour et al., 2019) to learn the causal graphs implied by learned models. The road to discovering the causal structures implied by ML models starts

with improving the existing causal discovery methods. For example, much is still unknown about the finite sample properties of different causal discovery algorithms.

5.3 Future directions in healthcare

Developing novel methods backed up by rigorous theory in machine learning and causal inference is important in its own right. However, for these tools to have a significant impact in healthcare settings, they need to address real medical needs. One of the most promising directions in healthcare research is designing decision-guidance tools that *complement* rather than replace existing human expertise. For example, a physician might be well trained in giving prognoses based on the patient's symptoms, but she cannot quickly ascertain the viral load that her patient has been exposed to through the patient's daily contacts.

To complement existing expertise, we must first understand what they are, and where the areas of true need exist. We conclude this thesis with the remark that meaningful progress in the area of machine learning, causality and healthcare has to be the result of deep and continuing collaborations between computer scientists and medical experts. To move forward, we must first understand each other's concerns, needs, limitations, and abilities.

Appendix A

Appendix to chapter 2

A.1 Proof of theorem 2.3.1

Before delving into the proof, we define the empirical proportion overestimated:

Definition A1. *For $f \in \mathcal{F}$, $\gamma > 0$, a sample $z = \{x_i, y_i\}_i^n$ drawn from a fixed but unknown distribution p_t, known weights $\boldsymbol{w}$, we define the empirical risk when the distribution with respect to p:*

$$\underline{\epsilon}_f^{\boldsymbol{w}}(z, \gamma) = \sum_i w(x) \mathbb{1}\{\underline{r}_f(x, y) < \gamma\}.$$

To construct the proof, we will first study the overestimation risk when there are no training set violations (Lemma A3). To extend our results to cases where there are training set violations, we rely on a technique, presented in Shawe-Taylor and Cristianini (2002) and used in Schölkopf et al. (2001), which allows us to ignore small violations in the training data at the cost of a more complex function space. This function space (formally defined in definition A2) is constructed by creating an "auxiliary function" that picks specific points to have a non-zero violation. Its complexity depends on the allowable violations. By augmenting the result from lemma A3 with

the auxiliary function space, we get theorem A1, a general version of theorem 2.3.1, which gives a bound on the overestimation risk for general sturdy function spaces. Finally, we give the proof for linear function spaces, which is presented in theorem 2.3.1 in the main text.

To build up to lemma A3, we restate the following two previously established results.

Lemma A1. *Due to Shawe-Taylor and Williamson (1999): Let $\mathcal{F}$ be a sturdy function class, then for each $N \in \mathbb{N}^+$ and any fixed sequence $X \in \mathcal{X}^n$ the infimum*

$$\inf\{\gamma : \mathcal{N}(\gamma, \mathcal{F}, X) < N\}$$

is attained

We assume that f_l^1, f_l^0, f_l^0 and f_u^0 belong to a sturdy function class, as defined in definition 2.3.4.

The following lemma due to Cortes et al. (2010a) bounds the second moment of the weighted loss.

Lemma A2. *Due to Cortes et al. (2010a). For $x \in \mathcal{X}$, a weighting function w_t on $\mathcal{X}$, a loss function ℓ, and some function $f \in \mathcal{F}$, the second moment of the importance weighted loss can be bounded as follows:*

$$\mathbb{E}_{X|T}\left[w_t^2(X)\ell_f^2(X) \mid T = t\right] \leq d_2(p||p_t)$$

We now study the overestimation error when there are no training set violations, i.e., when $D = 0$. A direct analogy can be drawn between the following lemma (lemma A3) and hard margin one-class SVMs studied in Schölkopf et al. (2001), whereas theorem 2.3.1 is analogous to the soft margin case.

Lemma A3. *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in (2.1). For $f_l^t \in \mathcal{F}$,*

and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_t}(z, f_t^1, \gamma) = 0$. With a probability $1 - \delta$ over the draw of random samples, we have that:

$$\underline{R}_{f_t^l}(\gamma) \leq \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}. \tag{A.1}$$

where, for $t \in \{0, 1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma, \mathcal{F}, 2n_t) \right\rceil .$$

Proof. For a given $f_l^1 \in \mathcal{F}$:

$$\begin{aligned} P\Big(\underline{R}_{f_l^1}(\gamma) - \underline{\epsilon}^{\boldsymbol{w}}_{f_l^1}(z, \gamma) > \varepsilon\Big) &= P\Big(\underline{R}_{f_l^1}(\gamma) > \varepsilon\Big) \\ &\leq 2P\Big(\underline{\epsilon}^{\boldsymbol{w}'}_{f_l^1}(z', \gamma) > \frac{\varepsilon}{2}\Big), \end{aligned}$$

where the equality follows from the fact that the empirical error on the estimation data will always be 0 by definition of γ. And the inequality follows from applying the double (ghost) sample trick. Suppose that such an f_l^1 exists. Pick a fixed k such that

$$\gamma_k = \inf\{\gamma : \mathcal{N}(\gamma, \mathcal{F}, 2n_1) \leq 2^k\} \leq \gamma \ .$$

By Lemma A1, and assumption of sturdiness, we have that this γ_k exists. Consider the γ_k-covering, U. There exists another $f_\bullet \in U$ such that the distance between f_l^1 and $f_\bullet$ is $\leq \gamma_k \leq \gamma$, meaning $f_\bullet$ satisfies:

$$P\Big(\underline{\epsilon}^{\boldsymbol{w}'}_{f_l^1}(z', \gamma) > \frac{\varepsilon}{2}\Big) = P\Big(\underline{\epsilon}^{\boldsymbol{w}'}_{f_\bullet}(z', 0) > \frac{\varepsilon}{2}\Big)$$

This limits the complexity of the function class from infinite to having a covering number $= \mathcal{C}_{\mathcal{F}}^{\gamma}$. Swapping samples between the estimation and the ghost sample, this will create a random variable $S' = \frac{1}{M}(\underline{\epsilon}^{\boldsymbol{w}'_1}_{f_\bullet}(z'_1, 0) + \ldots + \underline{\epsilon}^{\boldsymbol{w}'_m}_{f_\bullet}(z'_m, 0), + \ldots + \underline{\epsilon}^{\boldsymbol{w}'_M}_{f_\bullet}(z'_M, 0))$ for $M = 2^{n_1}$, where the subscripts of $\boldsymbol{w}'$ and z' denote the sample index. Note that $\mathbb{E}_{x\sim p_t}[S'] = \underline{R}_{f_\bullet}(0)$ and let S denote $S' - \mathbb{E}_{x\sim p_t}[S']$, with $\mathbb{E}_{x\sim p_t}[S] = 0$. Let $\sigma^2(S) =$

$\mathbb{E}[S^2] = \mathbb{E}[(S' - \mathbb{E}_{x\sim p_t}[S'])^2]$. By Lemma A2, we have that $\sigma^2(S') \leq d_2(p||p_1) - \underline{R}_{f_\bullet}(0)^2$. By Bernstein's inequality:

$$P\Big(\underline{R}_{f_\bullet}(0) - \underline{\epsilon}^{w'}_{f_\bullet}(z', 0) > \frac{\varepsilon}{2}\Big) \leq \exp\Big(\frac{-3n_1\varepsilon^2}{24\sigma^2(S) + 4C_1\varepsilon}\Big),$$

and a union bound over the function space:

$$\begin{aligned} &P\Big(\underline{R}_{f_\bullet}(0) - \underline{\epsilon}^{w'}_{f_\bullet}(z', 0) > \frac{\varepsilon}{2}\Big) \leq \\ &\mathcal{N}(\gamma, \mathcal{F}, 2n_1) \exp\Big(\frac{-3n_1\varepsilon^2}{24\sigma^2(S) + 4C_1\varepsilon}\Big) \end{aligned}$$

Putting it all together:

$$\begin{aligned} &P\Big(\underline{R}_{f_l^1}(\gamma) - \underline{\epsilon}^{w}_{f_l^1}(z, \gamma) > \varepsilon\Big) \\ &\leq 2P\Big(\underline{R}_{f_\bullet}(0) - \underline{\epsilon}^{w'}_{f_\bullet}(z', 0) > \frac{\varepsilon}{2}\Big) \\ &\leq 2\mathcal{N}(\gamma, \mathcal{F}, 2n_1) \exp\Big(\frac{-3n_1\varepsilon^2}{24\sigma^2(S) + 4C_1\varepsilon}\Big) \end{aligned}$$

Setting $\delta(\epsilon)$ to match the upper bound, inverting w.r.t. ϵ and removing the (negative) term $\underline{R}_{f_\bullet}(0)^2$ from the right-hand side, we get that stated bound with probability $1 - \delta$. □

Next, we define the auxiliary function space, which will allow us to study non-zero training set violations.

Definition A2. *[Restated from Schölkopf et al. (2001), definition 13] Let $L(\mathcal{X})$ be the set of real valued, non-negative functions f on $\mathcal{X}$ with support $supp(f)$ countable, that is the functions in in $L(\mathcal{X})$ are non-zero for at moust countably many points. We define the inner product of two functions $f, g \in L(\mathcal{X})$ by:*

$$f \cdot g \sum_{x\in supp(f)} f(x)g(x).$$

The 1-norm on $L(\mathcal{X})$ is defined by $||f||_1 = \sum_{x\in supp(f)} f(x)$. Let $L^D(\mathcal{X}) := \{f \in$

$L(\mathcal{X}) : ||f||_1 \leq D\}$. *Define a transformation, or embedding of* $\mathcal{X}$ *into the product space* $\mathcal{X} \times L(\mathcal{X})$ *as follows:*

$$\varpi : \mathcal{X} \to \mathcal{X} \times L(\mathcal{X})$$
$$\varpi : x \to (x, \Delta_x),$$

where

$$\Delta_x = \begin{cases} 1, & y = x, \\ 0, & otherwise \end{cases}$$

For a function $f \in \mathcal{F}$ *a set of training examples* z *of size* n*, define the function* $g_f \in L(\mathcal{X})$

$$g_f(\mathbf{y}) := \sum_{x,y \in z} w_1(x) \min\{0, \gamma - \underline{r}_{f_l^1}(x, y)\} \Delta_x(\mathbf{y}),$$

where $\mathbf{y} = \{y_i\}_{i=1}^n$

We can now state the risk of overestimation for general sturdy functions.

Theorem A1. *Let* $\mathcal{F}$ *be any sturdy function class defined over input space* $\mathcal{X}$*,* $z = \{x_i, y_i\}_{i:t_i=t}$*, where* $x_i, y_i \sim p_t(X, Y)$*, and* C_t *be as defined in* (2.1)*. For* $f_l^t \in \mathcal{F}$*, and any* $\gamma > 0$*, let the associated* $\underline{D}^{\boldsymbol{w}_t}(z, f_t^1, \gamma) = D > 0$*. With a probability* $1 - \delta$ *over the draw of random samples, we have that:*

$$\underline{R}_{f_t^l}(\gamma) \leq \frac{4C_t(k_t + \log \frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log \frac{1}{\delta})}{n_t}}. \tag{A.2}$$

where, for $t \in \{0, 1\}$*,*

$$k_t = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2n_t) + \log \mathcal{N}(\gamma/2, L^D(\mathcal{X}), 2n_t) \right\rceil .$$

Proof sketch. The proof extends lemma A3, replacing the function class $\mathcal{F}$ with the function class of the augmented space, that is $\mathcal{F}+L(\mathcal{X}) := \{f+g : f \in \mathcal{F}, g \in L(\mathcal{X})\}$. The details of the proof are identical to theorem 14 in Schölkopf et al. (2001), and are hence omitted.

The following lemma, restated from Shawe-Taylor and Cristianini (2002) gives a bound on the auxiliary function complexity for linear functions (defined in kernel spaces).

Lemma A4. *Due to Shawe-Taylor and Cristianini (2002). For $D > 0$, all $\gamma > 0$:*

$$\begin{aligned}&\log \mathcal{N}(\gamma, L^D(\mathcal{X}), n)\\ &\leq \left\lfloor \frac{D}{2\gamma} \right\rfloor \log\left(\frac{\exp(n + {}^{D}/_{2\gamma} - 1)}{{}^{D}/_{2\gamma}}\right)\end{aligned}$$

Finally, by replacing the auxiliary function term from theorem A1 (that is $\log \mathcal{N}({}^{\gamma}/_{2}, L^D(\mathcal{X}), 2n_t)$) with its bound for linear functions acquired from lemma A4 (that is $\log \frac{\exp(n_t + {}^{D}/_{\gamma} - 1)}{{}^{D}/_{\gamma}}$), we get the proof for theorem 2.3.1.

A.2 Proof of corollary

Corollary A1 (Restated Corollary 2.3.1)**.** *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z_t = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in expression (2.1). For $f_l^1, f_u^0 \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_1}(z_1, f_l^1, \gamma) = D_1 > 0$, and $\overline{D}^{\boldsymbol{w}_0}(z_0, f_u^0, \gamma) = D_0 > 0$ Define $\tilde{\tau}_l := f_l^1 - f_u^0$. With probability $1 - \delta$ over random samples, we have that:*

$$\underline{R}_{\hat{\tau}_l}(\gamma) \leq \sum_t \frac{4C_t(k_t + \log \frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log \frac{1}{\delta})}{n_t}}. \tag{A.3}$$

where, for $t \in \{0, 1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2n_t) + \log \mathcal{N}(\gamma/2, L^{D_t}(\mathcal{X}), 2n_t) \right\rceil .$$

Proof. Consider the event:

$$E = \left\{ x : \tau(x) < \tilde{\tau}_l(x) - 2\gamma \right\}$$

where $x \sim p$. Note that event E implies that one of the following two events must hold:

$$E_1 = \left\{ (x, y) : \underline{r}_{f_l^1}(x, y) < \gamma \right\}$$

for $t = 1$.

$$E_0 = \left\{ (x, y_0) : \overline{r}_{f_u^0}(x, y) < \gamma \right\}$$

for $t = 0$.

Note that $p(E_1) = \underline{R}_{f_l^1}(\gamma)$. So, theorem A1 implies that

$$p(E_1) \leq \frac{4C_t(k_t + \log \frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log \frac{1}{\delta})}{n_t}}$$

for k_t as defined in theorem A1. Similarly $p(E_0) = \overline{R}(f_u^0)$, and by a similar construction can obtain the bound on $p(E_0)$. Using a union bound we have that

$$\begin{aligned} p(E) &= p(E_1 \cup E_0) = p(E_1) + p(E_0) - p(E_1 \cap E_0) \\ &\leq p(E_1) + p(E_0), \end{aligned}$$

which completes the proof. □

A.3 Proof of Theorem 2.3.2

To build up to the proof of theorem 2.3.2, we first seek a bound on the fat-shattering dimension of functions defined in definition 2.3.7. This bound is constructed in a similar spirit to theorem 1.6 in Bartlett and Shawe-Taylor (1999). Specifically, to get a bound on the fat-shattering dimension, we rely on the lemmas A1 and A2. The former shows that the sum of any shattered set is far from the remainder of that set, the latter shows that the same sums cannot be too far apart.

Lemma A1. *Let $\mathcal{F}_u, \mathcal{F}_l, A, B$ be as defined in definition 2.3.7. Let $I = \{x_i\}_{i=1}^n$, where $x_i \sim p(X,Y)$.For a fixed $\gamma > 0$, if I is $\gamma-$shattered by $\mathcal{F}_l$ then every subset $I' \in I$ satisfies:*

$$\min_{q\in\{p,2\}} \left\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \right\|_q \geq \frac{2n\gamma}{A+B}$$

Proof. If I is γ shattered by $\mathcal{F}_l$, denote the corresponding "witness" vector by $\{s_i\}_{i=1}^n$, then for all $\boldsymbol{\sigma} = \{\sigma_1 \dots \sigma_i \dots \sigma_n\}$ there is an f with $\|f_l\| \leq A$ such that $\sigma_i \cdot (\theta^\top x_i - s_i) \geq \gamma$ for $i = 1 \dots n$. Suppose that:

$$\sum_{i\in I'} s_i \geq \sum_{i\in I\setminus' I} s_i \tag{A.4}$$

Then fix $\sigma_i = 1$ if $i \in I'$. In that case we have that

$$\langle f_l, x_i \rangle \geq s_i + \gamma \qquad \forall i \in I' \tag{A.5}$$

$$\langle f_l, x_i \rangle < s_i - \gamma \qquad \forall i \in I \setminus I'. \tag{A.6}$$

Pick $f_u \in \mathcal{F}_u$ such that $||f_u - f_l||_p = B' \leq B$, and:

$$\langle f_u - f_l, x_i \rangle \geq s_i + \gamma \qquad \forall i \in I' \tag{A.7}$$

$$\langle f_u - f_l, x_i \rangle < s_i - \gamma \qquad \forall i \in I \setminus I'. \tag{A.8}$$

Showing that such a function exists is trivial: simply take $f_u := f_l$. For that we have $||f_u - f_l|| = 0 \leq B$, which means that the function does exist in $\mathcal{F}_u$.

From expression A.5, we have that:

$$\Big\langle f_l, \sum_{i \in I'} x_i \Big\rangle = \sum_{i \in I'} \langle f_l, x_i \rangle \geq \sum_{i \in I'} s_i + Card(I')\gamma,$$

where $Card(.)$ denotes the cardinality. Similarly for $I \setminus I'$, we have that

$$\Big\langle f_l, \sum_{i \in I \setminus I'} x_i \Big\rangle < \sum_{i \in I \setminus I'} s_i + Card(I \setminus I')\gamma$$

Combining the expressions for I' and $I \setminus I'$, and from expression A.4:

$$\Big\langle f_l, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\rangle \geq n\gamma. \tag{A.9}$$

We now construct the same arguments for the distance. Let $f_d := f_u - f_l$. From expression A.7, we have that:

$$\Big\langle f_d, \sum_{i \in I'} x_i \Big\rangle = \sum_{i \in I'} \langle f_d, x_i \rangle \geq \sum_{i \in I'} s_i + Card(I')\gamma,$$

and from expression A.8:

$$\langle f_d, \sum_{i \in I \setminus I'} x_i \rangle < \sum_{i \in I \setminus I'} s_i + Card(I \setminus I')\gamma$$

Combining the two, and from expression A.4:

$$\langle f_d, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \geq n\gamma. \tag{A.10}$$

Putting expressions A.9 and A.10 together,

$$\langle f_l, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \tag{A.11}$$

$$+ \langle f_d, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \geq 2n\gamma. \tag{A.12}$$

Note that by Cauchy-Schwartz,

$$\begin{aligned}
\langle f_l, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle &\leq \|f_l\| \left\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \right\| \\
&\leq A \left\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \right\| \\
&\leq A \min_{q \in \{p,2\}} \left\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \right\|_q .
\end{aligned}$$

and,

$$\begin{aligned}\Big\langle f_d, \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i\Big\rangle &\le \|f_d\|_p \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_p \\ &\le B' \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_p \\ &\le B \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_p \\ &\le B \min_{q\in\{p,2\}} \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_q .\end{aligned}$$

For expression A.11 to hold:

$$\begin{aligned}& A \min_{q\in\{p,2\}} \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_q \\ & + B \min_{q\in\{p,2\}} \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_q \ge 2n\gamma \\ & (A+B) \min_{q\in\{p,2\}} \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_q \ge 2n\gamma \\ & \min_{q\in\{p,2\}} \Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\|_q \ge \frac{2n\gamma}{(A+B)},\end{aligned}$$

which completes the proof.

Lemma A2. *Let $\mathcal{F}_u, \mathcal{F}_l, r$ be as defined in definition 2.3.7. Let $I = \{x_i\}_{i=1}^n$, where $x_i \sim p(X,Y)$.For a fixed $\gamma > 0$, if I is $\gamma-$shattered by $\mathcal{F}_l$ then every subset $I' \in I$ satisfies:*

$$\Big\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \Big\| \le \sqrt{n} r$$

The proof is identical to Lemma 1.3 in Bartlett and Shawe-Taylor (1999), and is hence omitted.

Lemma A3. *Let $\mathcal{F}_u, \mathcal{F}_l, A, B, r$ be as defined in definition 2.3.7. For a fixed $\gamma > 0$, the $\gamma-$fat shattering dimension of $\mathcal{F}_l$ can be bounded as follows:*

$$fat(\gamma, \mathcal{F}_l) \leq \left(\frac{r \cdot (A+B)}{2\gamma} \right)^2$$

Combining the results from Lemmas A2 and A1, we get that:

$$\begin{aligned} \frac{2n\gamma}{A+B} &\leq \min_{q \in \{p,2\}} \left\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \right\|_q \\ &\leq \left\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \right\| \leq \sqrt{n} r, \end{aligned}$$

which gives us that:

$$\sqrt{n} \leq \frac{r(A+B)}{2\gamma},$$

which completes the proof.

Theorem A1. *Let $\mathcal{F}_l^t$, $\mathcal{F}_u^t$, A, B, and r be as defined in definition 2.3.7, z, and D as defined in theorem 2.3.1,and C_t be as defined in expression (2.1). For $f_l^t \in \mathcal{F}_l^t$, $f_u^t \in \mathcal{F}_u^t$ and any $\gamma > 0$, with a probability $1-\delta$ over the draw of random samples, we have that:*

$$\underline{R}_{f_t^l}(\gamma) \leq \frac{4C_t(k_t + \log \frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log \frac{1}{\delta})}{n_t}}. \tag{A.13}$$

where, for $t \in \{0, 1\}$,

$$\begin{aligned} k_t = \Bigg\lceil \left(\frac{2r(A+B)}{\gamma} \right)^2 \log \left(\frac{8n_t(b-a)^2}{\gamma^2} \right) \\ \log \left(\frac{4en_t(b-a)\gamma}{r^2(A+B)^2} \right) + \frac{D}{\gamma} \log \frac{e(n_t + {}^{D}/_{\gamma} - 1)}{{}^{D}/_{\gamma}} \Bigg\rceil . \end{aligned}$$

Using Corollary 3.8 Shawe-Taylor et al. (1998), we can $\log \mathcal{N}({}^{\gamma}/_{2}, \mathcal{F}, 2n_t)$ by its fat shattering dimension. Combining the results from lemma A3 and theorem 2.3.1, we get the final result.

A.4 Equivalence to quantile regression

Consider the following problem

$$\begin{aligned} \underset{f_u, f_l}{\text{minimize}} \quad & \ell^{(1)}_{\tilde{w}}(f_u(x_i), f_l(x_i)) \\ \text{subject to} \quad & \sum_{i:t_i=t} \tilde{w}_{t_i} \max[y_i - f_u(x_i), 0] \leq \beta \\ & \sum_{i:t_i=t} \tilde{w}_{t_i} \max[f_l(x_i) - y_i, 0] \leq \beta \\ & f_u(x_i) \geq f_l(x_i), \quad \forall i : t_i = t \end{aligned} \tag{A.14}$$

Theorem A1. *Assume that* (A.14) *is strictly convex and has a strictly feasible solution. Then, for any fixed quantile* $t \in (0.5, 1)$*, there is a parameter* $\beta \geq 0$ *such that the minimizer of* (A.14) *with weighted absolute loss and the minimizer of the werighted quantile loss, for quantiles* $(t, 1-t)$ *with non-crossing constraints, are equal and have false coverage rate* $1 - q$*.*

Proof. Problem (A.14) with absolute loss $\ell(y, y') = |y - y'|$ can be stated as

$$\begin{aligned} \underset{f_u, f_l}{\text{minimize}} \quad & \sum_{i:t_i=t} \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)| \\ \text{subject to} \quad & \sum_{i:t_i=t} \tilde{w}_{t_i} \max[y_i - f_u(x_i), 0] \leq \beta \\ & \sum_{i:t_i=t} \tilde{w}_{t_i} \max[f_l(x_i) - y_i, 0] \leq \beta \\ & f_u(x_i) \geq f_l(x_i), \quad \forall i : t_i = t \end{aligned}$$

Let $Q_\beta(f_u, f_l) = \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)|$ denote the objective and F the feasibility region. Introducing Lagrange multipliers for the first two constraints, we obtain the

regularized objective

$$
\begin{aligned}
L(f_u, f_l, \lambda_u, \lambda_l) = &\sum_{i:t_i=t} \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)| \\
&+ \frac{\lambda_u}{n} \sum_{i=1}^{n} \max(y_i - f_u(x_i), 0) - \beta \\
&+ \frac{\lambda_l}{n} \sum_{i=1}^{n} \max(f_l(x_i) - y_i, 0) - \beta
\end{aligned}
$$

and by convexity and strict feasibility, strong duality holds through Slater's condition,

$$
\min_{u,l \in F} Q_\beta(u, l) = \max_{\lambda_u, \lambda_l \geq 0} \min_{u \geq l} L(u, l, \lambda_u, \lambda_l) \ .
$$

By strict convexity, for each $\beta \geq 0$, the minimizers u^*, l^* on either side are equal for the maximizers λ_u^*, λ_l^*. Now, consider the following objective, equivalent in minima to $\tilde{L}(f_u, f_l, \lambda_u, \lambda_l)$,

$$
\begin{aligned}
\tilde{L}(f_u, f_l, \lambda_u, \lambda_l) := &\sum_{i:t_i=t} \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)| \\
&+ \lambda_u \sum_{i:t_i=t} \tilde{w}_{t_i} \max(y_i - f_u(x_i), 0) \\
&+ \lambda_l \sum_{i:t_i=t} \tilde{w}_{t_i} \max(f_l(x_i) - y_i, 0)
\end{aligned}
$$

We can separate $\tilde{L}$ into terms for which $y_i \geq f_u(x_i)$ and $y_i \geq f_l(x_i)$ respectively, adding and subtracting $\sum_i y_i$

$$
\begin{aligned}
&\tilde{L}(f_u, f_l, \lambda_u, \lambda_l) \\
&= (\lambda_u - 1) \sum_{y_i \geq u(x_i)} \tilde{w}_{t_i} (y_i - f_u(x_i)) - \sum_{y_i < f_u(x_i)} \tilde{w}_{t_i} (y_i - f_u(x_i)) \\
&+ (1 - \lambda_l) \sum_{y_i \geq f_l(x_i)} \tilde{w}_{t_i} (y_i - f_l(x_i)) - \sum_{y_i < f_l(x_i)} \tilde{w}_{t_i} (y_i - f_l(x_i))
\end{aligned}
$$

Now, let $\lambda_u = \lambda_l = 1/(1-q)$ for $q \in (0, 1)$, which means $(1 - q) \geq 0$. Multiplying by

$(1-q)$ leaves us with

$$
\begin{aligned}
&\tilde{L}(f_u, f_l, \lambda_u, \lambda_l) \\
&\propto \sum_{y_i \geq f_u(x_i)} q \cdot \tilde{w}_{t_i}(y_i - f_u(x_i)) + \\
&\sum_{y_i < f_u(x_i)} (q-1) \cdot \tilde{w}_{t_i}(y_i - f_u(x_i)) \\
&+ \sum_{y_i \geq f_u(x_i)} (1-q) \cdot \tilde{w}_{t_i}(y_i - f_l(x_i)) \\
&+ \sum_{y_i < f_u(x_i)} (-q) \cdot \tilde{w}_{t_i}(y_i - f_l(x_i)) \\
&\propto \sum_{i:t_i=t} \tilde{w}_{t_i} \max[q(y_i - f_u(x_i)), (q-1)(y_i - f_u(x_i)] \\
&+ \sum_{i:t_i=t} \tilde{w}_{t_i} \max[(1-q)(y_i - f_l(x_i)), (-q)(y_i - f_l(x_i)] \\
&= \sum_{i:t_i=t} \rho^{(q)}_{\tilde{w}_{t_i}}(y_i - f_u(x_i)) + \rho^{(1-q)}_{\tilde{w}_{t_i}}(y_i - f_l(x_i)) \ ,
\end{aligned}
$$

where $\rho^{(q)}_{\tilde{w}}$ is the weighted quantile loss for quantile q. Recalling that our original problem had the constraint $f_u(x_i) \geq f_l(x_i)$, we recover the non-crossing constraint. □

A.5 Experiments

A.5.1 Cross-validation details

For our BP method, we have 5 hyperparameters to pick. These are α, the regularization parameter, the kernel bandwidth, β_u and β_l which are the allowed violations. The last parameter, $\gamma_{BP} > 0$, as described in section 2.4.3. Note that the kernel bandwidth is only relevant for the experiments done on the ACIC data, but not the IST experiments since a linear kernel is used in the latter.

For the kernel regression (KR), we first split the training data into 2. On the first half, we do the typical 3-fold cross-validation to pick the model that minimizes the weighted empirical error. This allows us to pick the kernel bandwidth, and a regularization parameter the is multiplied by the L2 norm of the weights. Again, the kernel bandwidth is only relevant for the experiments done on the ACIC data, but not the IST experiments since a linear kernel is used in the latter. The intervals are then estimated in one of two ways. For KR-MI, we use the second part of the training data to estimate the residuals. We follow algorithm 2 in Lei et al. (2018) to get the final interval estimates. For KR-γ, we use the second half of the training data to estimate the FCR, $\hat{\nu}_{\gamma_{\text{KR}}}$, with γ_{KR} defined as the "shifting" parameter, where $\tilde{f}_u^{KR}(x_i) = \tilde{\mu}_t(x_i) + \gamma_{KR}$ and $\tilde{f}_l^{KR}(x_i) = \tilde{\mu}_t(x_i) - \gamma_{KR}$, for $\tilde{\mu}_t(x_i)$ being the predicted response value. We then pick the smallest γ_{KR} that does not violated the required FCR.

For the Gaussian process (GP), we pick the kernel bandwidth, the noise level added to the diagonal of the kernel. For BART models, we use the BartMachine package in R Kapelner and Bleich (2016). We do 3 fold cross-validation to pick the parameter k, which controls the prior probability that $\mathbb{E}(y|x)$ is contained in the interval (ymin,ymax), based on a normal distribution. We set the number of trees to be 200, since that did not seem to affect the results. For the CMGP, we pick the lengthscale of the RBF kernels of the two response surfaces as well as the variance and correlation parameters.

A.5.2 Small ACIC data results including CCI

Figures A-1, and A-2 are similar to figures 2-6 and 2-7 presented in the main text but they include the performance of CCI models.

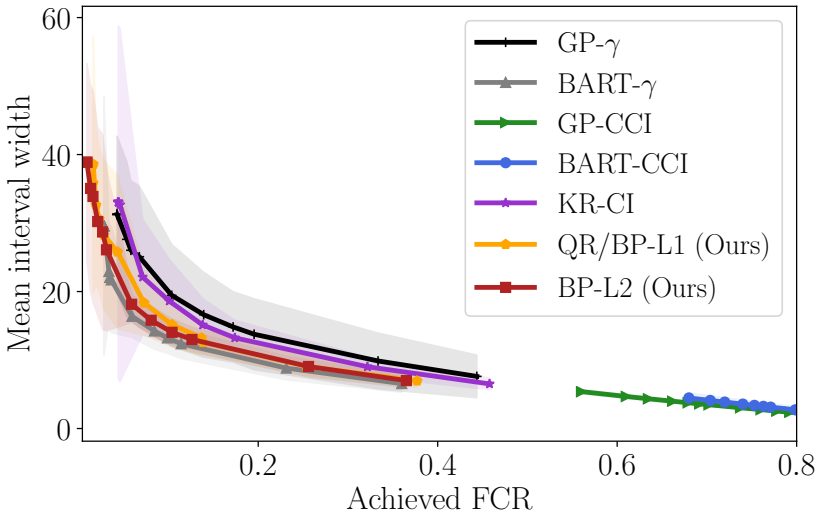

Figure A-1: Comparing tightness of estimated intervals

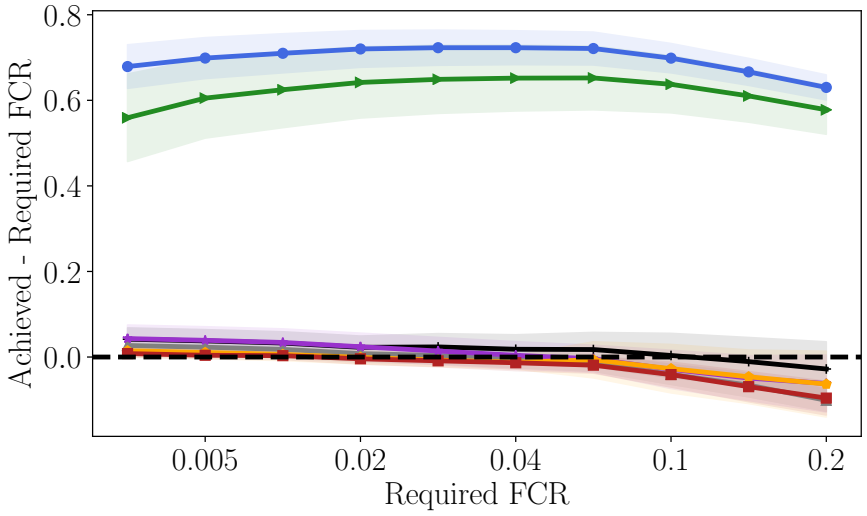

Figure A-2: Comparing violation to the required FCR. Legend is the same as that in figure A-1

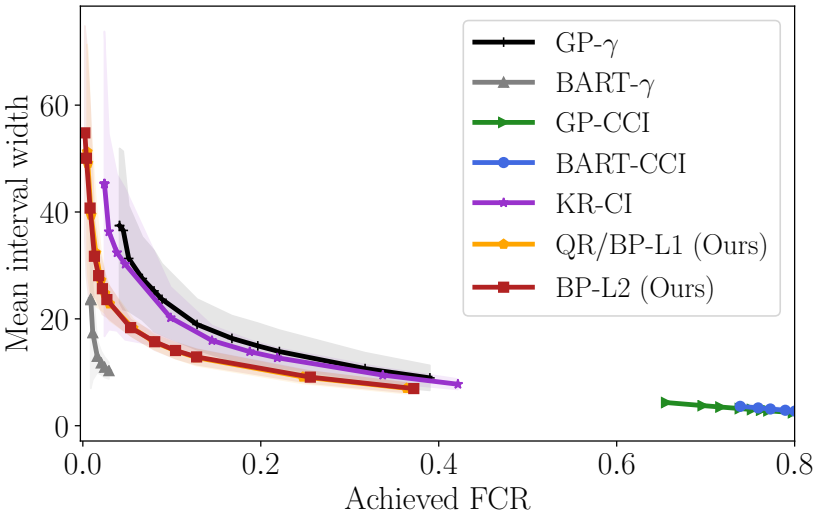

Figure A-3: Comparing tightness of estimated intervals

A.5.3 Large ACIC data results including CCI

Figures A-3, and A-4 are similar to figures 2-9 and 2-10 presented in the main text but they include the performance of CCI models.

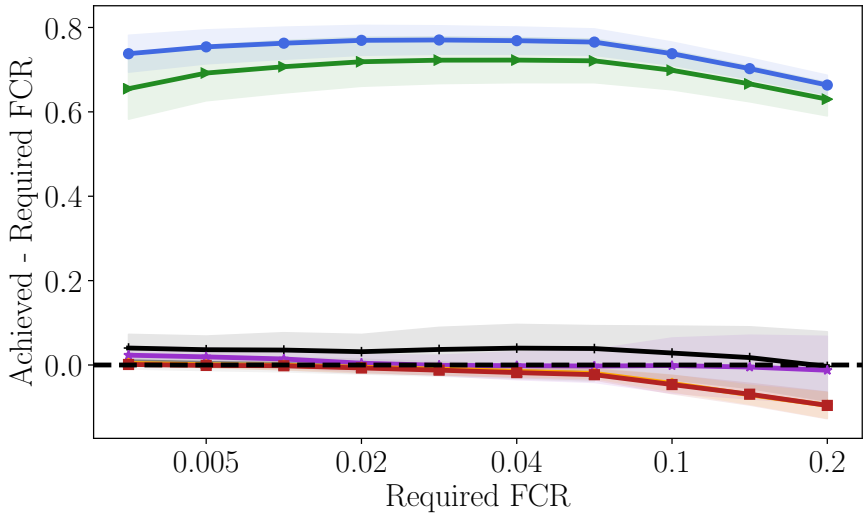

Figure A-4: Comparing violation to the required FCR. Legend is the same as that in figure A-4

Appendix B

Appendix to chapter 3

B.1 Proofs for section 3.2

Proposition A1 (Restated proposition 3.2.1)**.** *Under P°, the Bayes optimal predictor is (i) only a function of $\mathbf{X}^*$, and (ii) an optimal risk-invariant predictor f_{rinv} with respect to $\mathcal{P}$.*

Proof. Under P°, $\mathbf{X}^*$ d-separates Y from $\mathbf{X}$, so $\mathbb{E}_{P^\circ}[Y \mid \mathbf{X}] = \mathbb{E}_{P^\circ}[Y \mid \mathbf{X}^*]$. Thus, the population risk minimizer is only a function of $\mathbf{X}^*$.

By the assumption 3.2.1, we have that $\mathbf{X}^* = e(\mathbf{X})$ and hence $\mathbf{X}^*$ can be perfectly recovered from $\mathbf{X}$. This means that $\mathbb{E}_{P^\circ}[Y \mid \mathbf{X}^*]$ can be written as a function of $\mathbf{X}$, i.e., we can define $g(\mathbf{X}) = \mathbb{E}_{P^\circ}[Y \mid e(\mathbf{X})]$. Thus, the Bayes optimal classifier $f(\mathbf{X})$, which is a function of $\mathbb{E}_{P^\circ}[Y \mid \mathbf{X}] = \mathbb{E}_{P^\circ}[Y \mid e(\mathbf{X})]$, can be written (with some abuse of notation) as $f(\mathbf{X}^*)$ (that is, a function that only varies with the value of $\mathbf{X}^* = e(\mathbf{X})$).

Thus, the risk is also invariant. To see that note the following:

$$\begin{aligned} R^{\circ}(f) &= \int_{\mathbf{X},Y} \ell(f(\mathbf{X}), Y) P^{\circ}(\mathbf{X} \mid \mathbf{X}^*, V) P^{\circ}(\mathbf{X}^* \mid Y) P^{\circ}(Y) P^{\circ}(V) dY d\mathbf{X} \\ &= \int_{\mathbf{X}^*,Y} \ell(f(\mathbf{X}^*), Y) P^{\circ}(\mathbf{X}^* \mid Y) P^{\circ}(Y) dY d\mathbf{X}^* \\ &= \int_{\mathbf{X}^*,Y} \ell(f(\mathbf{X}^*), Y) P(\mathbf{X}^* \mid Y) P(Y) dY d\mathbf{X}^* \\ &= R_P(f), \end{aligned}$$

for any $P \in \mathcal{P}_t$.

Because this classifier is optimal under P°, no other risk invariant classifier can obtain a lower risk across $\mathcal{P}$; thus this classifier is an optimal risk invariant classifier. □

B.2 Proofs for section 3.3

We show that the reweighted risk is an unbiased estimator of the risk under P°, i.e., that

$$\mathbb{E}_{Ps}\left[\hat{R}^{\mathbf{u}}_{Ps}(f)\right] = R_{\circ}(f).$$

For any P_s, the $\mathbf{u}$-weighted risk is equal to the risk under the corresponding unconfounded distribution P°. That is, $R^{\mathbf{u}}_{Ps} := \mathbb{E}_{Ps}[u(Y,V)\ell(f(\mathbf{X},Y))] = R_{P^{\circ}}$.

To see this, note that the conditional distribution $P_s(\mathbf{X} \mid Y, V)$ is invariant across the family $\mathcal{P}$ defined in (3.1). Thus, the risk conditional on Y and V, $R_{Ps|y,v} :=$

$\mathbb{E}_{Ps}[\ell(f(\mathbf{X}), Y) \mid Y = y, V = v]$, does not change with P_s.

$$
\begin{aligned}
R^u_{Ps} &:= \mathbb{E}_{Ps}[u(Y,V)\ell(f(\mathbf{X},Y))] = \mathbb{E}_{Ps}[\mathbb{E}_{Ps}[u(Y,V)\ell(f(\mathbf{X},Y)) \mid Y = y, V = v]] \\
&= \sum_{y,v} P_s(Y = y, V = v)u(y,v)R_{P|y,v} = \sum_{y,v} P_s(Y = y)P_s(V = v)R_{Ps^\circ|y,v} \\
&= \mathbb{E}_{P^\circ}[\mathbb{E}_{P^\circ}[R_{P^\circ|y,v}]] = R_{P^\circ}.
\end{aligned}
$$

B.3 Proofs for section 3.4.1

Proposition A1. *(Restated proposition 3.4.1). Let $f(\mathbf{x}) = \sigma(\phi(\mathbf{x})) = \sigma(\mathbf{w}^\top \mathbf{x})$ be a function contained in $\mathcal{F}_{L_2,\text{MMD}}$. Then,*

$$\|\mathbf{w}_\perp\| \leq \frac{\tau}{\|\Delta\|}. \tag{B.1}$$

Proof. Note that τ must be non-negative. If not, let ω' be the function that achieves the max difference $\tau' < 0$. Then we can define $\omega'' = -\omega'$, which achieves $\tau'' = -\tau' > 0$, which is a contradiction. This means that for all $\omega \in \Omega$,

$$\tau \geq |\mathbb{E}[\omega(\mathbf{x}_i) \mid v_i = 0] - \mathbb{E}[\omega(\mathbf{x}_i) \mid v_i = 1]|$$

Taking $\omega(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$,

$$\tau \geq \left|\mathbb{E}[\mathbf{w}^\top \mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{w}^\top \mathbf{x}_i \mid v_i = 1]\right| = \left|\mathbf{w}^\top \Delta\right|.$$

Note that $\|\mathbf{w}_\perp\| = \frac{|\mathbf{w}^\top \Delta|}{\|\Delta\|}$, which completes our proof. □

Proposition A2. *(Restated proposition 3.4.2) For $\mathcal{D} \sim P^\circ$, and for any for any $\mathcal{F}_{L_2}$ such that $f_{rinv} \in \mathcal{F}_{L_2}$, there exists a $\mathcal{F}_{\text{MMD},L_2} \subseteq \mathcal{F}_{L_2}$ such that $f_{rinv} \in \mathcal{F}_{\text{MMD},L_2}$. And the smallest $\mathcal{F}_{\text{MMD},L_2}$ such that $f_{rinv} \in \mathcal{F}_{\text{MMD},L_2}$ has* $\text{MMD} = 0$.

Proof. We prove the existence of a subset, $\mathcal{F}_{\text{MMD},L_2} \subset \mathcal{F}_{L_2}$ by giving an example of

such a subset. Consider

$$\mathcal{F}_{L_2,\text{MMD}} = \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A,\ \text{MMD}(P^\circ_{\phi_0}, P^\circ_{\phi_1}) = 0\},$$

Clearly, $\mathcal{F}_{\text{MMD},L_2} \subset \mathcal{F}_{L_2}$. We will now show that any $f_{\text{rinv}} \in \mathcal{F}_{L_2}$ is also $\in \mathcal{F}_{\text{MMD},L_2}$.

By the definition of f_{rinv}, any $f_{\text{rinv}} \in \mathcal{F}_{L_2}$ must satisfy $f_{\text{rinv}}(\mathbf{x}) \perp\!\!\!\perp v$. Then $T_1(f_{\text{rinv}}(\mathbf{x})) \perp\!\!\!\perp T_2(v)$ for any transformations T_1, T_2. Taking T_1 to be the inverse of the sigmoid function, σ^{-1}, and T_2 to be the identity transformation, we get that $\sigma^{-1}(f_{\text{rinv}}(\mathbf{x})) = \sigma^{-1}(\sigma(\mathbf{w}^\top \mathbf{x})) = \mathbf{w}^\top \mathbf{x} \perp\!\!\!\perp v$. This implies that $p(\mathbf{w}^\top \mathbf{x}|v = 0) = p(\mathbf{w}^\top \mathbf{x}|v = 1)$, which in turn implies that $\text{MMD}(P^\circ_{\phi_0}, P^\circ_{\phi_1}) = 0$, where $\phi(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. □

Proposition A3. *(Restated Proposition 3.4.3). Let* $\mathbf{x}_\perp := \Pi\mathbf{x}$, $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$. *For training data* $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, $\sup_{\mathbf{x}_\perp} \|\mathbf{x}_\perp\|_2 \leq B_\perp$, $\sup_{\mathbf{x}_\parallel} \|\mathbf{x}_\parallel\|_2 \leq B_\parallel$,

$$\mathfrak{R}(\mathcal{F}_{L_2}) \leq \frac{A\sqrt{B_\parallel^2 + B_\perp^2}}{\sqrt{n}},$$

and

$$\mathfrak{R}(\mathcal{F}_{\text{MMD},L_2}) \leq \frac{A \cdot B_\parallel + \tau \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}}.$$

Proof. First, we derive the bound on $\mathfrak{R}(\mathcal{F}_{L_2})$

$$\begin{aligned}\mathfrak{R}(\mathcal{F}) &= \mathbb{E}_\mathcal{D}\mathbb{E}_\epsilon\left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq A} \frac{1}{n}\sum_i \epsilon_i \mathbf{w}^\top \mathbf{x}_i\right] \\ &= \mathbb{E}_\mathcal{D}\mathbb{E}_\epsilon\left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2 \leq A} \frac{1}{n}\sum_i \epsilon_i \mathbf{w}^\top (\mathbf{x}_{\perp i} + \mathbf{x}_{\parallel i})\right]\end{aligned}$$

Following the usual derivations (e.g., see Mohri et al. (2018)), we get the desired result for $\mathfrak{R}(\mathcal{F}_{L_2})$. Next, we derive the bound on $\mathfrak{R}(\mathcal{F}_{\text{MMD},L_2})$.

$$
\begin{aligned}
\mathfrak{R}(\mathcal{F}) &= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \mathbf{w}^\top \mathbf{x}_i\right] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \big(\Pi\mathbf{w}^\top \mathbf{x}_i + (1-\Pi)\mathbf{w}^\top \mathbf{x}_i\big)\right] \\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}\left[\sup_{\substack{\mathbf{w}_\parallel:\|\mathbf{w}_\parallel\|_2\leq A \\ \mathbf{w}_\perp:\|\mathbf{w}_\perp\|_2\leq A}} \frac{1}{n}\sum_i \epsilon_i \mathbf{w}_\perp^\top \mathbf{x}_{\perp i} + \epsilon_i \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i}\right] \\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}_\perp:\|\mathbf{w}_\perp\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \mathbf{w}_\perp^\top \mathbf{x}_{\perp i}\right] + \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}_\parallel:\|\mathbf{w}_\parallel\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i}\right],
\end{aligned}
$$

where the last inequality follows by the subadditivity of the supremum. Again, following the usual derivations (e.g., see Mohri et al. (2018)), we get the required result for $\mathfrak{R}(\mathcal{F}_{\mathrm{MMD},L_2})$ □

Proposition A4. *(Restated Proposition 3.4.4) Let* $\mathcal{F}'_{L_2,\mathrm{MMD}} := \{f : \mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x}), \|\mathbf{w}\|_2 \leq A,\ \mathrm{MMD}(P_{\phi_0}, P_{\phi_1}) \leq \tau'\}$ *be the smallest function class that contains* f_{rinv}*. Then* $\tau' = c \cdot A$ *for some* $c > 0$*, and the corresponding generalization error on* P° *is*

$$
R^\circ(f) \leq \hat{R}_P^{\mathbf{u}}(f) + L \cdot \frac{A \cdot B_\parallel + c \cdot A\frac{B_\perp}{\|\Delta\|}}{\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}},
$$

Proof. By proposition 3.4.2, we have that the smallest MMD regularized function class that contains f_{rinv} when $\mathcal{D} \sim P^\circ$ has MMD = 0. And by proposition 3.4.1 we have in that function class $\|\mathbf{w}_\perp\| = 0$, i.e., $\mathbf{w}_\perp$ is the 0 vector.

$$
\begin{aligned}
\tau' &\geq \left\|\mathbb{E}[\mathbf{w}^\top \mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{w}^\top \mathbf{x}_i \mid v_i = 1]\right\| \\
&= \left\|\mathbb{E}[\mathbf{w}_\perp^\top \mathbf{x}_{\perp i} + \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 0] - \mathbb{E}[\mathbf{w}_\perp^\top \mathbf{x}_{\perp i} + \mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 1]\right\| \\
&= \left\|\mathbb{E}[\mathbf{w}_\parallel^\top \mathbf{x}_{\parallel i} \mid v_i = 0] - \mathbb{E}[\mathbf{w}_\parallel^\top \mathbf{x}_{\parallel,i} \mid v_i = 1]\right\| \\
&= \left\|\mathbf{w}_\parallel \big(\mathbb{E}[\mathbf{x}_{\parallel i} \mid v_i = 0] - \mathbb{E}[\mathbf{x}_{\parallel,i} \mid v_i = 1]\big)\right\| \\
&= \left\|\mathbf{w}_\parallel (1 - \Pi) \big(\mathbb{E}[\mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{x}_i \mid v_i = 1]\big)\right\| \\
&= \|\mathbf{w}_\parallel\| \left\|(1 - \Pi) \big(\mathbb{E}[\mathbf{x}_i \mid v_i = 0] - \mathbb{E}[\mathbf{x}_i \mid v_i = 1]\big)\right\| \\
&= A \left\|(1 - \Pi)\Delta_P\right\|,
\end{aligned}
$$

where the fifth equality holds because the two vectors are scalar multiples of the same vector (they are both projections onto the vector orthogonal to Δ) so Cauchy-Schwartz holds with equality. Also note that $\|(1 - \Pi)\Delta_P\| = 0$ if and only if $\Delta_P = \Delta$, i.e., $P = P^\circ$. So $\|(1 - \Pi)\Delta_P\| > 0$.

The generalization error bound follows immediately by plugging in the upper bound on the Rademacher complexity into equation 3.7. □

B.3.1 Proof of Proposition 3.4.5

The proof of Proposition 3.4.5 applies the techniques for estimating the generalization error of reweighted estimators presented in Cortes et al. (2010b). To apply the Cortes results, we need the hypothesis space to be finite, which is not true for $\mathcal{F}_{L_2,\mathrm{MMD}}$. To address that, we construct a discretization or a covering of $\mathcal{F}_{L_2,\mathrm{MMD}}$, defined next.

Definition A1. *Given any function class $\mathcal{F}$, a metric D on the elements of $\mathcal{F}$, and $\varepsilon > 0$, we define a covering number $\mathcal{N}(\mathcal{F}, D, \varepsilon)$ as the minimal number m of functions*

$f_1, f_2, \ldots, f_m \in \mathcal{F}$, *such that for all* $f \in \mathcal{F}$, $\min_{i=1,\ldots,m} D(f_i, f) \leq \varepsilon$, *with*

$$D(f, f') = \sqrt{\frac{1}{n}\sum_i (f(\mathbf{x}_i) - f'(\mathbf{x}_i))^2}.$$

Our statement also makes use of Gaussian complexities, defined next.

Definition A2. *For a function family* $\mathcal{F}$*, the empirical Gaussian complexity is defined as:*

$$\mathfrak{G}(\mathcal{F}) = \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\eta}\left[\sup_{f\in\mathcal{F}} \eta_i f(\mathbf{x}_i)\right]$$

We are now ready to present the metric entropy of the discretized hypothesis space in this next lemma.

Lemma A1. *Let* $\mathbf{x}_\perp := \Pi\mathbf{x}$, $\mathbf{x}_\parallel := (I - \Pi)\mathbf{x}$, $\sup_{\mathbf{x}_\perp} \|\mathbf{x}_\perp\|_2 \leq B_\perp$, $\sup_{\mathbf{x}_\parallel} \|\mathbf{x}_\parallel\|_2 \leq B_\parallel$, D, ε *as is defined in A1. For* $\varepsilon, c', c'' > 0$*:*

$$\begin{aligned}
&\log(\mathcal{N}(\mathcal{F}_{L_2,\mathrm{MMD}}, D, \varepsilon)) \\
&\leq c'' \left(\frac{c'\sqrt{\log(n)} \cdot \left(A \cdot B_\parallel + \tau \frac{B_\perp}{\|\Delta\|} \right)}{\varepsilon} \right)^2
\end{aligned}$$

Proof. We construct our argument relying on Sudakov's minoration, and the bound between Gaussian and Rademacher complexities. Specifically, by Ledoux (1996), for some $c' > 0$:

$$\begin{aligned}
\mathfrak{G}_m(\mathcal{F}_{L_2,\mathrm{MMD}}) &\leq c'\sqrt{\log(n)} \cdot \mathfrak{R}(\mathcal{F}_{L_2,\mathrm{MMD}}) \\
&\leq c'\sqrt{\log(n)} \cdot \frac{A \cdot B_\parallel + \tau \frac{B_\perp}{\|\Delta\|}}{\sqrt{n}},
\end{aligned}$$

where the last inequality follows from plugging in the results from proposition 3.4.3.

By Sudakov's minoration (see Ledoux (1996) theorem 3.18), for some universal constant $c'' > 0$,

$$\log(\mathcal{N}(\mathcal{F}_{L_2,\text{MMD}}, D, \varepsilon)) \leq c''\left(\frac{\sqrt{n} \cdot \mathfrak{G}_m(\mathcal{F}_{L_2,\text{MMD}})}{\varepsilon}\right)^2$$
$$\leq c''\left(\frac{c'\sqrt{\log(n)} \cdot \left(A \cdot B_\| + \tau\frac{B_\perp}{\|\Delta\|}\right)}{\varepsilon}\right)^2$$

□

The final generalization error of the reweighted estimator is next.

Proposition A5. *(Restated proposition 3.4.5) For $\mathcal{D} \sim P$, with $P \in \mathcal{P}$, and $\mathbf{u}$ as defined in 3.2, C_P as defined in 3.9,*

$$R^\circ(f) \leq \hat{R}_P^{\mathbf{u}}(f) + \frac{2C_P(\kappa(\mathcal{F}_{\text{MMD},L_2}) + \log\frac{1}{\delta})}{2n} + \sqrt{\frac{\Lambda(P^\circ||P) \cdot (\kappa(\mathcal{F}) + \log\frac{1}{\delta})}{n}},$$

where

$$\kappa(\mathcal{F}_{\text{MMD},L_2}) = c''\left(\frac{c'\sqrt{\log(n)} \cdot \left(A \cdot B_\| + \tau\frac{B_\perp}{\|\Delta\|}\right)}{\varepsilon}\right)^2$$

Proof. Using the bound on the metric entropy derived in lemma A1, the proof becomes a direct application of Theorem 2 in Cortes et al. (2010b) □

This concludes the proof for proposition 3.4.5.

B.3.2 Proof for proposition 3.4.6

Lemma A2. *For training data $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$, $\mathcal{D} \sim P^\circ$, and a corresponding learned $f = h(\phi(\mathbf{x}))$ with expected risk $R^\circ(f)$, suppose that y is $\phi-$representable, i.e.,*

that there exists $g(\phi(\mathbf{x})) = y$*, and that* $g(\phi)\ell(\phi) \in \Omega$*. Then for all* y*:*

$$P(Y = y)[R^\circ_{0y} - R^\circ_{1y}] \leq \tau,$$

where $R^\circ_{vy} := \mathbb{E}_{\mathbf{x}\sim P^\circ}[\ell(f(\mathbf{x}), y)|V = v, Y = y]$

Proof. Without loss of generality, suppose that for $y = 1$,

$$P(Y = y)[R^\circ_{0y} - R^\circ_{1y}] = \tau_1 > \tau.$$

Then due to the fact that MMD $\leq \tau$, and by assumption that $\ell \in \Omega$

$$\begin{aligned} P(Y = 0)[R^\circ_{00} - R^\circ_{10}] &\leq \tau - \tau_1 \\ P(Y = 0)[R^\circ_{00} - R^\circ_{10}] &< 0 \\ R^\circ_{00} - R^\circ_{10} &< 0. \end{aligned}$$

Using the shorthand $R^\circ_{\Delta 0} := R_{00} - R_{10}$, the above inequality implies that $-R^\circ_{\Delta 0} > 0$.

Let $\dot{\ell}(\phi) = (2g(\phi) - 1) \cdot \ell(\phi)$, and $\dot{R}^\circ := \mathbb{E}[\dot{\ell}(\phi)]$. By assumption, we have that $\dot{\ell}$ is also $\in \Omega$. However,

$$\begin{aligned} \text{MMD}(\dot{\ell}, P_{\phi_0}, P_{\phi_1}) &= P(Y = 0)[\dot{R}^\circ_{00} - \dot{R}^\circ_{10}] + P(Y = 1)[\dot{R}^\circ_{01} - \dot{R}^\circ_{11}] \\ &= P(Y = 0)[-(R^\circ_{00} - R^\circ_{10})] + P(Y = 1)[R^\circ_{01} - R^\circ_{11}] \\ &= P(Y = 0)[-R^\circ_{\Delta 0}] + \tau_1 > \tau. \end{aligned}$$

This contradicts the MMD condition; that for all functions in Ω, MMD $\leq \tau$

□

Proposition A6 (Restated proposition 3.4.6)**.** *For training data* $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$*,* $\mathcal{D} \sim P^\circ$*, and a corresponding learned* $f = h(\phi(\mathbf{x}))$ *with expected risk* R°*, suppose that* y *is* $\phi-$*representable, i.e., that there exists* $g(\phi(\mathbf{x})) = y$*, and that* $g(\phi)\ell(\phi) \in \Omega$*. For*

some β that depends on P, such that $-2 < \beta < 2$, and $\beta = 0$ if $P = P^\circ$,

$$R_P \leq R^\circ + \beta \cdot \tau$$

Proof. We will use $P_{v|y}(v) := P(V = v|Y = v)$, $P_y(y) = P(Y = y)$, and $P_v(v) = P(V = v)$.

Note that

$$\begin{aligned} R_\circ &= \sum_y P_y^\circ(y)\big[P_{v|y}^\circ(0)R_{0,y}^\circ + P_{v|y}^\circ(1)R_{1,y}^\circ\big] \\ &= \sum_y P_y(y)\big[P_v^\circ(0)R_{0,y}^\circ + P_v^\circ(1)R_{1,y}^\circ\big]. \end{aligned}$$

And:

$$R_P = \sum_y P_y(y)\big[P_{v|y}(0)R_{0,y}^\circ + P_{v|y}(1)R_{1,y}^\circ\big].$$

Taking the difference between the two:

$$\begin{aligned} R_P - R^\circ &= \sum_y P_y(y)\big[\big(P_{v|y}(0) - P_v^\circ(0)\big)R_{0,y}^\circ + \big(P_{v|y}(1) - P_v^\circ(1)\big)R_{1,y}^\circ\big] \\ &= \sum_y P_y(y)\big[\big(P_{v|y}(0) - P_v^\circ(0)\big)R_{0,y}^\circ + \big(\big(1 - P_{v|y}(0)\big) - \big(1 - P_v^\circ(0)\big)\big)R_{1,y}^\circ\big] \\ &= \sum_y P_y(y)\big[\big(P_{v|y}(0) - P_v^\circ(0)\big)R_{0,y}^\circ - \big(P_{v|y}(0) - P_v^\circ(0)\big)R_{1,y}^\circ\big] \\ &= \sum_y P_y(y)\big[\beta_y R_{0,y}^\circ - \beta_y R_{1,y}^\circ\big] \\ &= \sum_y \beta_y P_y(y)\big[R_{0,y}^\circ - R_{1,y}^\circ\big] \\ &\leq \sum_y \beta_y \tau \\ &= \beta \cdot \tau \end{aligned}$$

where in the fourth equality, we use the shorthand $\beta_y := P_{v|y}(0) - P_v^\circ(0)$, and $-1 <$

$\beta_y < 1$. The first inequality follows from lemma A2, and the last equality follows from setting $\beta = \sum_y \beta_y$. □

B.4 Additional experiments

Here we present the results using the full (noisy) background images. Results from the main analysis largely hold, with the exception of the results from the training setting where the data are sampled from the ideal distribution P°. Because the backgrounds are noisy, we see an overall higher variance in performance, so the models perform equally well with no clear "winner".

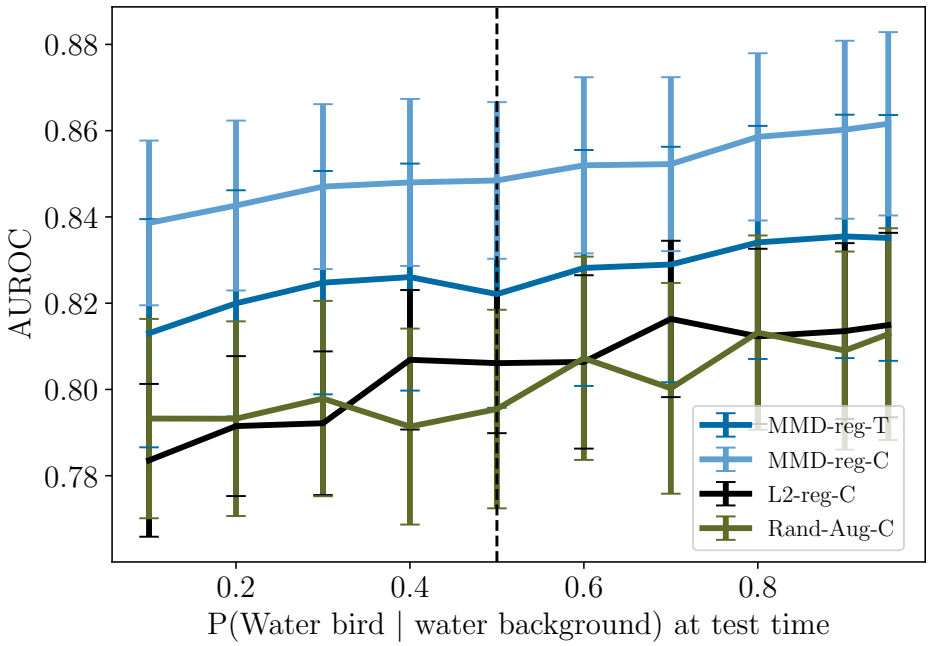

Figure B-1: Training data sampled from P°, with $P^\circ(Y|V=1) = P^\circ(Y|V=0) = 0.5$ and backgrounds are sampled from a noisy set of images. x-axis shows $P(Y|V)$ at test time under different shifted distributions. y-axis shows AUROC on test data. Vertical dashed line shows training data. MMD-regularized models outperform baselines within, and outside the training distribution.

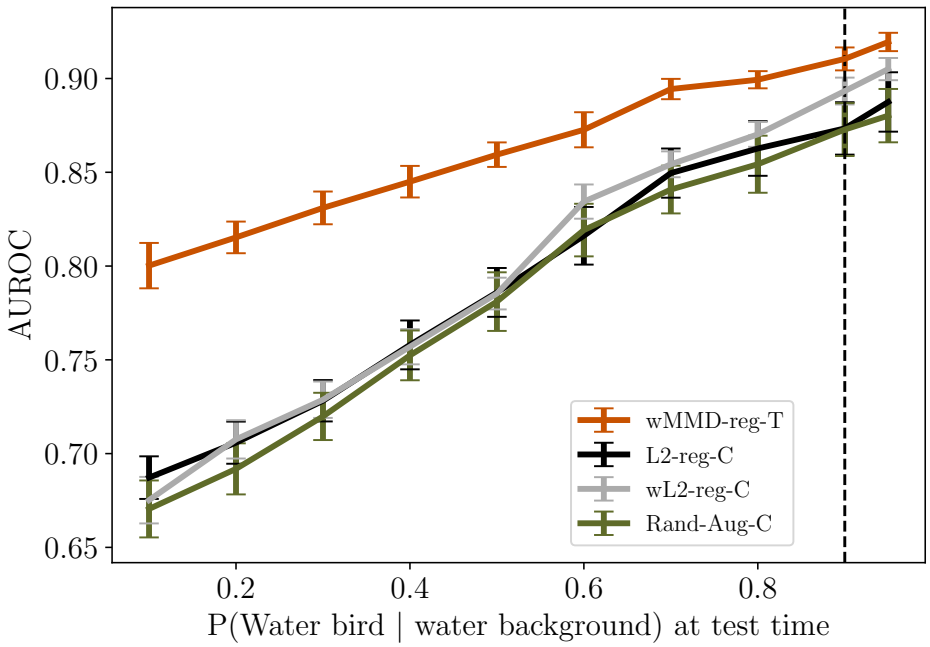

Figure B-2: Training data sampled from P, with $P(Y = 1|V = 1) = P^{\circ}(Y = 0|V = 0) = 0.9$, and backgrounds are sampled from a noisy set of images. Vertical dashed line shows training data. x, y axes similar to figure B-1. MMD-regularized models outperform baselines showing better robustness against distribution shifts at test time.

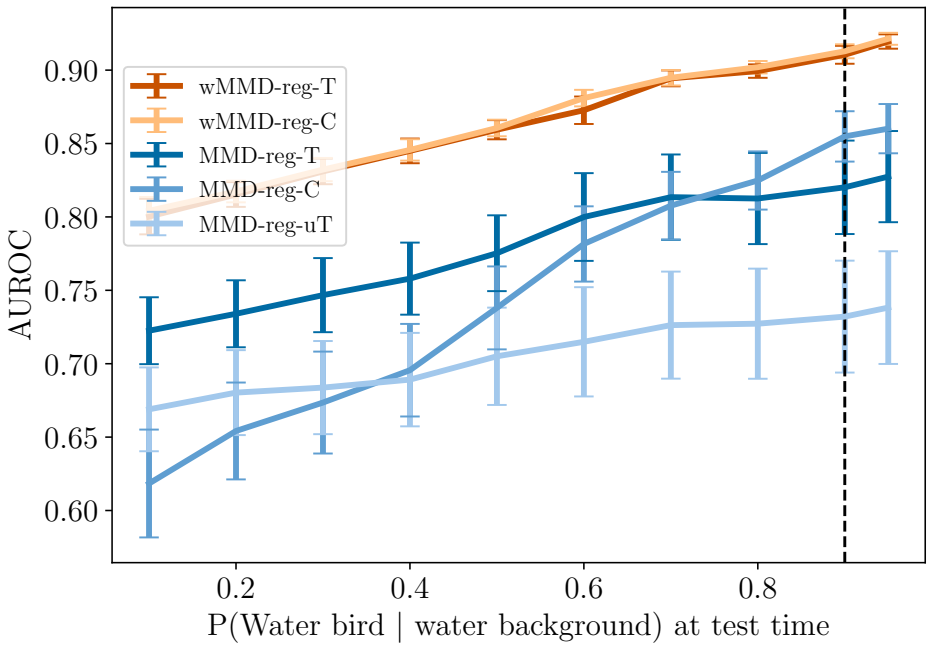

Figure B-3: Training data sampled from P, with $P(Y = 1|V = 1) = P^{\circ}(Y = 0|V = 0) = 0.9$. x, and backgrounds are sampled from a noisy set of images. y axes similar to fig B-1. An ablation study to show how different components of our suggested approach (wMMD-reg-T) contribute to improved performance.

Appendix C

Appendix to chapter 4

C.1 Architecture and cross-validation

The core network architecture is kept constant for all models and experiments outlined in the main text. All models have a core recurrent neural network (RNN). The RNN takes in the individual characteristics, the infection state from the previous time point, and an estimate of the exposure (except for the No exposure model, NEM, which ignores exposure). For our model (MIINT), the exposure estimate is based on the imputed values according to Q, whereas for the optimistic model (OM) it is the sum of observed neighbor infection states from the previous time point (assuming untested is uninfected). For the Oracle model (ORM), the estimate of exposure is the sum of true neighbor infection states from the previous time point.

The RNN inputs are passed through one fully dense layer with a tanh activation, giving an intermediate layer of dimension = 64 units. The output is then passed to another 64-unit dense layer which is finally passed through a sigmoid to give the final probability of infection.

For the simulation experiment, we found that different values of τ did not affect the

prediction much. So we set $\tau = .5$, but pick λ based on cross-validation using a grid of values = Cross-validation is done to pick the value of λ from the candidate values $[0, 1e^{-1}, 1e^{-2}, 1, 1e^{1}, 1e^{2}, 1e^{3}]$. This is done via 2-fold cross validation. We pick the final value to be the one that maximizes the AUROC defined with respect to the observed labels in a held out validation set.

Cross-validation for the real data experiment is similar, though her we found that values of τ are important, so in addition to λ, we also pick the value of τ from the candidate values $[0.001, 0.01, 0.1, 0.5]$.

C.2 Real data

Inclusion Criteria. Similar to Oh et al. (2018); Makar et al. (2018), we exclude all hospitalizations of patients younger than 18. We do so because predicting pediatric CDI is a significantly different task from that of the adult population. We also exclude patients with suspected community acquired infections, since predicting nosocomial infections (i.e., hospital associated infections) is a significantly different task that that of community-acquired infections. Again, we follow Oh et al. (2018); Makar et al. (2018) in defining community acquired infections as those who get the CDI diagnosis in the first 2 days of their visit, and those who have had a CDI infection in the 14 days prior to their hospitalization.

Patient Features. Similar to Oh et al. (2018); Makar et al. (2018), we include patient demographics, which are available upon admission such as age, gender, number and length of previous hospitalizations, reason and source of visit (e.g., transferred from a Skilled Nursing Facility or admitted through the emergency room). We capture medical history by including all ICD-9 procedure and diagnosis codes from prior visits that happened at most 90 days prior to the main (index) visit. We collect data from the index visit up to one day before the prediction date. This includes medications, lab tests ordered and their results.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 129–138. PMLR, 10–15 Jul 2018.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems 30*, pages 3424–3432. Curran Associates, Inc., 2017.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439): 1171–1176, 1997.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel methods—support vector learning*, pages 43–54, 1999.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Zhihong Cai, Manabu Kuroki, Judea Pearl, and Jin Tian. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701, 2008.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Victor Chernozhukov and Christian Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney K Newey. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems 23*, pages 442–450. Curran Associates, Inc., 2010a.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010b.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pages 2803–2813. PMLR, 2020.

Kai Fan, Chunyuan Li, and Katherine Heller. A unifying variational inference framework for hierarchical graph-coupled hmm with an application to influenza infection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3828–3834, 2016.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, pages 7538–7550, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

International Stroke Trial Collaborative Group. The international stroke trial (ist): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349(9065):1569–1581, 1997.

Joseph D Janizek, Gabriel Erion, Alex J DeGrave, and Su-In Lee. An adversarial approach for the robust classification of pneumonia from chest radiographs. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 69–79, 2020.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. *To appear in AISTATS*, 2019.

Adam Kapelner and Justin Bleich. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016. doi: 10.18637/jss.v070.i04.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Michel Ledoux. Isoperimetry and gaussian analysis. In *Lectures on probability theory and statistics*, pages 165–294. Springer, 1996.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2003.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124, 2019.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Shelley S Magill, Jonathan R Edwards, Wendy Bamberg, Zintars G Beldavs, Ghinwa Dumyati, Marion A Kainer, Ruth Lynfield, Meghan Maloney, Laura McAllister-Hollod, Joelle Nadle, et al. Multistate point-prevalence survey of health care–associated infections. *New England Journal of Medicine*, 370(13):1198–1208, 2014.

Maggie Makar, John Guttag, and Jenna Wiens. Learning the probability of activation in the presence of latent spreaders. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In *International Conference on Machine Learning*, pages 6661–6671. PMLR, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects, 2017.

Elizabeth L Ogburn, Oleg Sofrygin, MJ van der Laan, and I Diaz. Causal inference for social network data with contagion. *ArXiv e-prints*, 2017.

Jeeheh Oh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin E Ryan, Laraine Washer, Lauren R West, Vincent B Young, John Guttag, et al. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *infection control & hospital epidemiology*, 39(4):425–433, 2018.

J Origüen, L Corbella, MA Orellana, M Fernandez-Ruiz, F Lopez-Medrano, R San Juan, M Lizasoain, T Ruiz-Merlo, A Morales-Cartagena, G Maestro, et al. Comparison of the clinical course of clostridium difficile infection in glutamate dehydrogenase-positive toxin-negative patients diagnosed by pcr to those with a positive toxin test. *Clinical Microbiology and Infection*, 24(4):414–421, 2018.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, USA, 2000. ISBN 0521773628.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Christopher R Polage, Clare E Gyorke, Michael A Kennedy, Jhansi L Leslie, David L Chin, Susan Wang, Hien H Nguyen, Bin Huang, Yi-Wei Tang, Lenora W Lee, et al. Overdiagnosis of clostridium difficile infection in the molecular test era. *JAMA internal medicine*, 175(11):1792–1801, 2015.

Michelle M Riggs, Ajay K Sethi, Trina F Zabarsky, Elizabeth C Eckstein, Robin LP Jump, and Curtis J Donskey. Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic clostridium difficile strains among long-term care facility residents. *Clinical infectious diseases*, 45(8):992–998, 2007.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392, 2007.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

JM Robins. 1997 proceedings of the section on bayesian statistical science. 1998.

Joshua Robinson, Stefanie Jegelka, and Suvrit Sra. Strength from weakness: Fast learning using weak supervision. *arXiv preprint arXiv:2002.08483*, 2020.

Paul R Rosenbaum. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Donald B. Rubin. Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with "Censoring" Due to Death. *Statistical Science*, 21(3):299 – 309, 2006. doi: 10.1214/088342306000000114. URL `https://doi.org/10.1214/088342306000000114`.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020a. URL `http://proceedings.mlr.press/v119/sagawa20a.html`.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *arXiv preprint arXiv:2005.04345*, 2020b.

Robert S Sargent, Keebom Kang, and David Goldsman. An investigation of finite-sample behavior of confidence interval estimators. *Operations Research*, 40(5): 898–913, 1992.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, July 2001. ISSN 0899-7667.

Matthias Seeger. Learning with labeled and unlabeled data. Technical report, 2000.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3076–3085, Sydney, Australia, 2017. PMLR.

John Shawe-Taylor and Nello Cristianini. On the generalisation of soft margin algorithms. *IEEE Transactions on Information Theory*, 48(10):2721–2735, 2002.

John Shawe-Taylor and Robert C. Williamson. Generalization performance of classifiers in terms of observed covering numbers. In *Computational Learning Theory*, pages 274–285, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Oleg Sofrygin and Mark J van der Laan. Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. *Journal of causal inference*, 5(1), 2016.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3231–3239. Curran Associates, Inc., 2015a.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015b.

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *Journal of machine learning research*, 7(Jul):1231–1264, 2006.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

Alexander B Tsybakov et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Jiaxuan Wang, Jeeheh Oh, Haozhu Wang, and Jenna Wiens. Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2417–2426, 2018.

Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2021.

Jenna Wiens, Eric Horvitz, and John V Guttag. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Advances in Neural Information Processing Systems*, pages 467–475, 2012.

Wikipedia contributors. Cleromancy — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Cleromancy&oldid=1018059389`, 2021.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *arXiv preprint arXiv:1906.08386*, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.