Structural and Statistical Uncertainty in Observational Causal Machine Learning at Scale

Andrew Jesson

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

In loving memory of Mat Neuman

Acknowledgements

I would like to express my gratitude for the incredible support I have received from my family, Kathy and Don, Brad, Henry, Jen, and Mat.

I cannot overlook the contributions of those who have supported me from the beginning of a pivot that started 13 years ago in Montreal: Wilson, Tara, Phil, Laura, Marc, the Marosys, Laura M., Chris, Meredith, Dalia, Monique, The Michaels, Gab, Alex, Blaž, Matina, Neda, Pi Cafe, the dedicated McGill Gym and Grounds teams, and Imagia Canexia Health.

I am grateful for the supportive community I have found at Oxford. In addition to everyone at OATML and Linacre College, I would particularly like to extend my thanks to Lisa Schut, Milad Alizadeh, Lewis Smith, Joost van Amersfoort, Luisa Zintgraf, Panagiotis Tigas, Mizu Nishikawa-Toomey, Tim Rudner, Jannik Kossen, Muhammed Razzak, Lars Holdijk, Lorenz Kuhn, Kelsey Doerksen, and Eveline Lupien for their friendship and support.

I would also like to extend my appreciation to all collaborators: Sören Mindermann, Panagiotis Tigas, Alyson Douglas, Peter Manshausen, Joost Van Amersfoort, Andreas Kirsch, Lewis Smith, Oscar Key, Sebastian Farquhar, Parmida Atighehchian, Frederic Branchaud-Charron, Duncan Watson-Parris, Arash Mehrjou, Ashkan Soleymani, Pascal Notin, Stefan Bauer, Patrick Schwab, Yashas Annadani, Bernhard Schölkopf, Nicolai Meinshausen, Maëlys Solal, Philip Stier, Desi Ivanova, Adam Foster, Clare Lyle, Miruna Oprescu, Marah Ghoummaid, Jacob Dorn, Nathan Kallus, Myrl Marmarelis, Elizabeth Haddad, Neda Jahanshad, Aram Galstyan, Greg Ver Steeg, Shreshth Malik, Salem Lahlou, Moksh Jain, Nikolay Malkin, Tristan Deleu, Yoshua Bengio, Chris Lu, Angelos Filos, Gunshi Gupta, and Jakob Foerster.

Lastly, I would like to offer special thanks to Uri Shalit and my supervisor, Yarin Gal.

Abstract

Causal machine learning (Causal ML) tackles various tasks, including causal-effect inference, causal reasoning, and causal structure discovery. This thesis explores uncertainty for Causal ML methods that scale to large datasets and complex, high-dimensional input/output modalities, such as images, text, time series, and videos. Scalability is essential for efficiently processing vast amounts of information and predicting complex relationships.

As we scale and achieve greater modeling flexibility, communicating the unknown becomes increasingly important. We examine two primary types of uncertainty: statistical and structural. Statistical uncertainty arises when fitting machine learning models to finite datasets. Addressing this uncertainty allows predicting a range of plausible causal effects that shrink with more training examples, facilitating better-informed decision-making and indicating areas needing improved understanding. Structural uncertainty arises from imprecise knowledge of the causal structure and generally requires further assumptions about the data-generating process or interaction with the world.

In this thesis, we develop scalable Causal ML methods that navigate statistical and structural uncertainty effectively. We demonstrate the importance of considering scalability and uncertainty in Causal ML algorithm design and application, enhancing decision making and knowledge acquisition. Our contributions aim to advance the Causal machine learning field and provide a foundation for future research.

Contents

1 Why do Scalability and Uncertainty Matter in Causal Machine Learning? 1

1.1 On Structural and Statistical Uncertainty . . . 2
1.2 Objective and Aims . . . 4
1.3 Thesis Structure . . . 4

2 Causal-effect Inference from Observational Data 7

2.1 Observational Data . . . 7
2.2 Causal Estimands . . . 8
2.3 Identifiability Conditions . . . 9
2.4 Statistical Causal-Effect Estimands . . . 9
2.5 Machine Learning for Causal-Effect Estimation . . . 10
2.6 Uncertainty for Causal-Effect Inference . . . 11
2.6.1 Statistical Uncertainty . . . 11
2.6.1.1 Challenges . . . 15
2.6.1.2 Uncertainty Aware Machine Learning Methods for Causal-Effect Estimation . . . 15
2.6.2 Structural Uncertainty . . . 16
2.6.2.1 Causal Sensitivity Analysis Methods . . . 16

3 Methodological Background 18

3.1 Scalable Causal-Effect Inference . . . 18
3.1.1 Scalable Measures of Statistical Uncertainty . . . 21
3.2 Scalable Statistical Uncertainty Quantification . . . 22
3.2.1 Scalable Uncertainty-Aware Machine Learning . . . 23
3.2.1.1 Bayesian Neural Networks . . . 23
3.2.1.2 Deep Ensembles . . . 23
3.2.1.3 Deep Kernel Learning . . . 23

3.2.1.4 Other Methods 24
3.3 Sensitivity Analysis for Structural Uncertainty 24
3.3.1 Marginal Sensitivity Model for Binary-Valued Interventions . . 24
3.3.2 CAPO and CATE Bounds Under the MSM 26
3.4 Active Learning and Experimental Design 28

4 Scalable Statistical Uncertainty for Causal Machine Learning 30
4.1 Statistical Uncertainty Estimands and Estimators 32
4.1.1 Statistical Uncertainty About the Conditional Average Treatment Effect (CATE) 32
4.1.2 CATE Statistical Uncertainty Estimators 34
4.1.2.1 Estimating Statistical CATE Uncertainty Using Bayesian Linear Regression, Neural Linear Models, Gaussian Processes, and Deep Kernel GPs 34
4.1.2.2 Estimating Statistical CATE Uncertainty Using Bayesian Neural Networks and Ensemble Methods. 36
4.1.3 Statistical Uncertainty About Conditional Average Potential Outcomes (CAPO) 37
4.1.4 CAPO Statistical Uncertainty Estimators 38
4.1.4.1 Estimating Statistical CAPO Uncertainty Using Bayesian Linear Regression, Neural Linear Models, Gaussian Processes, and Deep Kernel GPs 38
4.1.4.2 Estimating Statistical CAPO Uncertainty Using Bayesian Neural Network and Ensemble Methods. 39
4.2 Active Learning for Conditional Causal-Effects 39
4.2.1 τ-Acquisition Functions 40
4.2.2 μ-Acquisition Functions 41
4.2.3 ρ-Acquisition Functions 41
4.2.4 $\mu\rho$-Acquisition Functions 42
4.3 Applications 44
4.3.1 Deferral of Recommendations to An Expert 44
4.3.1.1 Experiments 45
4.3.1.2 Using Uncertainty When Overlap is Violated 46
4.3.1.3 Uncertainty Under Covariate Shift 48
4.3.2 Causal Active Learning of Conditional Average Treatment Effects 49
4.3.2.1 Experiments 52

5 Scalable Structural Uncertainty for Causal Machine Learning 56
5.1 Marginal Sensitivity Models . . . 58
5.1.1 Discrete-Valued Interventions . . . 59
5.1.2 Continuous-Valued Interventions . . . 60
5.2 Causal-Effect Bounds Without Ignorability . . . 62
5.3 CAPO and CATE Interval Estimators . . . 65
5.3.1 Solving For w . . . 67
5.4 Finite-Sample Causal-Effect Bounds Without Ignorability . . . 68
5.5 Scalable Treatment Effect Estimation Under Structural Uncertainty . 69
5.6 Interpreting Λ . . . 71
5.7 Applications . . . 73
5.7.1 Scalable Sensitivity Analysis for Conditional Average Treatment Effects . . . 73
5.7.1.1 Experiments . . . 73
5.7.2 Scalable Sensitivity Analysis for CAPOs. . . . 78
5.7.2.1 Experiments . . . 78
5.7.3 Estimating Aerosol-Cloud-Climate Effects from Satellite Data 80

6 Conclusions 84

A Publications and Pre-prints 86
A.1 Published Works . . . 86

B Datasets 88
B.1 Simulated Datasets . . . 88
B.1.1 Binary Treatment, Continuous Outcome . . . 88
B.1.2 Continuous Treatment, Continuous Outcome . . . 89
B.2 IHDP . . . 89
B.2.1 IHDP Covariate Shift. . . . 91
B.2.2 IHDP Hidden Confounding . . . 92
B.3 ACIC 2016 . . . 93
B.4 MNIST . . . 93
B.4.1 CEMNIST Overlap . . . 93
B.4.2 HC-MNIST . . . 94
B.4.3 CMNIST . . . 95
B.5 Satellite Climate Observations . . . 96

C Implementation Details **97**
C.1 Deferral of Recommendations to An Expert 97
C.1.1 Models 98

D Theory **100**
D.1 Mathematical Background 100
D.2 Marginal Sensitivity Models 102
D.2.1 CATE Interval Estimator 104
D.2.2 CMSM is OK 110
D.2.3 Alternative Estimators and Optimizers 113
D.2.3.1 Approximating integrals using Gauss-Hermite quadrature 113
D.2.3.2 Line search interval optimization 114
D.2.3.3 Gradient Descent Interval Optimization 114

Bibliography **116**

List of Figures

2.1 Graphical representation of the dependencies. 12
2.2 Graphical representation of hidden confounding. 16

3.1 Neural network design motif for causal effect estimation 20
3.2 Many different functions (purple lines) can explain finite data (blue dots) equally well. 22
3.3 Varying Λ for Marginal Sensitivity Model. Ground truth $\Lambda^* = 2.7$. . . 28

4.1 The purple shaded areas in the lower panes depict regions of ignorance about the response to intervention for units summarized by the covariate value, $\mathbf{X} = \mathbf{x}$. The training data density for the untreated and treated groups are shown in the upper panes. (Figure 5.1a) For ignorance due to measurements $\mathbf{x}$ without **representation** in the observed data, the region should get wider as the distance between $\mathbf{x}$ and the training data increases. (Figure 5.1b) For ignorance without **positivity**, the region should get wider as $P(\mathrm{T} = 0 \mid \mathbf{x})$ or $P(\mathrm{T} = 1 \mid \mathbf{x}) \to 1$. 30
4.2 **CEMNIST** evaluation. (a) Histogram of estimated propensity scores. Untreated nines account for the peaks on the left side. (b) Error rate for different rejection policies as we vary the rejection rate. (c) $\sqrt{\epsilon_{PEHE}}$ for different models at a fixed rejection rate $r_{\text{rej}} = 0.5$. Compared are the policies *random*, *propensity trimming*, and *epistemic uncertainty*. 47
4.3 Uncertainty based rejection policies yield significantly lower error rates while withholding fewer recommendations than propensity policies, on IHDP, IHDP Cov., and ACIC 2016. 48

4.4 Left: Observational data (Figure 4.4a). Top: data density of treatment (right) and control (left) groups. Middle: observed outcome response for treatment (circles) and control (x's) groups. Bottom: data density for active learned training set after a number of acquisition steps. Right: Visualizing CMNIST dataset (Fig. 4.4b). Model inputs are MNIST digits and assigned treatments. The MNIST digits are high-dimensional proxies for the latent confounding covariate ϕ. Digits are projected onto ϕ by ordering them first by image intensity and then by digit class (0 - 9). Methods must be able to implicitly learn this non-linear mapping in order to predict the conditional expected outcomes. 51

4.5 Causal–BALD acquisition functions: How the training set is biased and how this effects the CATE function with a fixed budget of 300 acquired points. 53

4.6 $\sqrt{\epsilon_{PEHE}}$ performance (shaded standard error) for DUE models. **(left to right) synthetic** (40 seeds), and **IHDP** (200 seeds). We observe that BALD objectives outperform the **random**, $\boldsymbol{\gamma}$ and **propensity** acquisition functions significantly, suggesting that epistemic uncertainty aware methods that target reducible uncertainty can be more sample efficient. 55

5.1 The purple shaded areas in the lower panes depict regions of ignorance about the response to intervention for units summarized by the covariate value, $\mathbf{X} = \mathbf{x}$. The training data density for the untreated and treated groups are shown in the upper panes. (Figure 5.1a) For ignorance due to measurements $\mathbf{x}$ without **representation** in the observed data, the region should get wider as the distance between $\mathbf{x}$ and the training data increases. (Figure 5.1b) For ignorance without **positivity**, the region should get wider as $P(\mathrm{T} = 0 \mid \mathbf{x})$ or $P(\mathrm{T} = 1 \mid \mathbf{x}) \to 1$. (Figure 5.1c) For ignorance without **ignorability**, the CATE estimator, $\delta(\mathbf{x}, \boldsymbol{\theta})$, can be arbitrarily biased, hence the discrepancy between the blue solid line and the black dashed line. Therefore, the ignorance region should include the true CATE $\delta(\mathbf{x})$ on the training data manifold where overlap is satisfied. (Figure 5.1d) All sources of ignorance jointly. 56

5.2 Graphical representation of hidden confounding. The causal parent, U, of T and Y exists, but is either not observable or unknown. 57

5.3 Neural network design motif for causal effect estimation 70
5.4 Interpreting Λ as a proportion (ρ) of the unexplained range of Y_t attributed to unobserved confounding variables. Here, AOD is the intervention variable, which will be defined in Section 5.7.2.1 72
5.5 Varying Γ for Marginal Sensitivity Model. Ground truth $\Gamma^* = 2.7$. While the bounds follow the true CATE $\tau(x)$ on the support of $p_{\mathcal{D}}(\mathbf{x})$, they become nonsensical for out-of-distribution data ($\mathbf{x} < -2.5$ and $\mathbf{x} > 2.5$) and when there is a lack of overlap. 74
5.6 **IHDP Hidden Confounding:** Error rate as we sweep over the percentage of deferred points. We propose that recommendations should be deferred when there is ignorance. On the x-axis we vary the share of recommendations deferred, simulating various levels of practitioner caution. *Ignorance* (ours) accounts for all lack of knowledge. *Uncertainty* [Jes+20] accounts only for insufficient similarity and overlap. *Sensitivity* only accounts for hidden confounding, without accounting for insufficient similarity and overlap. *Sensitivity Kernel* is the kernel method of Kallus, Mao, and Zhou [KMZ19], which does not account for other sources of ignorance. Results show that all sources of ignorance are important on IHDP with one hidden confounder. 77
5.7 Figures 5.7a to 5.7d: Synthetic data and ground truth functions. Figures 5.7e to 5.7h Causal uncertainty under hypothesized Λ values. Solid lines are ground truth; thick solid lines where the true λ^* is within the range of hypothesized Λ, thin solid lines otherwise. The dotted lines are the estimated CAPO. Shaded regions are estimated CMSM intervals. 78
5.8 Statistical and causal uncertainty, α is statistical significance level for the bootstrap. see Figure 5.7 for other details. 80
5.9 Causal diagrams. Figure 5.9a, a simplified causal diagram representing what we are reporting within; aerosol optical depth (AOD, regarded as the treatment T) modulates cloud optical depth (τ, Y), which itself is affected by hidden confounders (U) and the meteorological proxies (X). Figure 5.9b, an expanded causal diagram of ACI. The aerosol (a) and aerosol proxy (AOD), the true confounders (light blue), their proxies (dark blue), and the cloud optical depth (red). 81

5.10 Left: The values of the observed, true τ against the modeled τ. Right: The curve for continuous treatment outcome of our aerosol proxy (AOD) on cloud optical depth (τ). The darkest shaded region ($\Lambda = 1$) represents the uncertainty in the treatment outcome from the ensemble due to finite data. As the strength of confounders increases ($\Lambda > 1.0$), the range of uncertainty in the treatment outcome increases. 83

B.1 Synthetic data with hidden confounding 89
B.2 Workflow of observed clouds from satellite to ingestion by model. . . 96

Chapter 1

Why do Scalability and Uncertainty Matter in Causal Machine Learning?

Causal machine learning (CML) encompasses a variety of tasks, including causal-effect inference, causal reasoning, causal structure discovery, and causal representation learning. CML also enriches data-driven algorithms by offering a principled methodology for incorporating domain knowledge, a rich language for expressing modeling assumptions, and a theory for understanding why machine learning predictions fail. This thesis explores uncertainty in scalable CML methods that accommodate large datasets and handle complex, high-dimensional input and output modalities, such as images, text, time series, and videos. Scalability is crucial in the era of big data and complex real-world problems, as it allows CML algorithms to process and learn from vast amounts of information efficiently while modeling the context needed to predict complex relationships.

As we scale and achieve greater flexibility in modeling our world, communicating the unknown becomes increasingly important. The challenge lies in adapting principled methods for uncertainty analysis to scalable methods. Addressing uncertainty is essential both for making better-informed decisions and for identifying what we need to learn. With this in mind, we examine two primary types of uncertainty: statistical and structural.

Statistical uncertainty, often called epistemic uncertainty, arises when fitting machine learning models to finite datasets. Addressing this uncertainty enables predicting a range of plausible causal effects, the size of which shrinks with the number of training

examples. This range of values facilitates better-informed decision-making and indicates states or individuals for which we need to improve our understanding. However, informative statistical uncertainty presupposes that we start with a correct model of the world. Structural uncertainty, which arises from imprecise knowledge of the underlying causal structure in a problem, becomes relevant at this point. Generally, alleviating structural uncertainty requires further assumptions about the data-generating process or interaction with the world. Nevertheless, CML can communicate uncertainty about causal relationships given additional domain knowledge, thus better informing decision-making.

In this thesis, we develop novel methods and techniques for scalable CML that effectively navigate statistical and structural uncertainty. We demonstrate the importance of considering scalability, and uncertainty in designing and applying CML algorithms, as they enhance model robustness, and generalizability. Our contributions aim to advance the field of CML and provide a solid foundation for future research in this domain.

1.1 On Structural and Statistical Uncertainty

Data driven decision making is based on statistical estimands. When we do not have enough data, the range of possible values that such an estimand can take grows. When we do not have the correct model for the underlying data generating process, the estimand can be arbitrarily biased. In both cases, the quality of our decision making can be hindered, which can result in unforeseen harm. This motivates a need to communicate the uncertainty arising from these two scenarios.

Machine learning discussions on uncertainty typically focus on two types of uncertainty: epistemic and aleatoric [DD09; KG17]. *Epistemic uncertainty*, which we will call **statistical uncertainty**, arises from the fact that we fit estimators using finite data. It is often referred to as reducible uncertainty since observing more data can increase our confidence about the expected value of a given outcome. In contrast, *aleatoric uncertainty*, which is a subset of what we will call **structural uncertainty**, is commonly associated with error or noise in the outcome. This uncertainty may stem from rater disagreement, measurement errors in the outcome variable, or inherent unexplainability of the outcome given the covariates modelled.

A classic example that illustrates the difference between epistemic and aleatoric uncertainty is the simple coin flip. For instance, in an experimental trial, i, we may

flip a coin ten times, observe any combination of heads and tails (eight heads and two tails, or four heads and six tails, or ten tails, etc.), and estimate the probability that the coin shows heads ($\mathbb{P}_i(\mathrm{H}) = 0.8$, or $\mathbb{P}_i(\mathrm{H}) = 0.4$, or $\mathbb{P}_i(\mathrm{H}) = 0.0$, etc.). If we repeat the experiment n times, the average estimated probability of heads across trials, $\mathbb{P}_n(\mathrm{H}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_i(\mathrm{H})$, ought to converge with increasing n to the true expected value, $\mathbb{P}(\mathrm{H}) = 0.5$, but the variance of the estimated per-trial probability, $\mathrm{Var}_n\left(\mathbb{P}_\mathrm{I}(\mathrm{H})\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{P}_i(\mathrm{H}) - \mathbb{P}_n(\mathrm{H})\right)^2$, will be high. Such variance represents epistemic uncertainty, which is reducible when we increase the number of coin flips per trial because, with a larger sample size (e.g. 100 flips instead of 10), the per-trial estimates, $\mathbb{P}_i(\mathrm{H})$, get closer to the true value, $\mathbb{P}(\mathrm{H}) = 0.5$. On the other hand, the variance of the true probability of heads, $\mathrm{Var}\left(\mathbb{P}(\mathrm{H}) = 0.5\right) = 0.25$, represents aleatoric uncertainty, which is termed "irreducible" since it arises from the inherent unpredictability when we only observe the outcome. That is, for an ideal fair coin, the outcome of a given coin flip cannot be predicted with certainty, but we can confidently assume that half of the time it will be heads and half of the time it will be tails.

Structural uncertainty goes beyond the traditional definition of aleatoric uncertainty because it can arise from processes that arbitrarily bias the mean estimate. In our above example, the coin may not in fact be fair, but some slight-of-hand deception could result in our statistical estimates incorrectly concluding that it is fair. If a betting policy were informed by such estimates, the resulting decisions would unfairly benefit the deceiver.

When discussing uncertainty in causal machine learning, assigning specific names to different nodes in a causal path diagram, such as actions, states, and rewards, is helpful. While these names are not mandatory, they enable us to distinguish between variables that we can intervene upon (treatments), variables that we can only observe but not intervene upon (covariates), and variables that we observe to measure the effectiveness of our actions (outcomes). We could also use different names such as actions for treatments, states for covariates, and rewards for outcomes. If we limit our covariates to only those variables that are direct causal parents of both the action/treatment and reward/outcome, we could call that set of covariates 'confounders'. Moreover, note that outcomes referred to as rewards imply positive connotations, but in some cases, outcomes may be associated with negative consequences. Hence, we will use outcomes to remain agnostic. I hope this brief comment helps to ensure that this work reaches a wide audience across various disciplines, including

causal inference, machine learning, algorithmic decision making, and reinforcement learning.

1.2 Objective and Aims

Objective: To understand the limits of estimating causal effects from observational data in various application domains and to develop scalable and robust machine learning methodology to responsibly navigate those limits while maintaining high efficacy.

Aim 1: Understand the state of causal machine learning for effect inference from observational data.

Aim 2: Develop novel methodology to communicate uncertainty when inferring causal effects using scalable machine learning.

Aim 3: Develop novel, scalable active data acquisition strategies to efficiently reduce uncertainty in causal effect inference.

Aim 4: Validate proposed methodologies using existing benchmark datasets and develop new benchmarks tailored to explore sources of uncertainty.

1.3 Thesis Structure

Conditional Causal-Effect Inference from Observational Data.

In Chapter 2, we describe the central problem of conditional causal-effect inference from observational data and present the challenges that motivate our contributions.

Methodological Background.

In Chapter 3, we review the building blocks of scalable statistical and structural uncertainty for conditional causal effect inference. In Section 3.1, we review deep-learning methods for scalable causal-effect inference. In Section 3.2, we review machine learning methods for scalable statistical uncertainty quantification. In Section 3.3, we review causal sensitivity analysis methods to communicate structural uncertainty due to unobserved confounding. Finally, in Section 3.4, we review deep learning methods that leverage statistical uncertainty for active learning and experimental design.

Scalable Statistical Uncertainty for Conditional Causal-Effects.

In Chapter 4, we present our contributions to facilitating scalable approaches to quantify statistical uncertainty in conditional causal-effect inference. In Section 4.1, we present estimands and scalable estimators of statistical uncertainty for conditional causal effect inference. In Section 4.2, we show how the resulting estimators can be used to define acquisition functions for active learning of conditional causal effects. Finally, in Section 4.3 we present experimental results for the applications of automated decision making (Section 4.3.1) and active learning (Section 4.3.2). In both the decision making and active learning settings, we show that our methods provide improved robustness to violations in the "positivity" assumption and scale to complex input modalities like images. Chapter 4 builds off of the following two publications:

1. Andrew Jesson*, Sören Mindermann*, Uri Shalit, and Yarin Gal. "Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models." *NeurIPS*. (2020).

2. Andrew Jesson*, Panagiotis Tigas*, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. "Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data." *NeurIPS*. (2021).

Scalable Structural Uncertainty for Conditional Causal-Effects.

In Chapter 5, we present our contributions to facilitating scalable approaches to quantify structural uncertainty in conditional causal-effect inference. In Section 5.1 we present discrete and continuous generalizations of the marginal sensitivity model (MSM) [Tan06]. In Section 5.2 we derive bounds on conditional causal-effects for binary, discrete, and continuous-intervention marginal sensitivity models. In Section 5.3 we derive estimators for the conditional causal-effect bounds. In Section 5.4 we incorporate statistical uncertainty into our estimators. In Section 5.5 we show how these methods are incorporated into scalable machine learning. In Section 5.6 we provide intuition for how we can interpret our indicator of structural uncertainty. Finally, in Section 5.7 we present experimental results for the applications of decision making and climate science. Chapter 5 builds off of the following two publications:

1. Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. "Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding." *ICML*. (2021).

2. Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. "Scalable Sensitivity and Uncertainty Analysis for Causal-Effect Estimates of Continuous-Valued Interventions." *NeurIPS*. (2022).

Conclusions.

We discuss the relationship of our work to active areas of research and interesting directions for future development in Chapter 6.

Chapter 2

Causal-effect Inference from Observational Data

In this work we use the Neyman-Rubin potential outcomes framework[1] [Ney23; Rub74; Sek08]. We denote the potential outcome (or counterfactual [Pea09]) given an initial state, χ, by the vector-valued random variable $\mathrm{Y_T} = (\mathrm{Y}_1, \mathrm{Y}_2, \dots) \in \mathcal{Y}_\mathrm{T}$. Each dimension, $\mathcal{Y}_\tau$, of the outcome space, $\mathcal{Y}_\mathrm{T}$, corresponds to an outcome, υ, had intervention, τ, occurred given initial state χ. Instances of potential outcomes, $\mathrm{Y_T}$, are denoted by the variable, $\mathrm{y_T} = \mathrm{Y_T}(\upsilon, \chi)$. Potential outcomes are fundamentally metaphysical [Daw21] and cannot be fully observed because we only have access to the dimension corresponding to a factually occurring intervention [Hol86]. We denote interventions by the random variable $\mathrm{T} \in \mathcal{T}$ and instances of interventions are denoted by $\mathrm{t} = \mathrm{T}(\tau, \chi)$. The dimensionality of the potential outcome space, $\mathcal{Y}_\mathrm{T}$, is equal to the cardinality of intervention space, $|\mathcal{T}|$. We denote the in-principle observable dimension of a potential outcome by the random variable $\mathrm{Y_t} \in \mathcal{Y}_\mathrm{t}$ and denote instances of $\mathrm{Y_t}$ by $\mathrm{y_t} = \mathrm{Y_t}(\chi)$. We can read this as, "an outcome had intervention, t, been applied given initial state χ."

2.1 Observational Data

In this work, we focus on observational data comprised of covariates representing the initial state, χ, factual treatments, and observed outcomes. We denote observable covariates by the random variable $\mathbf{X} \in \mathcal{X}$. For clarity, we will assume that $\mathcal{X}$ is a d-dimensional continuous space: $\mathcal{X} \subseteq \mathbb{R}^d$, but this does not preclude more diverse

[1] Alternative approaches include decision theoretic causal inference [Daw00; Daw21], Single World Intervention Graphs (SWIGs) [RR13], and Structural Causal Models (SCMs) [Wri34; SS77; Pea09].

spaces. Instances of $\mathbf{X}$ are denoted by $\mathbf{x} = \mathbf{X}(\chi)$. We denote observable interventions by the random variable $\mathrm{T} \in \mathcal{T}$. We will consider binary-valued interventions, $\mathcal{T} \equiv \{0, 1\}$; discrete-valued interventions, $\mathcal{T} \subseteq \mathbb{N}$; and continuous-valued interventions, $\mathcal{T} \subseteq \mathbb{R}$. Instances of of interventions, T, are denoted by t. We denote observable outcomes by the random variable $\mathrm{Y} \in \mathcal{Y}$. In this thesis we consider discrete-valued outcomes, $\mathcal{Y} \subseteq \mathbb{N}$; and continuous-valued outcomes, $\mathcal{Y} \subseteq \mathbb{R}$. Instances of Y are denoted by $\mathrm{y} = \mathrm{Y}(\upsilon, \tau, \chi)$.

We let observational datasets, $\mathcal{D}_n$, consists of n realizations of the random variables, $\mathcal{D}_n = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)\}_{i=1}^n$. We let the observed outcome be the potential outcome of the assigned treatment level, $\mathrm{y}_i = \mathrm{y}_{\mathrm{t}_i}$, thus assuming non-interference and consistency [Rub80]. Moreover, we assume that the tuple of observed covariates, interventions, and outcomes $(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)$ are i.i.d. samples from the joint distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y}_\mathrm{T})$, where $\mathrm{Y}_\mathrm{T} = \{\mathrm{Y}_\mathrm{t} : \mathrm{t} \in \mathcal{T}\}$.

2.2 Causal Estimands

We focus on inference of the following causal estimands: (1) the **conditional average potential outcome (CAPO)** function, (2) the **average potential outcome (APO)** function, (3) the **conditional average treatment effect (CATE)** function, and (4) the **average treatment effect (ATE)**. Each function is defined as an expectation below.

Definition 2.1. *The Conditional Average Potential Outcome (CAPO),*

$$f_\mathrm{t}(\mathbf{x}) \coloneqq \mathbb{E}[\mathrm{Y}_\mathrm{t} \mid \mathbf{X} = \mathbf{x}].$$

Definition 2.2. *The Average Potential Outcome (APO),*

$$f_\mathrm{t} \coloneqq \mathbb{E}[\mathrm{Y}_\mathrm{t}] = \mathbb{E}[f_\mathrm{t}(\mathbf{X})].$$

Definition 2.3. *The Conditional Average Treatment Effect (CATE),*

$$\delta(\mathbf{x}) \coloneqq \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathbf{X} = \mathbf{x}].$$

Definition 2.4. *The Average Treatment Effect (ATE),*

$$\delta \coloneqq \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0] = \mathbb{E}[\delta(\mathbf{X})].$$

2.3 Identifiability Conditions

The above causal estimands are not identifiable from observational data without further assumptions.

Assumption 2.1. ***Stable Unit-Treatment Distribution Assumption (SUTDA)*** *[Daw21]. For a set of unit-intervention pairs, $\{\chi_i, \tau_i\}_{i=1}^n$, the distribution of the outcome, $P(\mathrm{Y}(\chi_i, \tau_i))$, depends only on the factual intervention, τ_i, corresponding to unit, χ_i.*

Assumption 2.2. ***Consistency****. Observed outcomes are identical to potential outcomes under the observed treatment.*

Assumption 2.3. ***Weak Ignorability****. The potential outcome, $\mathrm{Y_t}$, and observed interventions, T, are conditionally independent [Daw79] given measured covariates, $\mathbf{X}$:*

$$\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}.$$

Weak ignorability implies that $P(\mathrm{Y_t} = \mathrm{y_t} \mid \mathbf{X} = \mathbf{x}) = P(\mathrm{Y_t} = \mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, $\forall \mathrm{y_t} \in \mathcal{Y}_\mathrm{t}$, $\forall \mathbf{x} \in \mathcal{X}$, and $\forall \mathrm{t} \in \mathcal{T}$.

Assumption 2.4. ***Positivity****. For a given covariate measurement $\mathbf{X} = \mathbf{x}$, the probability or density for observing intervention $\mathrm{T} = \mathrm{t}$ is greater than zero for all intervention values.*

$$p(\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}) > \eta : \quad \eta > 0, \forall \mathrm{t} \in \mathcal{T}, \forall \mathbf{x} \in \mathcal{X}.$$

Assumptions 2.1 to 2.3 are necessary for the identification of causal estimands by statistical estimands, which we will discuss next. As such, the SUTDA, consistency, and ignorability assumptions are related to structural uncertainty. Assumption 2.4 (positivity) is more related to effect estimation and without it one relies on interpolation or extrapolation to unobserved values. Moreover, for finite data, estimates of effects under weak overlap (low propensity for treatment) may be subject to greater variance than effects under balanced overlap conditions. Correspondingly, the positivity assumption is related to statistical uncertainty.

2.4 Statistical Causal-Effect Estimands

Under Assumptions 2.1 to 2.3, we have that the distribution of potential outcomes given covariates and the distribution of observed outcomes given observed treatments

and covariates are equal, $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = P(\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. So, causal-effect inference as an expectation over potential outcomes under the distribution, $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$, reduces to statistical inference as an expectation over observed outcomes under the distribution, $P(\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. We define the statistical estimands below, which are identical to the causal estimands under Assumptions 2.1 to 2.3.

Definition 2.5. *The Conditional Average Potential Outcome (CAPO) Statistical Estimand,*

$$f(\mathbf{x}, \mathrm{t}) \coloneqq \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}].$$

Definition 2.6. *The Average Potential Outcome (APO) Statistical Estimand,*

$$f(\mathrm{t}) \coloneqq \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}] = \mathbb{E}[f(\mathbf{X}, \mathrm{t})].$$

Definition 2.7. *The Conditional Average Treatment Effect (CATE) Statistical Estimand,*

$$\delta(\mathbf{x}) \coloneqq \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}].$$

Definition 2.8. *The Average Treatment Effect (ATE) Statistical Estimand,*

$$\delta \coloneqq \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 1] - \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 0] = \mathbb{E}[\delta(\mathbf{X})].$$

2.5 Machine Learning for Causal-Effect Estimation

There is a vast literature on inferring causal effects from observational data using machine learning [Hil11; AV17; SJS17; Lou+17; AW19; Che+18; SBV19; NW21; Sch+20; Nie+21; YJV18; CV21b]. Künzel et al. [Kün+19] offers a vantage point from which to understand these methods through the lens of meta-learners. Causal-effect meta-learners allow for arbitrary machine learning estimators to be used for causal-effect inference. An elementary meta-learner is the S-Learner (S for single). The, S-Learner, fits one function, $\widehat{f}(\mathbf{x}, \mathrm{t})$, to an observational dataset, $\mathcal{D}$, to yield an estimator of the CAPO function, $f_{\mathrm{t}}(\mathbf{x})$. Estimates of the CATE function, $\delta(\mathbf{x})$, are then given by taking the difference in CAPO predictions between different treatment arms: $\widehat{\delta}_s(\mathbf{x}) = \widehat{f}(\mathbf{x}, 1) - \widehat{f}(\mathbf{x}, 0)$. Estimates of the APO are taken as the empirical mean of the CAPO estimates: $\widehat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(\mathbf{x}_i, \mathrm{t})$. And estimates of the ATE are taken as the empirical mean of the CATE function: $\widehat{\delta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\delta}(\mathbf{x}_i)$.

2.6 Uncertainty for Causal-Effect Inference

In practice, the identifiability conditions of Section 2.3 rarely hold for observational data. For example, due to observed data, $\mathcal{D}$, being finite, the cardinality of the intervention space, $\mathcal{T}$, and the dimension of the covariate space, $\mathcal{X}$, there will certainly be values, $\mathrm{T} = \mathrm{t}$, that are unobserved for a given covariate measurement, $\mathbf{X} = \mathbf{x}$, leading to violations or near violations of positivity (Assumption 2.4, also known as overlap). We call the uncertainty about treatment effect estimands that arises from this consideration, *statistical*, because it concerns the kind of finite-sample uncertainty common in statistical inference. Moreover, there will almost always be unobserved confounding variables, thus violating ignorability (Assumption 2.3, also known as unconfoundedness or exogeneity). We call uncertainty about treatment effect estimands that arises out of this consideration, *structural*, because it concerns ignorance about the structure defining the causal model. In the following two subsections we will illustrate these two types of uncertainty and motivate the need for scalable causal ML approaches to uncertainty quantification.

2.6.1 Statistical Uncertainty

To illustrate statistical uncertainty in the context of inferring heterogeneous treatment effects from observational data, let's consider a simple setting. Imagine that public health officials in a fantastical universe are interested in reducing heart disease risk. Anecdotal evidence indicates that regular use of a common pain reliever is associated with reduced risk, and a data scientist is tasked with understanding the true effect in order to inform a policy for disease risk reduction. In our fantastical universe there is a moratorium on experimentation until systemic discriminatory recruitment practices leading to adverse outcomes for marginalized groups are resolved; however, their society has democratically consented to a fair and robust data collection policy, and our data scientist has ERB approval to access the needed variables over a representative sample of the target population.

Our imagined universe is so fantastical that one's propensity for taking the pain reliever is completely described by a constant probability for everyone 70 years old and younger, and another constant probability for everyone older than 70 years. Moreover, it is known that heart disease is normally distributed about a mean that depends only on whether or not one is older than 70 years, possibly on the regular use of the pain reliever, and possibly on an interaction effect between the age and

pain reliever indicators. Resolving the latter two possibilities are the causal questions being asked by the data scientist.

Causal Knowledge. Their causal knowledge can be described by a directed acyclic graph (DAG) [Wri34; Dun75; Pea09] as in Figure 2.1, with the variable Y describing heart disease risk, T indicating regular pain reliever use, and X indicating age over 70 years. This diagram says: that age influences pain reliever use, but not that pain reliever use can change one's age; that age influences heart disease risk, but not that heart disease risk can change ones age; and that pain reliever use can influence heart disease risk, but not that heart disease risk influences ones pain reliever use.

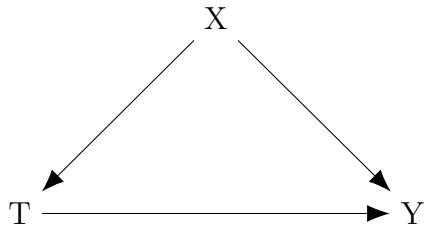


Figure 2.1: Graphical representation of the dependencies.

The causal knowledge can also be described by a system of probabilistic equations, akin to a structural equation model [Haa43; Dun75],

$$\begin{aligned}
\mathrm{x}_i &\sim p(\mathrm{x}), \\
\mathrm{t}_i &\sim \text{Bernoulli}\left(\mathrm{t} \mid (1+\exp(\alpha_0 + \alpha_\mathrm{x}\mathrm{x}_i))^{-1}\right), \\
\mathrm{y}_i &\sim \mathcal{N}\left(\mathrm{y} \mid \beta_0 + \beta_\mathrm{t}\mathrm{t}_i + \beta_\mathrm{x}\mathrm{x}_i + \beta_\mathrm{xt}\mathrm{x}_i\mathrm{t}_i, \sigma_\mathrm{y}^2\right).
\end{aligned}$$

Statistical Inference. Our fictional data scientist is now well positioned to collect a dataset, $\mathcal{D} = \{\mathrm{x}_i, \mathrm{t}_y, \mathrm{y}_i\}_{i=1}^n$, and use a statistical model,

$$\mathrm{y} = \beta_0 + \beta_\mathrm{t}\mathrm{t} + \beta_\mathrm{x}\mathrm{x} + \beta_\mathrm{xt}\mathrm{xt} + \epsilon,$$

to gain estimates for the ATE, δ,

$$\begin{aligned}
\delta &= \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0], \\
&= \beta_\mathrm{t} + \beta_\mathrm{xt}\mathbb{E}[\mathrm{X}],
\end{aligned}$$

and CATEs, $\delta(\mathrm{x})$, for each age group ($\mathrm{x} = 0$, if age ≤ 70 and $\mathrm{x} = 1$, otherwise),

$$\begin{aligned}
\delta(\mathrm{x}) &= \mathbb{E}[\mathrm{Y}_1 - \mathrm{Y}_0 \mid \mathrm{X} = \mathrm{x}], \\
&= \beta_\mathrm{t} + \beta_\mathrm{xt}\mathrm{x}.
\end{aligned}$$

First, they will construct an $n \times 4$ design matrix,

$$\Phi = \begin{bmatrix} 1 & \mathrm{t}_1 & \mathrm{x}_1 & \mathrm{x}_1\mathrm{t}_1 \\ 1 & \mathrm{t}_2 & \mathrm{x}_2 & \mathrm{x}_2\mathrm{t}_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{t}_n & \mathrm{x}_n & \mathrm{x}_n\mathrm{t}_n \end{bmatrix}, \tag{2.2}$$

and a vector of targets, $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_n)^\mathsf{T}$. They can obtain coefficient estimates, $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_\mathrm{t}, \widehat{\beta}_\mathrm{x}, \widehat{\beta}_\mathrm{xt})^\mathsf{T}$, by ordinary least squares (OLS) regression,

$$\widehat{\boldsymbol{\beta}} = \left(\Phi^\mathsf{T}\Phi\right)^{-1}\Phi^\mathsf{T}\mathbf{y}.$$

Finally, the causal estimates are determined from the coefficient estimates as,

$$\begin{aligned}
\widehat{\delta} &= \widehat{\beta}_\mathrm{t} + a\widehat{\beta}_\mathrm{xt} && \text{(ATE)},\\
\widehat{\delta}(\mathrm{X}=0) &= \widehat{\beta}_\mathrm{t} && \text{(CATE: age 70 years and below)},\\
\widehat{\delta}(\mathrm{X}=1) &= \widehat{\beta}_\mathrm{t} + \widehat{\beta}_\mathrm{xt} && \text{(CATE: age above 70 years)},
\end{aligned}$$

where we will assume that the probability of being above 70 years old, $a = \mathbb{E}[\mathrm{X}]$, is known for clarity of following illustrations.

So, is this enough information to inform new risk reduction policies? Say, if the data scientist finds that the ATE is negative — indicating an expected reduction in heart disease risk — then should the public health officials suggest that everyone regularly take the pain reliever? Or, if the CATE for under 70-year-olds is negative and the CATE for over 70-year-olds is positive, should only the below 70-year-olds use the pain reliever regularly? The short answer is, no, they do not have enough information. Even though the denizens of our fantastic universe have perfect structural causal knowledge of the problem, they are still making statistical inferences about causal estimands using finite data. Consequently, to say an estimated effect-size is positive or negative is meaningless without also communicating how that estimate can vary over datasets with equivalent sizes and properties. They must also ask the question, is the magnitude of the effect size significantly non-zero?

Hypothesis Testing. In OLS regression analysis, statisticians communicate uncertainty about the β coefficients through *standard error* estimates, given as the square root of the diagonal of the symmetric coefficient covariance matrix estimate,

$$\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix}
\widehat{\sigma}^2_{\widehat{\beta}_0} & \widehat{\mathrm{Cov}}\left(\widehat{\beta}_0, \widehat{\beta}_\mathrm{t}\right) & \widehat{\mathrm{Cov}}\left(\widehat{\beta}_0, \widehat{\beta}_\mathrm{x}\right) & \widehat{\mathrm{Cov}}\left(\widehat{\beta}_0, \widehat{\beta}_\mathrm{xt}\right)\\
\shortparallel & \widehat{\sigma}^2_{\widehat{\beta}_\mathrm{t}} & \widehat{\mathrm{Cov}}\left(\widehat{\beta}_\mathrm{t}, \widehat{\beta}_\mathrm{x}\right) & \widehat{\mathrm{Cov}}\left(\widehat{\beta}_\mathrm{t}, \widehat{\beta}_\mathrm{xt}\right)\\
\shortparallel & \shortparallel & \widehat{\sigma}^2_{\widehat{\beta}_\mathrm{x}} & \widehat{\mathrm{Cov}}\left(\widehat{\beta}_\mathrm{x}, \widehat{\beta}_\mathrm{xt}\right)\\
\shortparallel & \shortparallel & \shortparallel & \widehat{\sigma}^2_{\widehat{\beta}_\mathrm{xt}}
\end{bmatrix} = \widehat{\sigma}^2_\mathrm{y}\left(\Phi^\mathsf{T}\Phi\right)^{-1}, \qquad (2.3)$$

where $\widehat{\sigma}^2_\mathrm{y}$ is the outcome variance estimate,

$$\widehat{\sigma}^2_\mathrm{y} = \frac{1}{n-4}\sum_{i=1}^{n}\left(\mathrm{y}_i - \widehat{\boldsymbol{\beta}}^\mathsf{T}\Phi_i\right)^2.$$

They use this uncertainty estimate to decide whether a coefficient estimate is significantly non-zero through the process of hypothesis testing [Fis36]. Hypothesis tests consist of: (1) a null hypothesis (e.g. coefficients are zero, no treatment effect, etc); (2) a significance level, α, (the probability of rejecting the null when the null is actually true); (3) a test statistic, t; and (4) a distribution of the test statistic under the null. If the test statistic of the estimate, $t(\widehat{\beta}_{(\cdot)})$, lies outside of the null distribution of the test statistic, $P_0(T)$, with probability less than α, it can be said that the estimated value is significantly different from the null. Common significance levels are 0.01, 0.02, and 0.05. A common test statistic, t, is the ratio of the value estimate, $\widehat{\beta}_{(\cdot)}$, divided by the standard error estimate, $\widehat{\sigma}_{\widehat{\beta}_{(\cdot)}}$. For example, in the case of β_{t}, $t(\widehat{\beta}_{\mathrm{t}}) = \widehat{\beta}_{\mathrm{t}}/\widehat{\sigma}_{\widehat{\beta}_{\mathrm{t}}}$. Intuitively, we can see that the test statistic of the estimate grows as the standard error (uncertainty) of the estimate shrinks. In this setting, the common choice for the test statistic distribution under the null is a Student's t-distribution with degrees of freedom parameter, $n - d_\Phi$, where d_Φ is the number of columns in Φ. Letting, $p_0(t(\widehat{\beta}_{\mathrm{t}})) = P_0(T \geq t(\widehat{\beta}_{\mathrm{t}}))$, be the probability of the test statistic for the estimate under the null, the p-value of the 2-sided test (significantly greater than or less than 0) is given by, $p_{\mathrm{val}} = 2 * \min(p_0(t(\widehat{\beta}_{\mathrm{t}})), 1 - p_0(t(\widehat{\beta}_{\mathrm{t}})))$. One rejects the null-hypothesis of zero treatment effect if, $p_{\mathrm{val}} \leq \alpha$.

It's easy to lose sight of uncertainty when getting into the details about hypothesis testing. Remember that the coefficients themselves are random variables, and we can estimate their variance via Equation (2.3). This induces an estimated distribution over a given coefficient estimate, and should that distribution have significant mass about zero, we may not have sufficient evidence to trust the sign of our estimate.

Let's return to our data scientist now they have the tools necessary to evaluate the significance of their results. For the ATE estimate, $\widehat{\delta} = \widehat{\beta}_{\mathrm{t}} + a\widehat{\beta}_{\mathrm{xt}}$, we have:

$$\widehat{\mathrm{Var}}(\widehat{\delta}) = \widehat{\sigma}^2_{\widehat{\beta}_{\mathrm{t}}} + a^2\widehat{\sigma}^2_{\widehat{\beta}_{\mathrm{xt}}} + 2a\widehat{\mathrm{Cov}}\big(\widehat{\beta}_{\mathrm{t}}, \widehat{\beta}_{\mathrm{xt}}\big),$$

which is straightforwardly computed given the entries in the matrix of Equation (2.3). The test statistic of the estimated ATE is then given by $t(\widehat{\delta}) = \widehat{\delta}/\sqrt{\widehat{\mathrm{Var}}(\widehat{\delta})}$, and the hypothesis testing procedure described above can be followed to establish significance. For the CATE estimates, $\widehat{\delta}(\mathrm{X} = 0) = \widehat{\beta}_{\mathrm{t}}$ and $\widehat{\delta}(\mathrm{X} = 1) = \widehat{\beta}_{\mathrm{t}} + \widehat{\beta}_{\mathrm{xt}}$, we have,

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\delta}(\mathrm{X} = 0)), &= \widehat{\sigma}^2_{\widehat{\beta}_{\mathrm{t}}} \\
\widehat{\mathrm{Var}}(\widehat{\delta}(\mathrm{X} = 1)), &= \widehat{\sigma}^2_{\widehat{\beta}_{\mathrm{t}}} + \widehat{\sigma}^2_{\widehat{\beta}_{\mathrm{xt}}} + \widehat{\mathrm{Cov}}\big(\widehat{\beta}_{\mathrm{t}}, \widehat{\beta}_{\mathrm{xt}}\big),
\end{aligned}$$

from which we can obtain corresponding test statistic values and complete the hypothesis tests.

2.6.1.1 Challenges

This example glosses over some major challenges:

1. The functional form of the model is typically unknown.
2. The observational data can contain complex spatiotemporal modalities, such as images, text, or sensor arrays.
3. The total number of observations, n, may be very large.

In Chapter 4 we present our contributions addressing these challenges.

2.6.1.2 Uncertainty Aware Machine Learning Methods for Causal-Effect Estimation

Several machine learning methods for quantifying statistical uncertainty in the context of causal-effect estimation have been proposed prior to to the contributions of this thesis. Notably, Bayesian Additive Regression Trees (BART) [Hil11], Causal Multi-task Gaussian Processes (CMGP) [AV17], and Causal Forests [AW19]. While these methods provide solutions that are amenable to arbitrary, non-linear response surfaces, they still require that the covariates, $\mathbf{x}$, are tabular in nature and are not well-suited to handling raw input modalities such as images, text, or other complex spatiotemporal signals. Moreover, they may not scale to very large n.

Active Learning of Treatment Effects So now the data scientist can report their results to the public health officials, but what can they do if those results fail to show significance? Deng, Pineau, and Murphy [DPM11] propose the use of Active Learning for recruiting patients to assign treatments that will reduce the uncertainty of an Individual Treatment Effect model. Sundin et al. [Sun+19] propose using a Gaussian process (GP) to model the individual treatment effect and use the expected information gain over the S-type error rate, defined as the error in predicting the sign of the CATE, as their acquisition function. Although GPs are suitable for quantifying uncertainty, they do not work well on high-dimensional input spaces. In this work, we use Neural network methods to obtain uncertainty: Deep Ensembles [LPB17] and DUE [Van+21], a Deep Kernel Learning GP, both of which work well even on high dimensional inputs. Recent work by Qin, Wang, and Zhou [QWZ21] looks at budgeted heterogeneous effect estimation but does not factor weak or limited overlap into their acquisition function.

2.6.2 Structural Uncertainty

A common source of structural uncertainty in treatment effect estimation from observational data is hidden confounding. A confounding variable is a causal parent of both the intervention, T, and the outcome, Y. Throughout our discussion thus far, we have treated the observed covariates, $\mathbf{X}$, as confounders or proxies for confounders. We depict a hidden confounder, U, graphically in Figure 2.2. In contrast to the observed confounder, X, the hidden confounder, U, is either unknown, cannot be observed, or is not included in the observational dataset, $\mathcal{D} = \{\mathbf{x}, \mathrm{t}, \mathrm{y}\}_{i=1}^{n}$. A hidden confounder results in a violation of Assumption 2.3 (ignorability). That is, while the conditional independence relationship, $\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathrm{X}, \mathrm{U}$, may hold, the relationship, $\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathrm{X}$, does not. This violation of ignorability can result in an arbitrarily biased causal effect estimate. This is a particularly acute problem, since such violations cannot be identified from observational data alone. Understanding how ignorability violations can bias treatment effect estimates is the domain of causal sensitivity analysis.

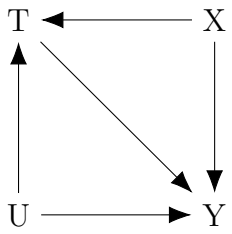


Figure 2.2: Graphical representation of hidden confounding.

2.6.2.1 Causal Sensitivity Analysis Methods

Causal sensitivity analysis includes a diverse family of frameworks, whose common goal is to give bounds on the treatment-effect under the assumption of some "level" of unobserved confounding, either at a population level for the ATE [RR83; RRS00; Imb03; Ros14; Dor+16; FDF19; VZ20; DG23] or at the level of individuals for the CATE [Yad+18; KMZ19; Opr+23].

Continuous Valued Treatments. The prior literature for continuous-valued treatments has focused largely on parametric methods assuming linear treatment/outcome, hidden-confounder/treatment, and hidden-confounder/outcome relationships [CHH16; Dor+16; Mid+16; Ost19; CH20b; CH20a]. In addition to linearity, these parametric methods need to assume the structure and distribution of the unobserved confounding variable(s). [Cin+19] allows for sensitivity analysis for arbitrary structural causal models under the linearity assumption. The methods we present in this thesis address both the distributional and linearity assumptions.

A two-parameter sensitivity model based on Riesz-Fréchet representations of the target functionals, here the APO and CAPO, is proposed by [Che+21] as a way to incorporate confidence intervals and sensitivity bounds. In contrast, we use the theoretical background of the marginal sensitivity model to derive a one-parameter sensitivity model. [Mar+23] derive a sensitivity model that bounds the partial derivative of the log density ratio between complete and nominal propensity densities. Bounding the effects of continuous valued interventions has also been explored using instrumental variable models [KKS20; Hu+21; Pad+22].

Chapter 3

Methodological Background

3.1 Scalable Causal-Effect Inference

Let's review our desiderata. We would (1) like methods for causal-effect inference that scale to very-large dataset sizes because uncertainty about a statistical estimand under a well specified model generally decreases with increasing n. We would (2) like methods for causal-effect inference that scale to high-dimensional, multi-modal datasets because real-world data are often not given as a design matrix without preprocessing. Neural networks have these desirable traits and have been extensively adapted for estimating ATEs, CATEs, APOs, and CAPOs [SJS17; Lou+17; YJV18; SBV19; Sch+20]

The simplest form of a neural network can be described by the function,

$$f(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}_1^\mathsf{T} \psi(\boldsymbol{\theta}_0^\mathsf{T} \mathbf{x}),$$

where $\boldsymbol{\theta}_0$ is an m-by-d matrix of real valued scalars, $\psi(\cdot)$ is a non-linear function that operates element-wise on $\boldsymbol{\theta}_0^\mathsf{T}\mathbf{x}$, and $\boldsymbol{\theta}_1$ is a 1-by-m matrix of real valued scalars. The neural-network parameters are "fit" to an observed dataset, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^n$, by minimizing the negative log-likelihood of the data under the network,

$$-\log p(\mathbf{y} \mid \mathrm{X}, \boldsymbol{\theta}) = -\log \prod_{i=1}^{n} p(\mathrm{y}_i \mid, \mathbf{x}_i, \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p(\mathrm{y}_i \mid f(\mathbf{x}_i, \boldsymbol{\theta})).$$

Here, $\mathbf{y} = (\mathrm{y}_i, \ldots, \mathrm{y}_n)^\mathsf{T}$, and $\mathbf{X}$ is a matrix with rows corresponding to each observed $\mathbf{x}_i$. Typical choices for the likelihood function, $p(\mathrm{y} \mid f(\mathbf{x}_i, \boldsymbol{\theta}))$, are a normal density, $\mathcal{N}(\mathrm{y} \mid f(\mathbf{x}, \boldsymbol{\theta}), \sigma_\mathrm{y}^2)$, for regression problems; or a Bernoulli distribution, $\mathrm{Bern}(\mathrm{y} \mid$

$\text{sig}(f(\mathbf{x}, \boldsymbol{\theta}))$ for classification problems[1]. In both cases, the expected value of either distribution is parametrized by the network output, $f(\mathbf{x}, \boldsymbol{\theta})$.

Neural networks are able to satisfy our first desideratum of scaling to large n because optimization is done via stochastic gradient descent [RM51; Bot98]. The elementary form of stochastic gradient descent proceeds by sequentially and randomly sampling subsets of the training data, $\mathcal{D}_t = \{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^{b} : b \ll n$, and updating the weights according to,

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \frac{1}{b} \sum_{i=1}^{b} \widehat{\nabla}_{\boldsymbol{\theta}} \log p(\mathrm{y}_i \mid f(\mathbf{x}_i, \boldsymbol{\theta}_{1-t})),$$

where $\widehat{\nabla}_{\boldsymbol{\theta}}$ is a gradient estimator employing the backpropagation algorithm [Wer74; RHW86]. For modern neural network optimization, the Adam algorithm [KB14] is commonly used instead of the elementary update equation above.

Neural networks for causal-effect estimation are fit to datasets comprised of observed covariates, interventions, and outcomes, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^{n}$. Under Assumptions 2.1 to 2.4, the objective function is the negative log-likelihood of the data under the neural network,

$$-\log p(\mathbf{y} \mid \mathbf{t}, \mathrm{X}, \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p(\mathrm{y}_i \mid f(\mathbf{x}_i, \mathrm{t}_i, \boldsymbol{\theta})),$$

where the variables $\mathbf{y}$ and X are as defined above, and $\mathbf{t} = (\mathrm{t}_1, \ldots, \mathrm{t}_n)^{\mathsf{T}}$ is a vector of observed treatments. As such, neural networks can be optimized in the regular way and both satisfy our first scalability desideratum and also yield an estimator for the CAPO, $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta})$. For discrete valued interventions, the CATE estimates are then given by $\delta(\mathbf{x}, \boldsymbol{\theta}) = f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}) - f(\mathbf{x}, \mathrm{t}_0, \boldsymbol{\theta})$, for a given intervention, t, and a reference intervention, t_0. For example, in the binary treatment regime we would have, $\mathrm{t} = 1$ and $\mathrm{t}_0 = 0$.

To see how such models can scale to arbitrary data modalities and satisfy our second desideratum, we focus on the CAPO estimate (or logits of the CAPO estimate in the case of discrete outcomes), $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta})$, and a motif in architecture design for causal-effect inference. The design motif is illustrated in Figure 3.1. Here, the covariates, $\mathbf{x}$, are first passed through the neural network function, $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E)$. The resulting output is then concatenated with the intervention indicator, t, and passed through a second neural network function, $\boldsymbol{\phi}(\cdot, \boldsymbol{\theta}_M)$. We denote the composed operation as,

$$\boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_B) = \boldsymbol{\phi}(\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E), \mathrm{t}, \boldsymbol{\theta}_M).$$

[1] $\text{sig}(\cdot)$ denotes the logistic sigmoid function, $\text{sig}(\mathrm{z}) = (1 + \exp(-\mathrm{z}))^{-1}$

The CAPO function is finally given as, $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}) = \boldsymbol{\theta}_f^\intercal \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_B)$. There are two main components of this motif: (1) the transformation of of the covariates to an m-dimensional feature representation, $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E) : \mathcal{X} \to \mathbb{R}^m$; and (2) the subsequent transformation of the feature representation and treatment to the outcome space, $\boldsymbol{\theta}_f^\intercal \boldsymbol{\phi}(\cdot, \boldsymbol{\theta}_M) : \mathbb{R}^{m+d_t} \to \mathcal{Y}$, where d_t is the dimensionality of the intervention space. The first transformation, $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E)$, we call the encoder of the network. The second transformation, $\boldsymbol{\theta}_f^\intercal \boldsymbol{\phi}(\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E), \mathrm{t}, \boldsymbol{\theta}_M)$, we call the treatment-effect meta-learner. From Section 2.5, you can recognize the meta-learner depicted in this motif as an S-Learner [Kün+19]. T-Learners and more elaborate configurations are also commonly used [SJS17; Lou+17; YJV18; SBV19; Sch+20; CV21a; CV21b].

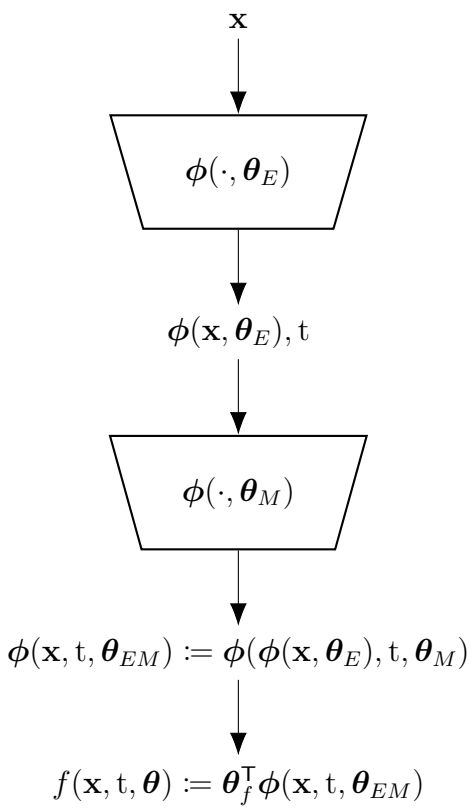


Figure 3.1: Neural network design motif for causal effect estimation

The beauty in this motif is the flexibility that the encoder provides for scaling to arbitrary input spaces, $\mathcal{X}$. For example, if images comprise the input space, then one can use a Convolutional Neural Network (CNN) [LeC+98; KSH12; He+16] or Vision Transformer (ViT) [Dos+20]. If text comprises the input space, then one could use an LSTM [HS97] or Transformer [Vas+17]. If the input space is multi-modal, then one could use a Perceiver [Jae+21]. The choice will largely depend on the dependencies (temporal, spatial, etc.) that need to be captured in the input space to derive appropriate vector representations of any confounders that may lie latent in the input data. The two main advantages of using the S-Learner are (1) it generalizes to binary, discrete, continuous, or even multi-dimensional intervention spaces, and (2) it can capture shared structure among response surfaces between different intervention levels.

3.1.1 Scalable Measures of Statistical Uncertainty

For uncertainty-aware ML methods, the mutual information [Sha48; Mac03] of labels, Y, and functions, F, (or labels, Y, and neural-network weights, $\boldsymbol{\Theta}$) is a common measure of statistical uncertainty about a function, $f(\mathbf{x})$, for observed covariates, $\mathbf{x}$, given training data, $\mathcal{D}$. Mutual information is typically denoted as $\mathrm{I}(\mathrm{Y}; \mathrm{F} \mid \mathbf{X} = \mathbf{x}, \mathcal{D})$ or $\mathrm{I}(\mathrm{Y}; \boldsymbol{\Theta} \mid \mathbf{X} = \mathbf{x}, \mathcal{D})$ when conditioned on $\mathbf{x}$. Using F-notation for illustration, mutual information can be expressed in terms of entropies as either

$$\mathrm{I}(\mathrm{Y}; \mathrm{F} \mid \mathbf{x}, \mathcal{D}) = \mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathcal{D}) - \mathrm{H}(\mathrm{Y} \mid \mathrm{F}(\mathbf{x}), \mathcal{D}), \tag{3.1}$$

or

$$\mathrm{I}(\mathrm{Y}; \mathrm{F} \mid \mathbf{x}, \mathcal{D}) = \mathrm{H}(\mathrm{F}(\mathbf{x}) \mid \mathcal{D}) - \mathrm{H}(\mathrm{F}(\mathbf{x}) \mid \mathrm{Y}, \mathcal{D}). \tag{3.2}$$

In general, entropy is given as $\mathrm{H}(\mathrm{A}) = -\int \log(p(\mathrm{a}))dP(\mathrm{a})$ and conditional entropy is given as $\mathrm{H}(\mathrm{A} \mid \mathrm{B}) = -\iint \log(p(\mathrm{a} \mid b))dP(\mathrm{a} \mid \mathrm{B} = \mathrm{b})dP(\mathrm{b})$, where $p(\mathrm{a})$ is either the probability distribution or density function for discrete or continuous random variables, A, respectively. In the context of machine learning, mutual information can be read as, "the information (in bits or nats) that we would gain about the expected value of the outcome, F, given an input, $\mathbf{X} = \mathbf{x}$, and dataset, $\mathcal{D}$, if we could observe and instance of the outcome, Y, for input, $\mathbf{x}$."

The law of total variance [WHH06] can also be used to derive a scalable measure of statistical uncertainty,

$$\mathrm{Var}(\mathrm{Y} \mid \mathbf{x}, \mathcal{D}) = \mathbb{E}\left[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F}(\mathbf{x}), \mathcal{D})\right] + \mathrm{Var}\left(\mathbb{E}[\mathrm{Y} \mid \mathrm{F}(\mathbf{x}), \mathcal{D}]\right), \tag{3.3}$$

which decomposes the total variance into its irreducible and statistical parts. Since in our setting the random variable F represents the expected value of Y, the statistical component can be shortened to $\mathrm{Var}\left(\mathrm{F}(\mathbf{x}) \mid \mathcal{D}\right)$. We will call the statistical component the **variance information** between random variables, Y and F.

Definition 3.1. *The variance information between an outcome,* Y*, and its expected value,* F*, given a measurement,* $\mathbf{x}$*, and dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^n$ *is defined as,*

$$\mathrm{V}(\mathrm{Y}; \mathrm{F} \mid \mathbf{x}, \mathcal{D}) \coloneqq \mathrm{Var}\left(\mathrm{F}(\mathbf{x}) \mid \mathcal{D}\right). \tag{3.4}$$

Whether we use the mutual information or variance information, we need to treat the mean functions, $f(\mathbf{x})$ as random variables, $\mathrm{F}(\mathbf{x})$, and any of the scalable methods presented in Section 3.2 can be used.

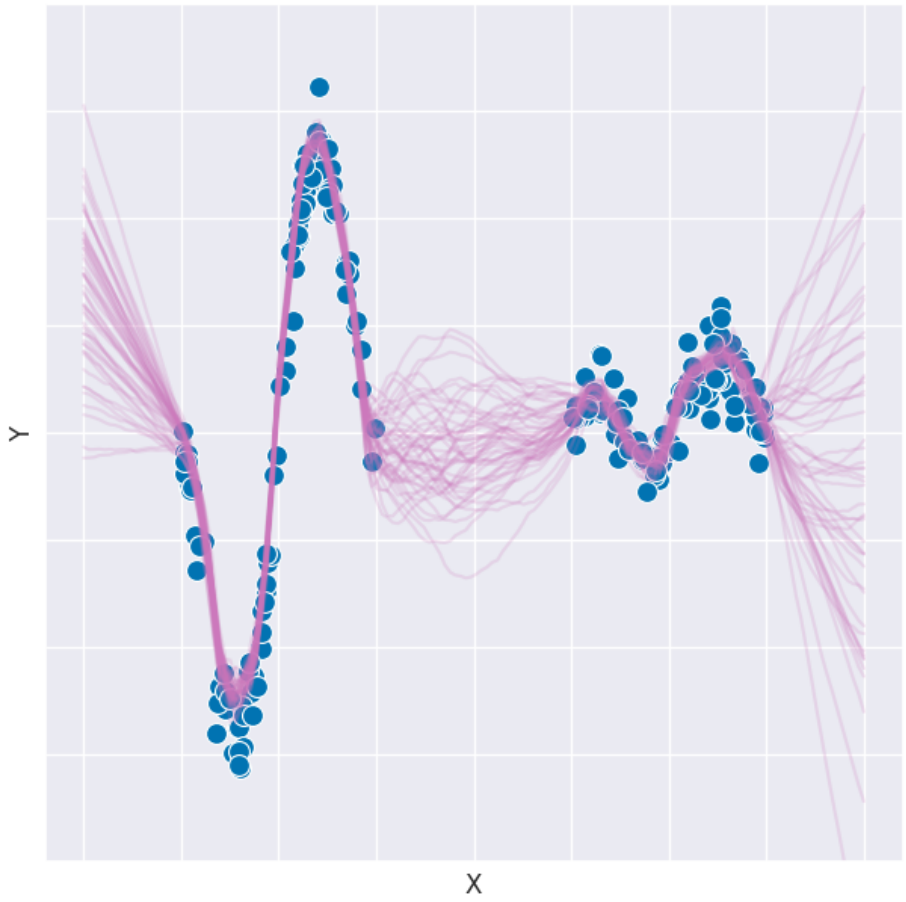


Figure 3.2: Many different functions (purple lines) can explain finite data (blue dots) equally well.

3.2 Scalable Statistical Uncertainty Quantification

Statistical uncertainty arises because there may be any number of functions, $f(\mathbf{x})$, that fit a finite dataset, $\mathcal{D}_n = \{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^n$, equally well. The set of functions depends on both the model class of the function being fit, and the optimization objective used to fit the function to the data. Uncertainty-aware machine learning methods (ensembles [Bre01; LPB17], Gaussian-processes (GPs) [WR06; HMG15; Wil+16; Van+21], Bayesian neural networks (BNNs) [Nea95; Mac95a; Nea12; KW13; GG16], etc.) aim to infer a set, $\mathrm{F} = \{f_i(\mathbf{x})\}_{i=1}^{|\mathrm{F}|}$, or distribution, $P(\mathrm{F} \mid \mathcal{D})$, of such functions. Figure 3.2 illustrates such a set of functions where blue dots correspond to observed $\mathbf{x}$-y pairs, and purple lines correspond to different estimates of the mean function, $f(\mathbf{x})$.

We will consider sets of functions as random variables. For example, we will use the notations, $\mathrm{F} \equiv f(\mathbf{x}; \boldsymbol{\Theta})$, interchangeably to denote a random variable over functions

conditioned on a finite dataset, $\mathcal{D}$. These random variables will be distributed according to $P(\mathrm{F} \mid \mathcal{D}) \equiv P(\boldsymbol{\Theta} \mid \mathcal{D})$, and will have instances $f(\mathbf{x}) \equiv f(\mathbf{x}; \boldsymbol{\theta})$. We use both notations because 'F-notation' is more compact for some concepts and more suitable for ensemble and GP methods, where '$\boldsymbol{\Theta}$-notation' is more compact for other concepts and more suitable to describe BNNs with weights, $\boldsymbol{\theta}$.

3.2.1 Scalable Uncertainty-Aware Machine Learning

3.2.1.1 Bayesian Neural Networks

Similarly to Bayesian linear regression, Bayesian Neural-Networks [Nea95; Mac95a; Mac95b; Bis95] model statistical uncertainty via inference of a posterior distribution, $p(\boldsymbol{\theta} \mid \mathcal{D})$, over network parameters, $\boldsymbol{\theta}$, given an observed dataset, $\mathcal{D}$. Exact posterior inference is often impractical for the large networks needed to effectively model the kind of data we would like to scale to, so we focus instead on the subclass of methods that use approximate Bayesian inference and yield an approximate posterior, $q(\boldsymbol{\theta} \mid \mathcal{D})$ [Hof+13] [HA15; LHT15; GG16; Her+16; Kha+18; Sun+18]. For example, The *dropout as Bayesian approximation* framework estimates the integral over the log evidence lower bound objective using Monte-Carlo integration. For a single sample from the approximate posterior density, $\boldsymbol{\theta} \sim q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}, p_d)$, the integrand is of the general form:

$$\mathcal{L} = \sum_{i=1}^{|\mathcal{D}|} \log p(\mathrm{y} \mid \mathbf{x}, \boldsymbol{\theta}) - \mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta})). \tag{3.5}$$

Such BNNs can be trained via stochastic gradient descent of the negative log evidence lower bound, and so satisfy our scale constraints.

3.2.1.2 Deep Ensembles

Instead of learning a posterior over functions, Deep Ensembles [LPB17; Izm+18] instead train multiple networks over the training data, which results in a set of neural network weights, $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \mid \mathcal{D}\}$, which induce a set of functions.

3.2.1.3 Deep Kernel Learning

Deep kernel learning [Wil+16; Liu+20; Van+21] combines a neural network feature extractor, with inducing point inference [HMG15], and Gaussian Processes [RN10] to provide a scalable method for function space variational inference.

3.2.1.4 Other Methods

Neural linear models [Sno+15], neural processes [Gar+18], epistemic neural networks [Osb+21], and transformers [Mül+21] each provide exciting scalable alternatives for modelling statistical uncertainty.

3.3 Sensitivity Analysis for Structural Uncertainty

Tan [Tan06] introduces the Marginal Sensitivity Model (MSM) as a means to express estimated treatment effect bias, $\mathbb{E}[\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = x] - \mathbb{E}[\mathrm{Y_t} \mid \mathbf{X} = x]$, induced by violations of Assumption 2.3 (weak-ignorability), $\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$, in terms of an assumed discrepancy between the conditional *potential outcome* distribution, $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$, and the conditional outcome distribution, $P(\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. The MSM will serve as the foundation for our scalable quantification of structural uncertainty in Chapter 5, as it is interpretable in discrete intervention settings, it makes minimal assumptions about the underlying sources of structural uncertainty, and it is amenable to scalable treatment-effect estimators. In Section 3.3.1, we detail Tan [Tan06]'s original formulation in terms of binary-valued treatments. In Section 5.1 we present two extensions of the MSM: (1) an MSM for arbitrary discrete-valued treatments in Section 5.1.1, and (2) an MSM for continuous-valued treatments in Section 5.1.2.

Kallus, Mao, and Zhou [KMZ19] use the MSM to provide a tractable algorithm for CATE bounds in the context of kernel regression methods when the ignorability assumption is relaxed. In Section 3.3.2, we show how they frame bounds for the CAPO, $f_{\mathrm{t}}(\mathbf{x})$ and CATE, $\delta(\mathbf{x})$, under the MSM for binary interventions. In Section 5.2, we show how their methodology can be extended to scalable machine learning methods for CATE and CAPO bounds under both discrete and continuous interventions. In Section 5.3 we provide estimators for the proposed bounds. And in Section 5.4, we further incorporate statistical uncertainty for finite-sample bounds on the CAPO, APO, CATE, and ATE estimands.

3.3.1 Marginal Sensitivity Model for Binary-Valued Interventions

For binary treatments, $\mathcal{T}_B = \{0, 1\}$, the (nominal) propensity score, $e(\mathbf{x}) \equiv p(\mathrm{T} = 1 \mid \mathbf{X} = \mathbf{x})$, states how the treatment status, t, depends on the covariates, $\mathbf{x}$, and is identifiable from observational data. The potential outcomes, $\mathrm{Y_0}$ and $\mathrm{Y_1}$, conditioned on the covariates, $\mathbf{x}$, are distributed as $P(\mathrm{Y_0} \mid \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_1} \mid \mathbf{X} = \mathbf{x})$,

respectively. Each of these conditional distributions can be written as mixtures of the factual and counterfactual conditional outcome distributions with weights based on the propensity score:

$$P(\mathrm{Y}_0 \mid \mathbf{X} = \mathbf{x}) = (1 - e(\mathbf{x}))P(\mathrm{Y}_0 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}) + e(\mathbf{x})P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}),$$
$$P(\mathrm{Y}_1 \mid \mathbf{X} = \mathbf{x}) = (1 - e(\mathbf{x}))P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}) + e(\mathbf{x})P(\mathrm{Y}_1 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}).$$

Without further assumptions, the conditional distributions of each potential outcome given the observed treatment, $P(\mathrm{Y}_0 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$, are identifiable from observational data, but the conditional distributions of each potential outcome given the counterfactual treatment, $P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$ are not. Under Assumption 2.3 (ignorability), $\mathrm{Y}_\mathrm{t} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X} = \mathbf{x}$; therefore, $P(\mathrm{Y}_\mathrm{t} \mid \mathbf{X} = \mathbf{x}) = P(\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, which implies that the factual and counterfactuals are equivalent: $P(\mathrm{Y}_0 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y}_1 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$. Therefore, any deviation from these equalities will be indicative of hidden confounding. However, because the distributions $P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$ are unidentifiable, the MSM can only postulate a relationship between each pair of identifiable and unidentifiable components.

The MSM assumes that $P(\mathrm{Y}_\mathrm{t} \mid \mathrm{T} = 1-t, \mathbf{X} = \mathbf{x})$ is absolutely continuous with respect to $P(\mathrm{Y}_\mathrm{t} \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})$ for all $\mathrm{t} \in \mathcal{T}_B$. Therefore, given that $P(\mathrm{Y}_\mathrm{t} \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_\mathrm{t} \mid \mathrm{T} = 1-t, \mathbf{X} = \mathbf{x})$ are σ-finite measures, by the Radon-Nikodym theorem, there exists a function $\lambda_B(\mathrm{y}_\mathrm{t}, \mathbf{x}) : \mathcal{Y} \to [0, \infty)$ such that,

$$P(\mathrm{Y}_\mathrm{t} \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x}) = \int_{\mathcal{Y}} \lambda_B(\mathrm{y}_\mathrm{t}, \mathbf{x}) dP(\mathrm{y}_\mathrm{t} \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x}). \tag{3.6}$$

Rearranging terms allows for the expression of the integrand, $\lambda_B(\mathrm{y}_\mathrm{t}, \mathbf{x})$, as the Radon-Nikodym derivative,

$$\lambda_B(\mathrm{y}_\mathrm{t}, \mathbf{x}) = \frac{dP(\mathrm{y}_\mathrm{t} \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x})}{dP(\mathrm{y}_\mathrm{t} \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}. \tag{3.7}$$

Thus, it can also be expressed as a ratio of probabilities or probability densities, $\frac{p(\mathrm{y}_\mathrm{t} \mid \mathrm{T}=1-t, \mathbf{X}=\mathbf{x})}{p(\mathrm{y}_\mathrm{t} \mid \mathrm{T}=t, \mathbf{X}=\mathbf{x})}$, depending on whether the outcome is discrete or continuous. Then, by Bayes's rule, $\lambda_B(\mathrm{y}_0, \mathbf{x})$ and $\lambda_B(\mathrm{y}_1, \mathbf{x})$ are expressed as odds ratios,

$$\begin{aligned} \lambda_B(\mathrm{y}_0, \mathbf{x}) &= \frac{1 - e(\mathbf{x})}{e(\mathbf{x})} \Big/ \frac{1 - e(\mathrm{y}_0, \mathbf{x})}{e(\mathrm{y}_0, \mathbf{x})}, \\ \lambda_B(\mathrm{y}_1, \mathbf{x}) &= \frac{e(\mathbf{x})}{1 - e(\mathbf{x})} \Big/ \frac{e(\mathrm{y}_1, \mathbf{x})}{1 - e(\mathrm{y}_1, \mathbf{x})}, \end{aligned} \tag{3.8}$$

where $e(\mathrm{y_t}, \mathbf{x}) \equiv P(\mathrm{T} = 1 \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})$ is the unidentifiable complete propensity for treatment.

Finally, the MSM postulates that the odds of receiving the treatment $\mathrm{T} = 1$ for subjects with covariates $\mathbf{X} = \mathbf{x}$ can only differ from $e(\mathbf{x})/(1 - e(\mathbf{x}))$ by at most a factor of Λ,

$$\Lambda^{-1} \leq \lambda_B(\mathrm{y_t}, \mathbf{x}; t) \leq \Lambda. \tag{3.9}$$

We can then define the MSM for binary treatments as a set of functions $\mathcal{P}_{\mathrm{B}}(\Lambda)$ that satisfy this inequality.

Definition 3.2. *Marginal Sensitivity Model (MSM)*

$$\mathcal{P}_B(\Lambda) \coloneqq \left\{ w(\mathrm{y_t}, \mathbf{x}) \coloneqq \frac{1}{e(\mathrm{y_t}, \mathbf{x})} : \alpha_B(\mathbf{x}, \Lambda) \leq w(\mathrm{y_t}, \mathbf{x}) \leq \beta_B(\mathbf{x}, \Lambda) \right\},$$

$\forall \mathrm{y} \in \mathcal{Y}_t, \forall \mathbf{x} \in \mathcal{X}$. *Where,*

$$\alpha_B(\mathbf{x}, \Lambda) = \frac{1}{\Lambda e(\mathbf{x})} + 1 - \frac{1}{\Lambda}, \quad \textit{and} \quad \beta_B(\mathbf{x}, \Lambda) = \frac{\Lambda}{e(\mathbf{x})} + 1 - \Lambda.$$

This set of functions is defined in terms of the identifiable propensity for treatment, $e(\mathbf{x})$, and a user defined parameter, Λ, that encodes a belief in the degree of divergence from the ignorability assumption.

3.3.2 CAPO and CATE Bounds Under the MSM

Kallus, Mao, and Zhou [KMZ19] use the following factorization of the CAPO in order to derive CATE and CAPO under the MSM,

$$f_{\mathrm{t}}(\mathbf{x}) = \frac{\int_{\mathcal{Y}} \mathrm{y_t} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, \tag{3.10}$$

which elegantly expresses the unbiased conditional expectation of the potential outcome in terms of the unidentifiable inverse complete propensity $w(\mathrm{y_t}, \mathbf{x}) = 1/e(\mathrm{y_t}, \mathbf{x})$ and the conditional probability or density $p(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})$ of the outcome. We provide a derivation of this factorization in Lemma D.10.

Defining the r.h.s. term as,

$$f_{\mathrm{t}}\big(\mathbf{x}, w(\mathrm{y_t})\big) \coloneqq \frac{\int_{\mathcal{Y}} \mathrm{y_t} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, \tag{3.11}$$

they then define lower, $\underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)$, and upper, $\overline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)$, bounds on the CAPO under Λ.

Definition 3.3. *CAPO Bounds,*

$$\underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda) \coloneqq \inf_{w(\mathrm{y_t}, \mathbf{x}) \in \mathcal{P}(\Lambda)} \left\{ f_{\mathrm{t}}\big(\mathbf{x}, w(\mathrm{y_t})\big) \right\}, \tag{3.12a}$$

$$\overline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda) \coloneqq \sup_{w(\mathrm{y_t}, \mathbf{x}) \in \mathcal{P}(\Lambda)} \left\{ f_{\mathrm{t}}\big(\mathbf{x}, w(\mathrm{y_t})\big) \right\}. \tag{3.12b}$$

The CAPO bounds then completely determine the CATE bounds.

Definition 3.4. *CATE Bounds,*

$$\underline{\delta}(\mathbf{x}, \Lambda) \coloneqq \underline{f}_1(\mathbf{x}, \Lambda) - \overline{f}_0(\mathbf{x}, \Lambda), \tag{3.13a}$$

$$\overline{\delta}(\mathbf{x}, \Lambda) \coloneqq \overline{f}_1(\mathbf{x}, \Lambda) - \underline{f}_0(\mathbf{x}, \Lambda). \tag{3.13b}$$

Which gives an ignorance interval over possible CATE functions.

Definition 3.5. *CATE ignorance interval,*

$$\mathcal{U}(\mathbf{x}, \Lambda) = \left[\underline{\delta}(\mathbf{x}, \Lambda), \overline{\delta}(\mathbf{x}, \Lambda) \right]. \tag{3.14}$$

The ignorance interval $\mathcal{U}(\mathbf{x}; \Lambda)$ under an assumed Λ is completely defined with respect to identifiable estimands. For example, a parametric normal likelihood function in can be used to model the density $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$ and a model with Bernoulli distribution, $P(\mathrm{T} = \mathrm{t} \mid \mathbf{x}, \boldsymbol{\omega}) = \mathrm{Bern}\left(\mathrm{t} \mid \phi(\mathbf{x}, \boldsymbol{\omega})\right)$, can be used to model the identifiable nominal propensity for treatment $e(\mathbf{x})$. Kallus, Mao, and Zhou [KMZ19] use a non-parameteric kernel based method and discrete line search to learn a function that maps $\mathbf{x}$ to the identifiable CATE intervals: $\mathcal{U}(\mathbf{x}; \Lambda)$. Figure 3.3 illustrates the bounds given by such a model for given assumptions on Λ.

For average treatment effects, there are two approaches for interpreting the bounds on the CATE, $\delta(\mathbf{x})$ [Tan06]. One approach seeks the smallest value Λ_s such that the interval $[\underline{\delta}(\mathbf{x}; \Lambda_s), \overline{\delta}(\mathbf{x}; \Lambda_s)]$ crosses 0. This approach then reports that the CATE becomes sensitive to hidden confounding at Λ_s. The other approach sets a cutoff Λ_c and examines how the CATE changes for plausible λ values below Λ_c.

There are two main limitations of the approach of Kallus, Mao, and Zhou [KMZ19] that our contributions seek to address. First, as is evident in the regions of $\mathbf{x}$ that lie out of distribution ($\mathbf{x} < -2.5 \quad \text{or} \quad \mathbf{x} > 2.5$), the bounds become nonsensical (as expected), and there is no way to identify that a measurement $\mathbf{x}$ is actually out of

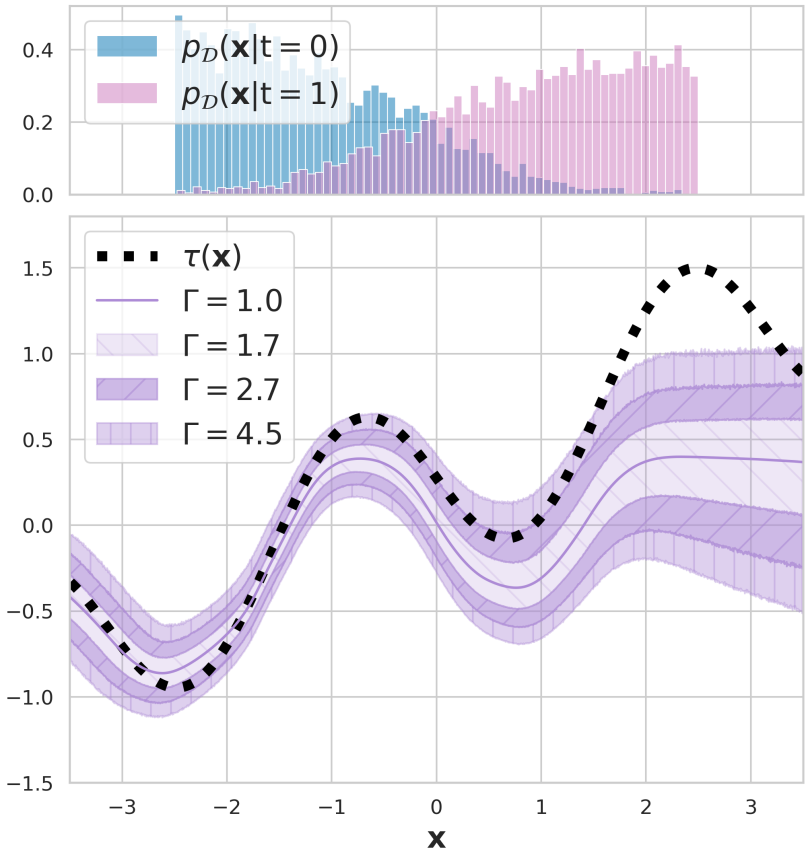


Figure 3.3: Varying Λ for Marginal Sensitivity Model. Ground truth $\Lambda^* = 2.7$.

distribution; more generally, it does not account for sources of ignorance other than violations of ignorability. Second, the method does not scale well computationally to large sample sizes, and does not scale well statistically to high-dimensional datasets as it relies on weighted kernel regression to estimate the outcome.

3.4 Active Learning and Experimental Design

Formally, an active learning setup consists of an unlabeled dataset $\mathcal{D}_{\text{pool}} = \{\mathbf{x}_i\}_{i=1}^{n_{\text{pool}}}$, a labeled training set $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^{n_{\text{train}}}$, and a predictive model with likelihood $p(\mathrm{y} \mid \mathbf{x}, \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})$. The setup also assumes that an oracle exists to provide outcomes y for any data point in $\mathcal{D}_{\text{pool}}$. After model training, a batch of data $\{\mathbf{x}_i^*\}_{i=1}^b$ is selected from $\mathcal{D}_{\text{pool}}$ using an acquisition function a according to the informativeness of the batch.

An intuitive way to define informativeness is using the estimated uncertainty of our model. In general, we can distinguish two sources of uncertainty: epistemic and

aleatoric uncertainty [DD09; KG17]. Epistemic (or model) uncertainty, arises from ignorance about the model parameters. For example, this is caused by the model not seeing similar data points during training, so it is unclear what the correct label would be. We focus on using epistemic uncertainty to identify the most informative points for label acquisition.

Bayesian Active Learning by Disagreement (BALD) [Hou+11] defines an acquisition function based on epistemic uncertainty. Specifically, it uses the mutual information (MI) between the unknown output and model parameters as a measure of disagreement:

$$\mathrm{I}(\mathrm{Y}; \boldsymbol{\theta} \mid \mathbf{x}, \mathcal{D}_{\text{train}}) = \mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathcal{D}_{\text{train}}) - \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})} \left[\mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \boldsymbol{\theta}) \right], \tag{3.15}$$

where H is the entropy function; a straightforward estimand for discrete outcomes with Bernoulli or categorical likelihoods.

The general acquisition function based on BALD for acquiring a batch of data points given the pool dataset and the model parameters is given by the joint mutual information between the set $\{\mathrm{Y}_i\}$ and the model parameters [KVG19]:

$$a_{\text{BALD}}(\mathcal{D}_{\text{pool}}, p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})) = \underset{\{\mathbf{x}_i\}_{i=1}^{b} \subseteq \mathcal{D}_{\text{pool}}}{\arg\max} \; \mathrm{I}(\{\mathrm{Y}_i\}; \boldsymbol{\theta} \mid \{\mathbf{x}_i\}, \mathcal{D}_{\text{train}}). \tag{3.16}$$

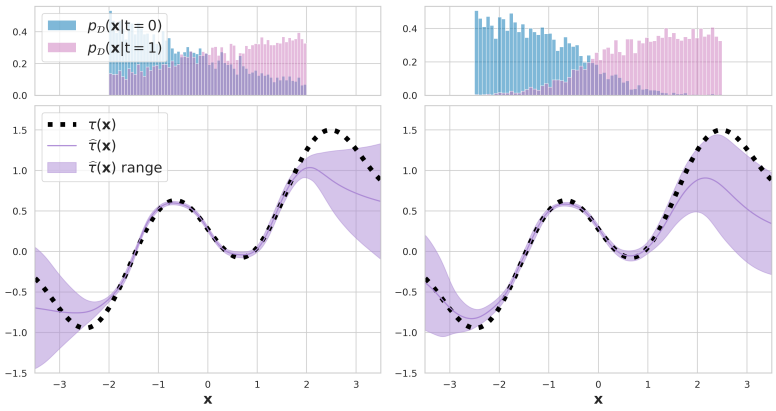
This batch acquisition function can be upper-bounded by scoring each point in $\mathcal{D}_{\text{pool}}$ independently and taking the top b; however, this bound ignores correlations between the samples. In fact, for datasets with significant repetition, this approach can perform worse than random acquisition, and computing the joint mutual information (introduced as *BatchBALD*) rectifies the issue [KVG19].

Estimating the joint mutual information can be computationally expensive, as evaluating the joint entropy over all possible outcomes (for classification) or a covariance matrix over all inputs (for regression) is required. But recent work has made estimation computationally efficient [Hol+23; Kir23] An alternative approach is to use softmax-BALD, which involves importance weighted sampling across $\mathcal{D}_{\text{pool}}$ with the individual importance weights given by BALD [Kir+21]. We use softmax-BALD for batch acquisition because it is computationally more efficient and performs competitively with BatchBALD.

Chapter 4

Scalable Statistical Uncertainty for Causal Machine Learning

In this chapter we will focus on scalable methods for **statistical uncertainty** quantification in conditional causal effect estimation. Statistical uncertainty is informative of when units summarized by measured covariates, $\mathbf{x}$, are not represented in the data, or when they are not represented in a given treatment arm (violations of Assumption 2.4 (positivity)).



(a) representation (b) positivity

Figure 4.1: The purple shaded areas in the lower panes depict regions of ignorance about the response to intervention for units summarized by the covariate value, $\mathbf{X} = \mathbf{x}$. The training data density for the untreated and treated groups are shown in the upper panes. (Figure 5.1a) For ignorance due to measurements $\mathbf{x}$ without **representation** in the observed data, the region should get wider as the distance between $\mathbf{x}$ and the training data increases. (Figure 5.1b) For ignorance without **positivity**, the region should get wider as $P(\mathrm{T} = 0 \mid \mathbf{x})$ or $P(\mathrm{T} = 1 \mid \mathbf{x}) \to 1$.

The work we present in this section is from the following two publications:

1. Andrew Jesson*, Sören Mindermann*, Uri Shalit, and Yarin Gal. "Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models." *NeurIPS*. (2022).

2. Andrew Jesson*, Panagiotis Tigas*, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. "Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data." *NeurIPS*. (2021).

In Section 4.1, we present estimands and scalable estimators of statistical uncertainty for conditional causal effect inference. In Section 4.2, we show how the resulting estimators can be used to define acquisition functions for active learning of conditional causal effects. Finally, in Section 4.3 we present experimental results for several applications. Specifically, in Section 4.3.1, we explore how statistical uncertainty estimators for the CATE function can inform deferral to experts in the setting of automated decision making. We show that our methods are more robust than alternatives to violations of Assumption 2.4 (positivity) and covariate shift. And in Section 4.3.2, we evaluate our proposed active learning acquisition functions fare in the setting of CATE inference when acquiring outcome measurements is expensive. Again, we show that our methods are more robust than previous methods when the positivity assumption is violated in the available data. In both the decision making and active learning settings, we show that our methods scale to complex input modalities like images.

Further publications that the author has contributed to or served and advisory role on during their DPhil on the topic of using statistical uncertainty to learn about causal-effects are:

1. Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. "GeneDisco: A Benchmark for Experimental Design in Drug Discovery." *ICLR*. (2022).

2. Panagiotis Tigas*, Yashas Annadani*, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. "Interventions, Where and How? Experimental Design for Causal Models at Scale." *NeurIPS*. (2022).

3. Clare Lyle*, Arash Mehrjou*, Pascal Notin*, Andrew Jesson, Stefan Bauer, Yarin Gal, Patrick Schwab. "DiscoBAX-Discovery of optimal intervention sets in genomic experiment design." *ICML*. (2023).

4. Yashas Annadani*, Panagiotis Tigas*, Desi R Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, Stefan Bauer. "Differentiable Multi-Target Causal Bayesian Experimental Design." *ICML*. (2023).

4.1 Statistical Uncertainty Estimands and Estimators

In this section we will present estimands and estimators for statistical uncertainty about the CATE, $\delta(\mathbf{x})$, and CAPO, $\mathrm{f_t}(\mathbf{x})$ functions.

4.1.1 Statistical Uncertainty About the Conditional Average Treatment Effect (CATE)

We focus first on estimands and estimators for statistical uncertainty about the CATE function, $\delta(\mathbf{x})$. To define statistical uncertainty about the CATE, we first consider a metaphysical dataset for which we could observe both potential outcomes, $\mathrm{y}_0 \equiv \mathrm{y}(0)$ and $\mathrm{y}_1 \equiv \mathrm{y}(1)$, such that, $\mathcal{D}_{10} = \{\mathbf{x}_i, \mathrm{y}_i(0), \mathrm{y}_i(1)\}_{i=1}^n$. When presenting measures based on mutual information, we will limit our analysis to normally distributed outcomes and normally distributed posterior-predictive distributions. While these assumptions allow for flexible modelling of a broad class of CAPO functions for continuous outcomes under homogeneous response variability (for example, using Gaussian processes), it is restrictive when looking beyond this setting. We will not make these assumptions when presenting measures based on the variance information (Definition 3.1).

Let's start with the hypothetical mutual information between the difference in potential outcomes, $\mathrm{Y}_1 - \mathrm{Y}_0$, and the CATE function, $\Delta(\mathbf{x})$, given our metaphysical dataset, $\mathcal{D}_{10}$.

Lemma 4.1. *Consider a metaphysical dataset, $\mathcal{D}_{10} = \{\mathbf{x}_i, \mathrm{y}_i(0), \mathrm{y}_i(1)\}_{i=1}^n$. Let the difference in potential outcomes be normally distributed as,*

$$\mathrm{Y}_1 - \mathrm{Y}_0 \sim \mathcal{N}(\delta^*(\mathbf{x}), \sigma^2_{\mathrm{y}_{10}}),$$

with posterior-predictive distribution,

$$\int p(\mathrm{y}_1 - \mathrm{y}_0 \mid \delta(\mathbf{x}))dP(\delta(\mathbf{x}) \mid \mathcal{D}_{10}) \sim \mathcal{N}\left(\int \delta(\mathbf{x})dP(\delta(\mathbf{x}) \mid \mathcal{D}_{10}), \sigma^2_{\mathrm{y}_{10}}\right),$$

and known variance, $\sigma^2_{\mathrm{y}_{10}}$*. Then, the hypothetical mutual information between the difference in potential outcomes,* $\mathrm{Y}_1 - \mathrm{Y}_0$*, and the CATE,* $\Delta(\mathbf{x})$ *is,*

$$\mathrm{I}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{10}) = \frac{1}{2} \log \left(1 + \sigma^{-2}_{\mathrm{y}_{10}} \operatorname{Var}(\Delta(\mathbf{x}) \mid \mathcal{D}_{10})\right). \tag{4.1}$$

This follows from Lemma D.5.

This gives an exact expression when the difference in potential outcomes variance, $\sigma^2_{\mathrm{y}_{10}}$, is known. We get the following lower-bound when the variance is unknown.

Lemma 4.2. *Under the same conditions as Lemma 4.1, but with unknown variance,* $\sigma^2_{\mathrm{y}_{10}}$*. Then, the hypothetical mutual information between the difference in potential outcomes,* $\mathrm{Y}_1 - \mathrm{Y}_0$*, and the CATE,* $\Delta(\mathbf{x})$*, is bounded from below by,*

$$\mathrm{I}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{10}) \geq \frac{1}{2} \log \left(1 + \widehat{\overline{\sigma}}^{-2}_{\mathrm{y}_{10}} \operatorname{Var}(\Delta(\mathbf{x}) \mid \mathcal{D}_{10})\right), \tag{4.2}$$

where,

$$\widehat{\overline{\sigma}}^2_{\mathrm{y}_{10}} = \int \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathrm{y}_i(1) - \mathrm{y}_i(0) - \delta(\mathbf{x}_i)\right)^2 dP(\delta(\mathbf{x}) \mid \mathcal{D}_{10}). \tag{4.3}$$

This follows from Lemma D.6 and refines the approximation given in Theorem 1 of our previous work [Jes+21a].

We can also define a realizable mutual information quantity that makes no appeal to metaphysical datasets or hypothetical observations.

Lemma 4.3. *Given an observed dataset,* $\mathcal{D}_n = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}$ *and Assumptions 2.1 to 2.4, the realizable mutual information between a potential outcome,* Y_t*, and the CATE,* $\Delta(\mathbf{x})$*, is*

$$\mathrm{I}(\mathrm{Y}_\mathrm{t}; \Delta \mid \mathbf{x}, \mathcal{D}_n) = \mathrm{H}(\Delta(\mathbf{x}) \mid \mathcal{D}_n) - \int \mathrm{H}(\Delta(\mathbf{x}) \mid \{\mathbf{x}, \mathrm{t}, \mathrm{y}\} \cup \mathcal{D}_n) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_n). \tag{4.4}$$

This refines the foundation for Theorem 3. in our previous work [Jes+21a].

Intuitively, the realizable mutual information quantifies how much we could learn about the CATE function if we were able to observe an instance of the potential outcome, Y_t.

Finally, we define the hypothetical variance information between the difference in potential outcomes, $\mathrm{Y}_1 - \mathrm{Y}_0$, and the CATE function, $\Delta(\mathbf{x})$, given our metaphysical dataset, $\mathcal{D}_{10}$.

Definition 4.1. *CATE variance information [Jes+20]:*

$$\mathrm{V}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{10}) \coloneqq \mathrm{Var}(\Delta(\mathbf{x}) \mid \mathcal{D}_{10}), \tag{4.5}$$

which follows from Lemma D.1 (law of total variance) and Definition 3.1 in Section 3.1.1.

4.1.2 CATE Statistical Uncertainty Estimators

The estimands presented in the preceding section are fundamentally unidentifiable without making cross-world assumptions [RR13]. That is, we cannot observe the dataset, $\mathcal{D}_{10}$, and thus cannot make substantive claims about the variance of the unit-level treatment effect, $\sigma^2_{\mathrm{y}_{10}}$. Even though we are more interested in the variance of the mean function, this does not allow us to escape cross-world assumptions as even the variance of the CATE function depends on an estimate of the unit-level treatment effect variance. This dependence is explicit in the analytic expressions for the CATE variance given by Bayesian linear regression or Gaussian processes. This can be seen in Equations (4.10) and (4.12), where we have to use a substitute term $\widehat{\sigma}^2_{\mathrm{y}}$ to approximate the metaphysical unit-level treatment variance $\sigma^2_{\mathrm{y}_{10}}$. Thus, the following estimators we present may always be biased, even under Assumptions 2.1 to 2.3 (SUDTA, consistency, and ignorability). Alaa and Van Der Schaar [AV18] give a detailed analysis of such bias in the context of Causal Multi-task Gaussian Processes, and we leave analogous analysis for the following estimators for future work. In spite of these concerns, we demonstrate that the following estimators have empirical utility in Section 4.3.

4.1.2.1 Estimating Statistical CATE Uncertainty Using Bayesian Linear Regression, Neural Linear Models, Gaussian Processes, and Deep Kernel GPs

We now define CATE statistical uncertainty estimators for traditional and deep variants of Bayesian linear regression and Gaussian processes. We will call these *kernel estimators*, and distinguish these estimators with the subscript, K.

Definition 4.2. *Under Assumptions 2.1 to 2.3, the kernel estimator for the mutual information between the difference in potential outcomes,* $\mathrm{Y}_1 - \mathrm{Y}_0$*, and the CATE function,* $\Delta(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{I}}_K(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{2} \log \left(1 + \frac{1}{2\widehat{\sigma}^2_{\mathrm{y}}} \left(\widehat{\sigma}^2_0 + \widehat{\sigma}^2_1 - 2\widehat{\sigma}_{10} \right) \right), \tag{4.6}$$

where,

$$\widehat{\overline{\sigma}}_{\mathrm{y}}^2 = \int \frac{1}{n-1} \sum_{i=1}^{n} \Big(\mathrm{t}_i \big(\mathrm{y}_i - f_1(\mathbf{x}_i) \big)^2 + (1 - \mathrm{t}_i) \big(\mathrm{y}_i - f_0(\mathbf{x}_i) \big)^2 \Big) dP(f_{\mathrm{t}}(\mathbf{x}) \mid \mathcal{D}_{\mathrm{t}}). \quad (4.7)$$

Definition 4.3. *Under Assumptions 2.1 to 2.3, the kernel estimator for the mutual information between the potential outcome,* $\mathrm{Y_t}$*, and the CATE function,* $\Delta(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{I}}_K(\mathrm{Y_t}; \Delta \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{2} \log \left(1 + \frac{\widehat{\sigma}_{\mathrm{t}}^2 - 2\widehat{\sigma}_{10}}{\widehat{\sigma}_{1-\mathrm{t}}^2} \right). \quad (4.8)$$

Definition 4.4. *Under Assumptions 2.1 to 2.3, the kernel estimator for the variance information between the difference in potential outcomes,* $\mathrm{Y}_1 - \mathrm{Y}_0$*, and the CATE function,* $\Delta(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{V}}_K(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}) \coloneqq \widehat{\sigma}_0^2 + \widehat{\sigma}_1^2 - 2\widehat{\sigma}_{10}. \quad (4.9)$$

Bayesian Linear Regression and the Neural Linear Model. For Bayesian linear regression given an m-dimensional basis function expansion, $\boldsymbol{\phi}(\mathrm{t}, \mathbf{x})$, or a neural linear model with an m-dimensional penultimate layer feature space, $\boldsymbol{\phi}(\mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$, the variance estimates in Definitions 4.2 and 4.4 are given by:

$$\begin{bmatrix} \widehat{\sigma}_0^2 & \widehat{\sigma}_{10} \\ \widehat{\sigma}_{10} & \widehat{\sigma}_1^2 \end{bmatrix} = \begin{bmatrix} \phi(0, \mathbf{x})^{\mathsf{T}} \\ \phi(1, \mathbf{x})^{\mathsf{T}} \end{bmatrix} \left(\widehat{\overline{\sigma}}_{\mathrm{y}}^{-2} \Phi^{\mathsf{T}} \Phi + \sigma_p^{-2} I_m \right)^{-1} \begin{bmatrix} \phi(0, \mathbf{x})^{\mathsf{T}} \\ \phi(1, \mathbf{x})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}, \quad (4.10)$$

with design matrix,

$$\Phi = \begin{bmatrix} 1 & \phi_1(\mathrm{t}_1, \mathbf{x}_1) & \cdots & \phi_m(\mathrm{t}_1, \mathbf{x}_1) \\ 1 & \phi_1(\mathrm{t}_2, \mathbf{x}_2) & \cdots & \phi_m(\mathrm{t}_2, \mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathrm{t}_n, \mathbf{x}_n) & \cdots & \phi_m(\mathrm{t}_n, \mathbf{x}_n) \end{bmatrix}. \quad (4.11)$$

Gaussian Processes and Deep Kernel GPs. For Gaussian Processes or deep kernel GPs with kernel function $k(\phi_i, \phi_j)$, the variance estimates in Definitions 4.2 and 4.4 are given by:

$$\begin{bmatrix} \widehat{\sigma}_0^2 & \widehat{\sigma}_{10} \\ \widehat{\sigma}_{10} & \widehat{\sigma}_1^2 \end{bmatrix} = K\left(\Phi^*, \Phi^*\right) - K\left(\Phi^*, \Phi\right) \left(K(\Phi, \Phi) + \widehat{\overline{\sigma}}_{\mathrm{y}}^2 I_n \right)^{-1} K\left(\Phi, \Phi^*\right), \quad (4.12)$$

$$\Phi^* = \begin{bmatrix} 0 & \boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}} \\ 1 & \boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}} \end{bmatrix},$$

with design matrix,

$$\Phi = \begin{bmatrix} \mathrm{t}_1 & \boldsymbol{\phi}(\mathbf{x}_1)^\mathsf{T} \\ \mathrm{t}_2 & \boldsymbol{\phi}(\mathbf{x}_2)^\mathsf{T} \\ \vdots & \vdots \\ \mathrm{t}_n & \boldsymbol{\phi}(\mathbf{x}_n)^\mathsf{T} \end{bmatrix}, \quad \boldsymbol{\phi}(\mathbf{x}) = \begin{cases} \mathbf{x} & \text{Gaussian Process} \\ \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}) & \text{Deep Kernel GP} \end{cases}. \tag{4.13}$$

4.1.2.2 Estimating Statistical CATE Uncertainty Using Bayesian Neural Networks and Ensemble Methods.

We now turn to CATE statistical uncertainty estimators for Bayesian neural networks and ensemble methods. We will call these *ensemble estimators*, and distinguish these estimators with the subscript, E.

Definition 4.5. *Under Assumptions 2.1 to 2.3, the ensemble estimator for the mutual information between the difference in potential outcomes, $\mathrm{Y}_1 - \mathrm{Y}_0$, and the CATE function, $\Delta(\mathbf{x})$, given dataset, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$, is defined as:*

$$\widehat{\mathrm{I}}_E(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{2} \log \left(1 + \frac{1}{2(k-1)\widehat{\bar{\sigma}}_{\mathrm{y}}^2} \sum_{i=1}^{k} \Big(\delta(\mathbf{x}, \boldsymbol{\theta}_i) - \bar{\delta}(\mathbf{x}) \Big)^2 \right), \tag{4.14}$$

with the i-th set of parameters drawn from an approximate posterior, $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta} \mid \mathcal{D})$, or a member of an ensemble, $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \mid \mathcal{D}\}$; empirical mean, $\bar{\delta}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^k \delta(\mathbf{x}, \boldsymbol{\theta}_i)$, and CATE function, $\delta(\mathbf{x}, \boldsymbol{\theta}) = f\big(\boldsymbol{\phi}(1, \mathbf{x}, \boldsymbol{\theta})\big) - f\big(\boldsymbol{\phi}(0, \mathbf{x}, \boldsymbol{\theta})\big)$.

Definition 4.6. *Under Assumptions 2.1 to 2.3, the ensemble estimator for the mutual information between the in potential outcome, Y_t, and the CATE function, $\Delta(\mathbf{x})$, given dataset, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$, is defined as:*

$$\widehat{\mathrm{I}}_E(\mathrm{Y}_\mathrm{t}; \Delta \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{2} \log \left(\frac{\sum_{i=1}^k \Big(\delta(\mathbf{x}, \boldsymbol{\theta}_i) - \bar{\delta}(\mathbf{x}) \Big)^2}{\sum_{i=1}^k \Big(f\big(\boldsymbol{\phi}(1-\mathrm{t}, \mathbf{x}, \boldsymbol{\theta}_i)\big) - \bar{f}\big(\boldsymbol{\phi}(1-\mathrm{t}, \mathbf{x})\big) \Big)^2} \right), \tag{4.15}$$

with the i-th set of parameters drawn from an approximate posterior, $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta} \mid \mathcal{D})$, or a member of an ensemble, $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \mid \mathcal{D}\}$; empirical mean, $\bar{\delta}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^k \delta(\mathbf{x}, \boldsymbol{\theta}_i)$, and CATE function, $\delta(\mathbf{x}, \boldsymbol{\theta}) = f\big(\boldsymbol{\phi}(1, \mathbf{x}, \boldsymbol{\theta})\big) - f\big(\boldsymbol{\phi}(0, \mathbf{x}, \boldsymbol{\theta})\big)$.

Definition 4.7. *Under Assumptions 2.1 to 2.3, the ensemble estimator for the variance information between the difference in potential outcomes, $\mathrm{Y}_1 - \mathrm{Y}_0$, and the CATE function, $\Delta(\mathbf{x})$, given dataset, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$, is defined as:*

$$\widehat{\mathrm{V}}_E(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{k-1} \sum_{i=1}^{k} \Big(\delta(\mathbf{x}, \boldsymbol{\theta}_i) - \bar{\delta}(\mathbf{x}) \Big)^2, \tag{4.16}$$

with the same definitions as in Definition 4.5.

4.1.3 Statistical Uncertainty About Conditional Average Potential Outcomes (CAPO)

In this section we will derive estimands and estimators for statistical uncertainty about the CAPO function, $\mathrm{f_t}(\mathbf{x})$. In contrast to statistical uncertainty about the CATE function, we need not make any appeal to metaphysical datasets here. To define statistical uncertainty about the CAPO under Assumptions 2.1 to 2.3 (SUDTA, consistency, ignorability), we consider an observable dataset for which we observe the potential outcome, $\mathrm{y_t} \equiv \mathrm{y(t)}$ for intervention, $\mathrm{T} = \mathrm{t}$, such that, $\mathcal{D}_\mathrm{t} = \{\mathbf{x}_i, \mathrm{y}_i(\mathrm{t})\}_{i=1}^n$. Again, when presenting measures based on mutual information, we will limit our analysis to normally distributed outcomes and normally distributed posterior-predictive distributions, but we will not make this assumption when presenting measures based on the variance information (Definition 3.1).

Let's start with the mutual information between the the potential outcome, $\mathrm{Y_t}$, and the CAPO function, $\mathrm{F_t}(\mathbf{x})$, given the observable dataset, $\mathcal{D}_\mathrm{t}$.

Lemma 4.4. *Consider a dataset, $\mathcal{D}_\mathrm{t} = \{\mathbf{x}_i, \mathrm{y}_i(\mathrm{t})\}_{i=1}^n$. Let the potential outcome be normally distributed as,*

$$\mathrm{Y_t} \sim \mathcal{N}(f_\mathrm{t}^*(\mathbf{x}), \sigma_{\mathrm{y_t}}^2),$$

with posterior-predictive distribution,

$$\int p(\mathrm{y_t} \mid f_\mathrm{t}(\mathbf{x}))dP(f_\mathrm{t}(\mathbf{x}) \mid \mathcal{D}_\mathrm{t}) \sim \mathcal{N}\left(\int f_\mathrm{t}(\mathbf{x})dP(f_\mathrm{t}(\mathbf{x}) \mid \mathcal{D}_\mathrm{t}), \sigma_{\mathrm{y_t}}^2\right),$$

and known variance, $\sigma_{\mathrm{y_t}}^2$. Then, the mutual information between the potential outcome, $\mathrm{Y_t}$, and the CAPO, $\mathrm{F_t}(\mathbf{x})$ is,

$$\mathrm{I}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}_\mathrm{t}) = \frac{1}{2}\log\left(1 + \sigma_{\mathrm{y_t}}^{-2}\,\mathrm{Var}(\mathrm{F_t}(\mathbf{x}) \mid \mathcal{D}_\mathrm{t})\right). \tag{4.17}$$

This follows from Lemma D.5.

This gives an exact expression when the potential outcome variance, $\sigma_{\mathrm{y_t}}^2$, is known. We get the following lower-bound when the variance is unknown.

Lemma 4.5. *Under the same conditions as Lemma 4.4, but with unknown variance, $\sigma_{\mathrm{y_t}}^2$. Then, the hypothetical mutual information between the potential outcome, $\mathrm{Y_t}$, and the CAPO, $\mathrm{F_t}(\mathbf{x})$, is bounded from below by,*

$$\mathrm{I}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}_\mathrm{t}) \geq \frac{1}{2}\log\left(1 + \widehat{\overline{\sigma}}_{\mathrm{y_{y_t}}}^{-2}\,\mathrm{Var}(\mathrm{F_t}(\mathbf{x}) \mid \mathcal{D}_\mathrm{t})\right), \tag{4.18}$$

where,

$$\widehat{\overline{\sigma}}^2_{\mathrm{y_t}} = \int \frac{1}{n-1} \sum_{i=1}^{n} \big(\mathrm{y}_i(\mathrm{t}) - f_{\mathrm{t}}(\mathbf{x}_i)\big)^2 dP(f_{\mathrm{t}}(\mathbf{x}) \mid \mathcal{D}_{\mathrm{t}}). \tag{4.19}$$

This follows from Lemma D.6 and refines the approximation given in Theorem 2 of our previous work [Jes+21a].

We now define the variance information between the potential outcomes, $\mathrm{Y_t}$, and the CAPO function, $\mathrm{F_t}(\mathbf{x})$, given the dataset, $\mathcal{D}_{\mathrm{t}}$.

Definition 4.8. *CAPO variance information:*

$$\mathrm{V}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}_{\mathrm{t}}) \coloneqq \mathrm{Var}(\mathrm{F_t}(\mathbf{x}) \mid \mathcal{D}_{\mathrm{t}}), \tag{4.20}$$

which follows from Lemma D.1 (law of total variance) and Definition 3.1 [Jes+20].

Again, in contrast to the estimands presented for statistical uncertainty, all quantities in the CAPO statistical uncertainty estimands are identifiable under Assumptions 2.1 to 2.3 (SUDTA, consistency, and ignorability).

4.1.4 CAPO Statistical Uncertainty Estimators

In contrast to the CATE estimands, the CAPO statistical uncertainty estimands are *not* metaphysical because we can observe the dataset, $\mathcal{D}_t$, and thus we can make substantive claims about the variance, $\sigma^2_{\mathrm{y}_t}$. Thus, only the ensemble mutual information estimators are biased because of the expectation inside the non-linear log function.

4.1.4.1 Estimating Statistical CAPO Uncertainty Using Bayesian Linear Regression, Neural Linear Models, Gaussian Processes, and Deep Kernel GPs

We now define CAPO statistical uncertainty estimators for traditional and deep variants of Bayesian linear regression and Gaussian processes. We will again call these *kernel estimators*, and distinguish these estimators with the subscript, K.

Definition 4.9. *Under Assumptions 2.1 to 2.3, the kernel estimator for the mutual information between the potential outcome,* $\mathrm{Y_t}$*, and the CATE function,* $\mathrm{F_t}(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{I}}_K(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{2} \log\left(1 + \frac{\widehat{\sigma}^2_{\mathrm{t}}}{\widehat{\overline{\sigma}}^2_{\mathrm{y_t}}}\right). \tag{4.21}$$

Definition 4.10. *Under Assumptions 2.1 to 2.3, the kernel estimator for the variance information between the potential outcome,* $\mathrm{Y_t}$*, and the CAPO function,* $\mathrm{F_t}(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{V}}_K(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}) \coloneqq \widehat{\sigma}_{\mathrm{t}}^2. \tag{4.22}$$

The variance estimators, $\widehat{\sigma}_{\mathrm{t}}^2$, are defined as before in either Equation (4.10) or Equation (4.12).

4.1.4.2 Estimating Statistical CAPO Uncertainty Using Bayesian Neural Network and Ensemble Methods.

We now turn to CAPO statistical uncertainty estimators for Bayesian neural networks and ensemble methods. We will again call these *ensemble estimators* and distinguish these estimators with the subscript, E.

Definition 4.11. *Under Assumptions 2.1 to 2.3, the ensemble estimator for the mutual information between the potential outcome,* $\mathrm{Y_t}$*, and the CAPO function,* $\mathrm{F_t}(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{I}}_E(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{2} \log \left(1 + \frac{1}{(k-1)\widehat{\sigma}_{\mathrm{y}}^2} \sum_{i=1}^{k} \Big(f\big(\boldsymbol{\phi}(\mathrm{t}, \mathbf{x}, \boldsymbol{\theta}_i)\big) - \bar{f}\big(\boldsymbol{\phi}(\mathrm{t}, \mathbf{x})\big) \Big)^2 \right), \tag{4.23}$$

with empirical mean, $\bar{f}\big(\boldsymbol{\phi}(\mathrm{t}, \mathbf{x})\big) = \frac{1}{k} \sum_{i=1}^{k} f\big(\boldsymbol{\phi}(\mathrm{t}, \mathbf{x}, \boldsymbol{\theta}_i)\big)$.

Definition 4.12. *Under Assumptions 2.1 to 2.3, the ensemble estimator for the variance information between the potential outcome,* $\mathrm{Y_t}$*, and the CAPO function,* $\mathrm{F_t}(\mathbf{x})$*, given dataset,* $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$*, is defined as:*

$$\widehat{\mathrm{V}}_E(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}) \coloneqq \frac{1}{k-1} \sum_{i=1}^{k} \Big(f\big(\boldsymbol{\phi}(\mathrm{t}, \mathbf{x}, \boldsymbol{\theta}_i)\big) - \bar{f}\big(\boldsymbol{\phi}(\mathrm{t}, \mathbf{x})\big) \Big)^2, \tag{4.24}$$

with the same definitions as in Definition 4.5.

4.2 Active Learning for Conditional Causal-Effects

In this section we will present and refine the active learning acquisition functions we first published in Jesson et al. [Jes+21a]. The acquisition functions will be defined in terms of the scalable mutual information and variance information estimators we have just presented in Section 4.1.

By including the treatment, we depart from the standard active learning setting. For active learning of treatment effects, we define $\mathcal{D}_{\text{pool}} = \{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^{n_{\text{pool}}}$, a labeled training set $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^{n_{\text{train}}}$, and a predictive model with likelihood $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{train}})$. The acquisition function takes as input $\mathcal{D}_{\text{pool}}$ and returns a batch of data $\{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^{b}$ which are labelled using an oracle and added to $\mathcal{D}_{\text{train}}$.

4.2.1 τ-Acquisition Functions

The first set of acquisition functions we will discuss are based on the mutual information, $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D})$, and variance information, $\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D})$, between the difference in potential outcomes, $\mathrm{Y}_1 - \mathrm{Y}_0$, and CATE, $\Delta(\mathbf{x})$, given covariate value, $\mathbf{x}$, and dataset, $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^{n_{\text{train}}}$. Following the notation we established in [Jes+21a], we call this class "τ–acquisition functions."

Definition 4.13. *τ mutual information acquisition,*

$$\widehat{\mathrm{I}}(\tau; \mathbf{x}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}}), \tag{4.25}$$

where $\widehat{\mathrm{I}}_{(\cdot)}$ represents one of the estimators defined in Definitions 4.2 and 4.5.

Definition 4.14. *τ variance information acquisition,*

$$\widehat{\mathrm{V}}(\tau; \mathbf{x}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}}), \tag{4.26}$$

where $\widehat{\mathrm{V}}_{(\cdot)}$ represents one of the estimators defined in Definitions 4.4 and 4.7. Note that this acquisition under the power-BALD [Kir+21] sampling strategy is equivalent to what we called τ-BALD in [Jes+21a] under softmax-BALD [Kir+21] sampling strategy.

Intuition. We can interpret the mutual information estimator in Definition 4.13 as, *the information we would gain about the CATE function, $\Delta(\mathbf{x})$, if we could observe the corresponding difference in potential outcomes, $\mathrm{Y}_1 - \mathrm{Y}_0$, for a given covariate realization $\mathbf{x}$*. While, this can be a useful measure of uncertainty about the CATE, it should be clear why it does not make sense as an acquisition function. Namely, we do not get to see the difference in potential outcomes, just the potential outcome, Y_{t}, for the intervention applied, $\mathrm{T} = \mathrm{t}$, given the boundary conditions summarized by, $\mathbf{x}$. This issue is not absolved by the variance information formulation, for analogous reasons.

4.2.2 μ-Acquisition Functions

The second set of acquisition functions we will discuss are based on the mutual information, $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D})$, and variance information, $\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D})$, between the potential outcome, $\mathrm{Y_t}$, and CAPO, $\mathrm{F}_t(\mathbf{x})$, given covariate value, $\mathbf{x}$, intervention level, t, and dataset, $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^{n_{\text{train}}}$. Following the notation we established in [Jes+21a], we call this class "μ–acquisition functions."

Definition 4.15. *μ mutual information acquisition,*

$$\widehat{\mathrm{I}}(\mu; \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}_{\text{train}}), \tag{4.27}$$

where $\widehat{\mathrm{I}}_{(\cdot)}$ represents one of the estimators defined in Definitions 4.9 and 4.11.

Definition 4.16. *μ variance information acquisition,*

$$\widehat{\mathrm{V}}(\mu; \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}_{\text{train}}), \tag{4.28}$$

where $\widehat{\mathrm{V}}_{(\cdot)}$ represents one of the estimators defined in Definitions 4.10 and 4.12. Note that this acquisition under the power-BALD [Kir+21] sampling strategy is equivalent to what we called μ-BALD in [Jes+21a] under softmax-BALD [Kir+21] sampling strategy.

Intuition. In contrast to the τ–acquisition functions, the μ–acquisition functions are much more sensible. We can interpret the estimator, $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y_t}; \mathrm{F_t} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$, as *the information we would gain about the CAPO function, $\mathrm{F_t}(\mathbf{x})$, if we could observe the corresponding potential outcome, $\mathrm{Y_t}$, for a given covariate realization, $\mathbf{x}$, under intervention,* t. Under Assumptions 2.1 to 2.3 (SUDTA, consistency, and ignorability), all of these quantities are either observable or identifiable, so we do not run into the same issues.

4.2.3 ρ-Acquisition Functions

The third set of acquisition functions we will discuss are based on either a difference between mutual information estimates or a ratio of variance information estimates. Following the notation we established in [Jes+21a], we call this class "ρ–acquisition functions."

In the mutual information formulation, we look at the difference between the CATE mutual information and the "counterfactual" CAPO mutual information estimates.

Definition 4.17. *ρ mutual information acquisition,*

$$\widehat{\mathrm{I}}(\rho; \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}}) - \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_{1-\mathrm{t}}; \mathrm{F}_{1-\mathrm{t}} \mid \mathbf{x}, \mathcal{D}_{\text{train}}), \tag{4.29}$$

where $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.2 and 4.5 and $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_{1-\mathrm{t}}; \mathrm{F}_{1-\mathrm{t}} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.9 and 4.11.

In the variance information formulation, we look at the ratio between the CATE variance information and the "counterfactual" CAPO variance information estimates.

Definition 4.18. *ρ variance information acquisition,*

$$\widehat{\mathrm{V}}(\rho; \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \coloneqq \frac{\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}})}{\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_{1-\mathrm{t}}; \mathrm{F}_{1-\mathrm{t}} \mid \mathbf{x}, \mathcal{D}_{\text{train}})}, \tag{4.30}$$

where $\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.4 and 4.7 and $\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_{1-\mathrm{t}}; \mathrm{F}_{1-\mathrm{t}} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.10 and 4.12. Note that this acquisition under the power-BALD [Kir+21] sampling strategy is equivalent to what we called ρ-BALD in [Jes+21a] under softmax-BALD [Kir+21] sampling strategy.

Intuition. While the μ–acquisition functions consider only reducible uncertainty for observable or identifiable quantities, they *do not* take into account our current knowledge about the counterfactual. The ρ–acquisition functions seek to address this consideration. In both the mutual and variance information formulations we look at the problematic CATE uncertainty, and either subtract or divide the portion of that information that we cannot reduce under a factual intervention, $\mathrm{T} = \mathrm{t}$. These can be a different quantities from their μ-counterparts if the covariance estimate, $\widehat{\sigma}_{10}$, of the CATE function, $\Delta(\mathbf{x})$, is non-zero, and can account for ways in which knowledge about the response surface under one intervention value can inform the response surface under other intervention values. This motivates our use of the S-Learner structure.

4.2.4 $\mu\rho$-Acquisition Functions

The fourth and final set of acquisition functions we will discuss are based on either a sum of mutual information estimates or a product of variance information estimates. Following the notation we established in [Jes+21a], we call this class "$\mu\rho$–acquisition functions."

In the mutual information formulation, we look at the sum between the CATE mutual information and difference between the "factual" and the "counterfactual" CAPO mutual information estimates.

Definition 4.19. *$\mu\rho$ mutual information acquisition,*

$$\begin{aligned}\widehat{\mathrm{I}}(\mu\rho; \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}}), \\ + \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_\mathrm{t}; \mathrm{F}_\mathrm{t} \mid \mathbf{x}, \mathcal{D}_{\text{train}}) - \widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_{1-\mathrm{t}}; \mathrm{F}_{1-\mathrm{t}} \mid \mathbf{x}, \mathcal{D}_{\text{train}}),\end{aligned} \tag{4.31}$$

where $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.2 and 4.5 and $\widehat{\mathrm{I}}_{(\cdot)}(\mathrm{Y}_{(\cdot)}; \mathrm{F}_{(\cdot)} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.9 and 4.11.

In the variance information formulation, we look at the product between the CATE variance information and the ratio of the "factual" and "counterfactual" CAPO variance information estimates.

Definition 4.20. *$\mu\rho$ variance information acquisition,*

$$\widehat{\mathrm{V}}(\mu\rho; \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \coloneqq \widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}}) \frac{\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_\mathrm{t}; \mathrm{F}_\mathrm{t} \mid \mathbf{x}, \mathcal{D}_{\text{train}})}{\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_{1-\mathrm{t}}; \mathrm{F}_{1-\mathrm{t}} \mid \mathbf{x}, \mathcal{D}_{\text{train}})}, \tag{4.32}$$

where $\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_1 - \mathrm{Y}_0; \Delta \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.4 and 4.7 and $\widehat{\mathrm{V}}_{(\cdot)}(\mathrm{Y}_{(\cdot)}; \mathrm{F}_{(\cdot)} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$ represents one of the estimators defined in Definitions 4.10 and 4.12. Note that this acquisition under the power-BALD [Kir+21] sampling strategy is equivalent to what we called $\mu\rho$-BALD in [Jes+21a] under softmax-BALD [Kir+21] sampling strategy.

Intuition. There is a potential shortcoming of ρ–acquisitions that may lead to suboptimal data efficiency. Consider two examples in an observational pool dataset, $\mathcal{D}_{\text{pool}} = \{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^{n_{\text{pool}}}$, consisting of covariates and assigned interventions, $(\mathbf{x}_1, \mathrm{t}_1)$ and $(\mathbf{x}_2, \mathrm{t}_2)$, where

$$\operatorname*{Var}_{\boldsymbol{\theta}}(f(\mathbf{x}_1, \mathrm{t}_1, \boldsymbol{\theta})) = \operatorname*{Var}_{\boldsymbol{\theta}}(f(\mathbf{x}_1, (1 - \mathrm{t}_1), \boldsymbol{\theta})),$$

and

$$\operatorname*{Var}_{\boldsymbol{\theta}}(f(\mathbf{x}_2, \mathrm{t}_2, \boldsymbol{\theta})) = \operatorname*{Var}_{\boldsymbol{\theta}}(f(\mathbf{x}_2, (1 - \mathrm{t}_2), \boldsymbol{\theta})),$$

for each point, we are as uncertain about the conditional expectation given the factual treatment as we are uncertain given the counterfactual treatment. Further, let $\mathrm{Cov}_{\boldsymbol{\theta}}(f(\mathbf{x}_1, \mathrm{t}_1, \boldsymbol{\theta}), f(\mathbf{x}_1, (1-\mathrm{t}_1), \boldsymbol{\theta})) = \mathrm{Cov}_{\boldsymbol{\theta}}(f(\mathbf{x}_2, \mathrm{t}_2, \boldsymbol{\theta}), f(\mathbf{x}_2, (1-\mathrm{t}_2), \boldsymbol{\theta}))$. And finally, let

$$\operatorname*{Var}_{\boldsymbol{\theta}}(f(\mathbf{x}_1, \mathrm{t}_1, \boldsymbol{\theta})) > \mathrm{Var}(f_{\boldsymbol{\theta}}(\mathbf{x}_2, \mathrm{t}_2, \boldsymbol{\theta})).$$

That is, we are more uncertain about the conditional expectation given the factual treatment for data point $(\mathbf{x}_1, \mathrm{t}_1)$ than we are for data point $(\mathbf{x}_2, \mathrm{t}_2)$. Under these three conditions, ρ–BALD would rank these two points equally, and so this method would bias training data to the modes of $\mathcal{D}_{\text{pool}}$ when $(\mathbf{x}_2, \mathrm{t}_2)$ is more frequent than $(\mathbf{x}_1, \mathrm{t}_1)$. In practice, it may be more data-efficient to choose $(\mathbf{x}_1, \mathrm{t}_1)$ over $(\mathbf{x}_2, \mathrm{t}_2)$ as it would more likely be a point as yet unseen by the model. The $\mu\rho$–acquisitions are proposed to combine the positive attributes of μ-and ρ- acquisitions, while mitigating their shortcomings.

4.3 Applications

4.3.1 Deferral of Recommendations to An Expert

If there is insufficient knowledge about an individual, and a high cost associated with making errors, it may be preferable to defer the intervention recommendations to an expert or safe policy. It is therefore important to have an informed *deferral policy*. In our experiments, we defer, i.e. choose to make no intervention recommendation, when the epistemic uncertainty exceeds a certain threshold. In general, setting the threshold will be a domain-specific problem that depends on the cost of type I (incorrectly recommending treatment) and type II (incorrectly withholding treatment) errors. In the diagnostic setting, thresholds have been set to satisfy public health authority specifications, e.g. for diabetic retinopathy [Lei+17]. Some rejection methods additionally weigh the chance of algorithmic error against that of human error [Rag+19].

When using the propensity score for deferral, a simple policy is to specify η_0 and defer recommendation for points that do not satisfy Assumption 2.4 with $\eta = \eta_0$. Caliendo and Kopeinig [CK08] propose more sophisticated standard guidelines. These methods only account for the uncertainty about the CATE, $\delta(\mathbf{x})$, that is due to limited overlap and do not consider that uncertainty is also modulated by the availability of data on similar individuals (as well as the noise in this data).

We introduce two deferral policies based on the epistemic and predictive uncertainty estimates of an uncertainty aware CATE estimator. Both policies opt to defer if the uncertainty estimate is greater than a threshold that defers for a given proportion of the training data r_{defer}. The training data is used since there may not be a large enough test set in practice. For all policies, we determine thresholds on the training set to simulate a real-world individual-level recommendation scenario. The *epistemic*

uncertainty policy uses a sample-based estimator of the uncertainty in CATE using the variance information estimator given in Definition 4.7, where M Monte Carlo samples are taken from each of $q(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1|\mathcal{D})$. Note that, for the T-Learner, this posterior factorizes into two independent distributions $q(\boldsymbol{\theta}_0|\mathcal{D}), q(\boldsymbol{\theta}_1|\mathcal{D})$ because there are separate models for the outcome given treatment and no treatment. Furthermore, other models share parameters for $\mu, (\cdot; \boldsymbol{\theta}_0), \mu(\cdot; \boldsymbol{\theta}_1)$ so the individual parameters in $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ may overlap. The *predictive* uncertainty policy uses an estimator of of the total variance, $\widehat{\text{Var}}_{pred}[\text{Y}^1 - \text{Y}^0|\mathbf{x}_i]$, which has the same functional form as in Definition 4.7, but instead of being over the difference in expected values $f^{\hat{\boldsymbol{\theta}}^{\text{t}}_j}(\mathbf{x}_i)$ of the output distribution it is over samples $\text{y}^{\hat{\boldsymbol{\theta}}^{\text{t}}_j}(\mathbf{x}_i)$ of the output distribution.

We compare the utility of these policies to a random rejection baseline and two policies based on propensity scores. The first propensity policy (*propensity quantiles*) finds a two sided threshold on the distribution of estimated propensity scores such that a proportion $(1. - r_{\text{reject}})$ of the training data is retained. The second policy (*propensity trimming*) implements a trimming algorithm following the guidelines proposed by Caliendo and Kopeinig [CK08].

4.3.1.1 Experiments

In this section, we summarize empirical evidence from Jesson et al. [Jes+20] for the following claims: that our uncertainty aware methods are robust both to violations of the overlap assumption and a failure mode of propensity based trimming (Section 4.3.1.2); that they indicate high uncertainty when covariate shifts occur between training and test distributions (Section 4.3.1.3); and that they yield lower CATE errors while rejecting fewer points than propensity based trimming (Section 4.3.1.3). In Jesson et al. [Jes+20], we introduce a new, high-dimensional, individual-level causal effect prediction benchmark dataset called CEMNIST to demonstrate robustness to overlap and propensity failure (Section 4.3.1.2). Finally, we introduce a modification to the IHDP causal inference benchmark to explore covariate shift (4.3.1.3).

We evaluate our methods by considering *intervention recommendations*. A simple recommendation strategy is to intervene with treatment, $\text{T} = 1$, if the predicted CATE, $\widetilde{\delta}(\mathbf{x}; \boldsymbol{\theta})$, is positive, or intervene with treatment, $\text{T} = 0$, if negative. However, as stated in section 4.3.1, insufficient knowledge about an individual and high costs due to error necessitate informed *deferral policies* to formalize when a recommendation should be withheld. We compare four rejection policies: *epistemic uncertainty* using $\text{Var}\left(\widetilde{\delta}(\mathbf{x}; \boldsymbol{\Theta})\right)$, *propensity quantiles*, *propensity trimming* [CK08] and *random*

(implementation details of each policy are given in Appendix C.1). Policies are ranked according to the proportion of incorrect recommendations made, given a fixed recommendation deferral rate (r_{def}). This corresponds to assigning a cost of 1 to making an incorrect prediction and a cost of 0 for either making a correct recommendation or withholding an automated recommendation and deferring the decision to a human expert instead. We also report the Precision in Estimation of Heterogeneous Treatment Effect (PEHE) [Hil11; SJS17] over the non-rejected subset. The mean and standard error of each metric is reported over a dataset-dependent number of training runs.

We evaluate and compare each deferral policy using several uncertainty aware CATE estimators. The estimators are Bayesian versions of CEVAE [Lou+17], TARNet, CFR-MMD [SJS17], Dragonnet [SBV19], and a deep T-Learner [Kün+19]. Each model is augmented by introducing Bayesian parameter uncertainty and by predicting a distribution over model outputs. For imaging experiments, a two-layer CNN encoder is added to each model. Details for each model are given in Appendix C.1.1. In the result tables, each model's name is prefixed with a "B" for "Bayesian". We also compare to Bayesian Additive Regression Trees (BART) [Hil11].

4.3.1.2 Using Uncertainty When Overlap is Violated

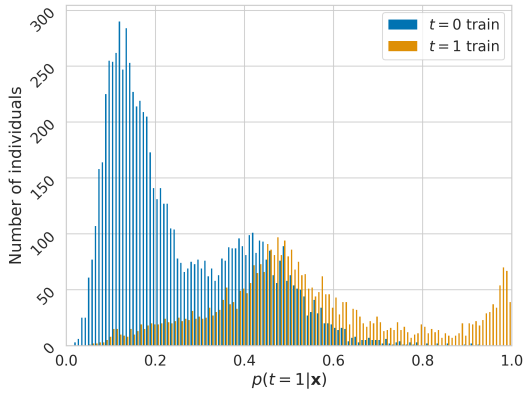
Causal Effect MNIST (CEMNIST). We introduce the CEMNIST dataset in Jesson et al. [Jes+20] using hand-written digits from the MNIST dataset [LeC+98; Den12] to demonstrate that our uncertainty measures capture non-overlap on high-dimensional data and that they are robust to a failure mode of propensity score rejection.

Table 4.1: **CEMNIST-Overlap** Description of "Causal effect MNIST" dataset.

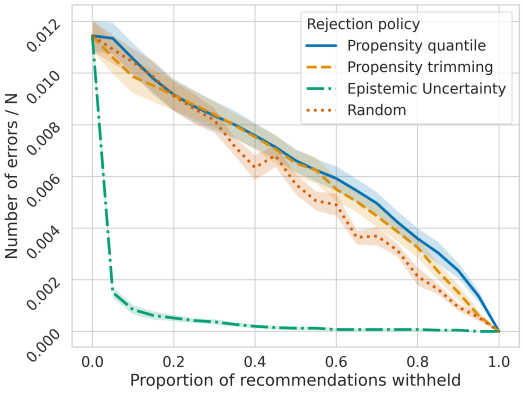
Digit(s)	$p(\mathbf{x})$	$p(t=1\|\mathbf{x})$	$p(y=1\|\mathbf{x},t=0)$	$p(y=1\|\mathbf{x},t=1)$	CATE
9	0.5	1/9	1	0	-1
2	0.5/9	1	0	1	1
other odds	0.5/9	0.5	1	0	-1
other evens	0.5/9	0.5	0	1	1

Table 4.1 depicts the data generating process for CEMNIST. In expectation, half of the samples in a generated dataset will be nines, and even though the propensity for treating a nine is relatively low, there are still on average twice as many treated nines as there are samples of other treated digits (except for twos). Therefore, it is reasonable to expect that the CATE can be estimated most accurately for nines.

For twos, there is strict non-overlap. Therefore, the CATE cannot be estimated accurately. For the remaining digits, the CATE estimate should be less confident than for nines because there are fewer examples during training, but more confident than for twos because there are both treated and untreated training examples.



(a) Propensity histogram



(b) Error rate vs. r_{rej}

$\sqrt{\epsilon_{PEHE}}$	**CEMNIST**($r_{\text{rej}} = 0.5$)		
Method / *Pol.*	***rand.***	***prop.***	***unct.***
BART	2.1±.0	2.1±.0	**2.0±.0**
BT-Learner	0.3±.0	0.2±.0	**0.0±.0**
BTARNet	0.2±.0	0.2±.0	**0.0±.0**
BCFR-MMD	0.3±.0	0.3±.0	**0.1±.0**
BDragonnet	0.2±.0	0.2±.0	**0.0±.0**
BCEVAE	0.3±.0	0.2±.0	**0.0±.0**

(c) Model comparison

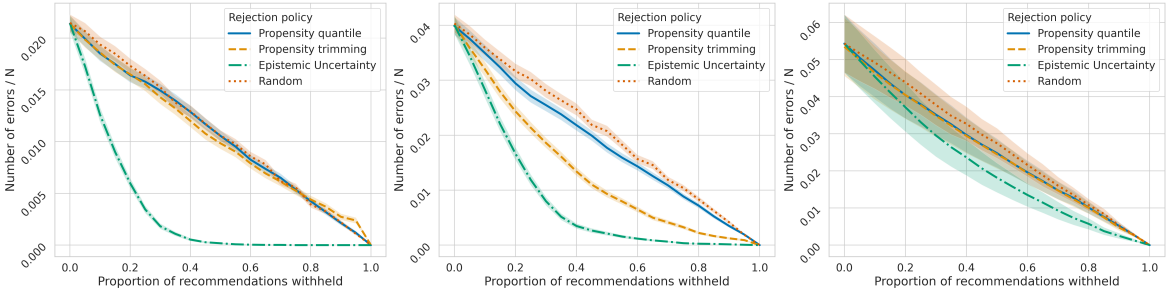
Figure 4.2: **CEMNIST** evaluation. (a) Histogram of estimated propensity scores. Untreated nines account for the peaks on the left side. (b) Error rate for different rejection policies as we vary the rejection rate. (c) $\sqrt{\epsilon_{PEHE}}$ for different models at a fixed rejection rate $r_{\text{rej}} = 0.5$. Compared are the policies *random*, *propensity trimming*, and *epistemic uncertainty*.

This experimental setup is chosen to demonstrate where the *propensity* based rejection policies can be inappropriate for the prediction of individual-level causal effects. Figure 4.2a shows the histogram over training set predictions for a deep propensity model on a realization of the CEMNIST dataset. A data scientist following the trimming paradigm [CK08] would be justified in choosing a lower threshold around 0.05 and an upper threshold around 0.75. The upper threshold would properly reject twos, but the lower threshold would start rejecting nines, which represent the population that the CATE estimator can be most confident about. Therefore, rejection choices can be worse than random.

Figure 4.2b shows that the recommendation-error-rate is significantly lower for the *epistemic uncertainty* policy (green dash-dot) than for both the *random* baseline policy (red dot) and the *propensity* based policies (orange dash and blue solid). BT-Learner is used for this plot. These results hold across a range of other SOTA CATE estimators for the $\sqrt{\epsilon_{PEHE}}$, as shown in Figure 4.2c. Details on the protocol generating these results are in Appendix B.4.

4.3.1.3 Uncertainty Under Covariate Shift

Infant Health and Development Program (IHDP). When deploying a machine learning system, we must often deal with a test distribution of $\mathbf{x}$ which is different from the training distribution $p(\mathbf{x})$. We induce a covariate shift in the semi-synthetic dataset IHDP [Hil11; SJS17] by excluding instances *from the training* set for which the mother is unmarried. Mother's marital status is chosen because it has a balanced frequency of 0.52 ± 0.00; furthermore, it has a mild association with the treatment as indicated by a log odds ratio of 2.22 ± 0.01; and most importantly, there is evidence of a simple distribution shift, indicated by a predictive accuracy of 0.75 ± 0.00 for marital status using a logistic regression model over the remaining covariates. We comment on the ethical implications of this experimental set-up, describe IHDP, and explain the experimental protocol in Appendix B.2.



(a) IHDP (b) IHDP Covariate Shift (c) ACIC 2016

Figure 4.3: Uncertainty based rejection policies yield significantly lower error rates while withholding fewer recommendations than propensity policies, on IHDP, IHDP Cov., and ACIC 2016.

We report the mean and standard error in recommendation-error-rates and $\sqrt{\epsilon_{PEHE}}$ over 1000 realizations of the IHDP Covariate-Shift dataset to evaluate each policy by computing each metric over the test set (both sub-populations included). We sweep r_{def} from 0.0 to 1.0 in increments of 0.05. Figure 4.3b shows, for the BT-Learner, that the *epistemic uncertainty* (green dash-dot) policy significantly outperforms the

uncertainty-oblivious policies across the whole range of rejection rates. The right hand section of Table 4.2 supports this claim by reporting the $\sqrt{\epsilon_{PEHE}}$ for each model at $r_{\text{def}} = 0.5$; the approximate frequency of the excluded population. Every model class shows improved rejection performance. However, comparisons between model classes are not necessarily appropriate since some models target different scenarios, for example, CEVAE targets *non*-synthetic data where confounders z aren't directly observed, and it is known to underperform on IHDP [Lou+17].

Table 4.2: Comparing *epistemic uncertainty*, *propensity trimming*, and *random* rejection policies for IHDP and IHDP Covariate Shift datasets with uncertainty-equipped SOTA models. Errors are reported on the remaining test-set recommendations after 50% or 10% of examples set to be rejected. *Epistemic uncertainty* policy leads to the lowest errors in CATE estimates (in bold).

$\sqrt{\epsilon_{PEHE}}$	**IHDP** ($r_{\text{rej}} = 0.1$)			**Cov. Shift** ($r_{\text{rej}} = 0.5$)		
Method / *Pol.*	***rand.***	***prop.***	***unct.***	***rand.***	***prop.***	***unct.***
BART	1.9±.2	1.9±.2	**1.6±.1**	2.6±.2	2.7±.3	**1.8±.2**
BT-Learner	1.0±.0	0.9±.0	**0.7±.0**	2.3±.2	2.3±.2	**1.3±.1**
BTARNet	1.1±.0	1.0±.0	**0.8±.0**	2.2±.3	2.0±.3	**1.2±.1**
BCFR-MMD	1.3±.1	1.3±.1	**0.9±.0**	2.5±.2	2.4±.3	**1.7±.2**
BDragonnet	1.5±.1	1.4±.1	**1.1±.0**	2.4±.3	2.2±.3	**1.3±.2**
BCEVAE	1.8±.1	1.9±.1	**1.5±.1**	2.5±.2	2.4±.3	**1.7±.1**

We report results for the unaltered IHDP dataset in Figure 4.3a and the l.h.s. of Table 4.2. This supports that uncertainty rejection is more *data-efficient*, i.e., errors are lower while rejecting less. This is further supported by the results on ACIC 2016 [Dor+19] (Figure 4.3c and Table 4.3). The preceding results can be reproduced using publicly available code[1].

4.3.2 Causal Active Learning of Conditional Average Treatment Effects

A problem in scalable machine learning is data efficiency. While modern methods are capable of impressive performance, they need a significant amount of labeled data. Acquiring labeled data can be expensive, requiring specialist knowledge or an invasive procedure to determine the outcome. Therefore, it is desirable to minimize the amount of labeled data needed to obtain a well-performing model. Active learning provides a principled framework to address this concern [CGJ96]. In active learning for treatment effects [DPM11; Sun+19; QWZ21] a model is trained on available

[1]Available at: `https://github.com/OATML/ucate`

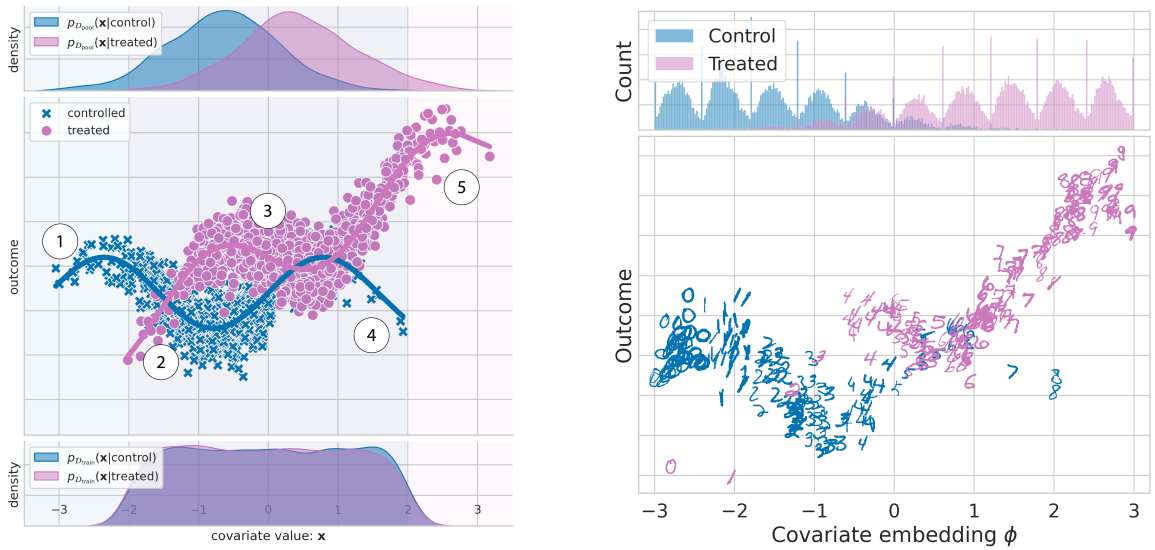
Table 4.3: Comparing *epistemic uncertainty*, *propensity trimming*, and *random* rejection policies for ACIC 2016 dataset with uncertainty-equipped SOTA models. Errors are reported on the remaining test-set recommendations after 10% of examples set to be rejected. *Epistemic uncertainty* policy leads to the lowest errors in CATE estimates (in bold).

$\sqrt{\epsilon_{PEHE}}$	**ACIC 2016** ($r_{\text{rej}} = 0.1$)		
Method / *Pol.*	***rand.***	***prop.***	***unct.***
BART	1.3±.1	1.2±.1	**0.9±.1**
BT-Learner	2.1±.1	2.0±.1	**1.5±.1**
BTARNet	1.8±.1	1.7±.1	**1.2±.1**
BCFR-MMD	2.3±.2	2.1±.1	**1.7±.1**
BDragonnet	1.9±.1	1.8±.1	**1.3±.1**
BCEVAE	3.3±.2	3.2±.2	**2.9±.1**

labeled data consisting of covariates, assigned treatments, and acquired outcomes. The model predictions decide the most informative examples from data comprised of only covariates and treatment indicators. Outcomes are acquired, e.g., by performing a biopsy for the selected patients, and the model is retrained and evaluated. This process repeats until either a satisfactory performance level is achieved or the labeling budget is exhausted.

At first sight, this might seem simple; however, active learning induces biases that result in divergence between the distribution of the acquired training data and the distribution of the pool set data [FGR21]. In the context of learning causal effects, such bias can have both positive and negative consequences. For example, while random acquisition active learning results in an unbiased sample of the training data, we demonstrate how it can lead to over-allocation of resources to the mode of the data at the expense of learning about underrepresented data. Conversely, while biasing acquisitions toward lower density regions of the pool data can be desirable, it can also lead to outcome acquisition for data with unidentifiable treatment effects, which leads to uninformed, potentially harmful, personalized recommendations.

To see how training data bias can benefit treatment effect estimation, consider one difference between experimental and observational data: the treatment assignment mechanism is unavailable for observational data. In observational data, variables that affect treatment assignment (an untestable condition) may be unobserved. Moreover, the relative proportion of treated to controlled individuals varies across different subpopulations of the data. Fig. 4.4a illustrates the latter point, where there are relatively equal proportions of treated and controlled examples for data in region 3. However,



(a) Simulated

(b) CMNIST

Figure 4.4: Left: Observational data (Figure 4.4a). Top: data density of treatment (right) and control (left) groups. Middle: observed outcome response for treatment (circles) and control (x's) groups. Bottom: data density for active learned training set after a number of acquisition steps. Right: Visualizing CMNIST dataset (Fig. 4.4b). Model inputs are MNIST digits and assigned treatments. The MNIST digits are high-dimensional proxies for the latent confounding covariate ϕ. Digits are projected onto ϕ by ordering them first by image intensity and then by digit class (0 - 9). Methods must be able to implicitly learn this non-linear mapping in order to predict the conditional expected outcomes.

the proportions become less balanced as we move to either the left or the right. In extreme cases, say if a group described by some covariate values were systematically excluded from treatment, the treatment effect for that group *cannot be known* [Pet+12]. Fig. 4.4a illustrates this in region 1, where only controlled examples reside, and in region 5, where only treated cases occur. In the language of causal inference, the necessity of seeing both treated and untreated examples for each sub-population corresponds to satisfying the overlap (or positivity) Assumption 2.4. The data available in the pool set limits overlap when treatments cannot be assigned. In this setting, regions 2 and 4 of Fig. 4.4a are very interesting because while either the treated or control group are underrepresented, there may still be sufficient coverage to estimate treatment effects. D'Amour and Franks [DF21] have described such regions as having weak overlap. Training data bias towards such regions can benefit treatment effect estimation for underrepresented data by acquiring low-frequency data with sufficient overlap.

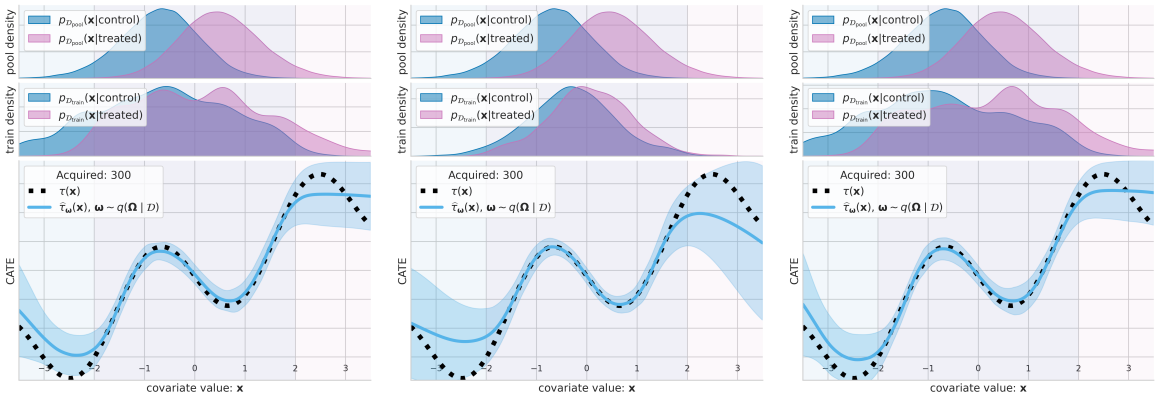
We hypothesize that the efficient acquisition of unlabeled data for treatment effect estimation focuses on only exploring regions with sufficient overlap, and uncertainty should be high for areas with non-overlapping support. The bottom pane of Fig. 4.4a imagines what a resulting training set distribution could look like at an intermediate active learning step. It is not trivial to design such acquisition functions: naively applying active learning acquisition functions results in suboptimal and sample inefficient acquisitions of training examples, as we show below. To this end, we develop epistemic uncertainty-aware methods for active learning of personalized treatment effects from high dimensional observational data. In contrast to previous work that uses only information gain as the acquisition objective, we propose ρBALD and $\mu\rho$BALD as "Causal BALD" objectives because they consider both the information gain and overlap between treated and control groups. We demonstrate the performance of the proposed acquisition strategies using synthetic and semi-synthetic datasets.

4.3.2.1 Experiments

In this section, we evaluate our acquisition objectives on synthetic and semi-synthetic datasets. Code to reproduce these experiments is available at `https://github.com/OATML/causal-bald`.

Datasets Starting from the hypothesis that different objectives can target different types of imbalances and degrees overlap, we construct a **synthetic** dataset [KMZ19] demonstrating the various biases. We depict this dataset graphically in Figure 4.4a. We use this dataset primarily for illustrative purposes. By design, we have constructed a primary data mode and have regions of weak or no overlap. Additionally, we study the performance of our acquisition functions on the **IHDP** dataset [Hil11; SJS17], which is a standard benchmark in causal treatment effect literature. Finally, we demonstrate that our method is suitable for high-dimensional, large-sample datasets on **CMNIST** [Jes+21b], an MNIST [Den12] based dataset adapted for causal treatment effect studies. In Fig. 4.4b, we see that CMNIST is an adaptation of the synthetic dataset. Model inputs are MNIST digits and assigned treatments, and the response surfaces are generated based on a projection of the digits onto a latent 1-dimensional manifold. The observed digits are high-dimensional proxies for the confounding covariate ϕ. Detailed descriptions of each dataset are available in Appendix B.4.3.

Model. Our objectives rely on methods that are capable of modeling uncertainty and handling high-dimensional data modalities. DUE [Van+21] is an instance of Deep Kernel Learning [Wil+16] that uses a deep feature extractor to transform the inputs and defines a Gaussian process (GP) kernel over the extracted feature representation. In particular, DUE uses a variational inducing point approximation [HMG15] and a constrained feature extractor that contains residual connections and spectral normalization to enable reliable uncertainty. Due obtains SotA results on IHDP [Hil11; Van+21]. In DUE, we distinguish between the model parameters $\boldsymbol{\theta}$ and the variational parameters ω, and we are Bayesian only over the ω parameters. Since DUE is a GP, we obtain a full Gaussian posterior over outcomes from which we can use the mean and covariance directly. When necessary, sampling functions from the deep kernel GP is efficient requiring only a single forward pass through the deep kernel encoder. All implementation details are made available by Jesson et al. [Jes+21a] with code to reproduce experiments provided at `https://github.com/OATML/causal-bald`.



(a) $\mu\pi$BALD (b) ρBALD (c) $\mu\rho$BALD

Figure 4.5: Causal–BALD acquisition functions: How the training set is biased and how this effects the CATE function with a fixed budget of 300 acquired points.

Baselines We compare against the following baselines: **Random.** This acquisition function selects points uniformly at random. **Propensity.** An acquisition function based on the propensity score (Eq. 2.4). We train a propensity model on the pool data, which we then use to acquire points based on their propensity score. Please note that this is a valid assumption as training a propensity model does not require outcomes. **γ (S-type error rate) [Sun+19].** This acquisition function is the S-type error rate based method proposed by Sundin et al. [Sun+19]. We have adapted the acquisition function to use with Bayesian Deep Neural Networks. The objective is defined as $\mathrm{I}(\boldsymbol{\gamma}; \boldsymbol{\Theta} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$, where $\gamma(x) = \text{probit}^{-1}\left(-\frac{|\mathbf{E}_{p(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}[\tau]|}{\sqrt{\text{Var}(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}}\right)$ and $\text{probit}^{-1}(\cdot)$ is the

Table 4.4: Summary of active learning parameters for each dataset.

Dataset	Init. size	Acq. size	Acq. steps	Pool Size	Valid Size
Synthetic	10	10	30	10k	1k
IHDP	100	10	38	471	201
CMNIST	250	50	55	35k	15k

cumulative distribution function of normal distribution. In contrast to the original formulation, we do not assume access to counterfactual observations at training time.

Results. For each of the acquisition objectives, dataset, and model we present the mean and standard error of empirical square root of precision in estimation of heterogenous effect (PEHE) [2]. We summarize each active learning setup in Section 4.3.2.1. The *warm up size* is the number of examples in the initial pool dataset. *Acquisition size* is the number of examples labeled at each acquisition step. *Acquisition steps* is the number of times we query a batch of labels. *Pool size* is the number of examples in the pool dataset. Finally, *valid size* is the number of examples used for model selection when optimizing the model at each acquisition step.

In Fig. 4.6, we see that epistemic uncertainty aware $\mu\rho$BALD outperforms the baselines, random, propensity, and S-Type error rate ($\boldsymbol{\gamma}$). As discussed in Section 4.2.4, we expect this improvement as our acquisition objectives target reducible uncertainty – that is, epistemic uncertainty when there is overlap between treatment and control. Additionally, $\mu\rho$BALDshows superior performance over the other objectives in the high-dimensional dataset CMNIST verifying our qualitative analysis in Figure 4.5c.

Each of (μBALD, ρBALD, $\mu\pi$BALD, and $\mu\rho$BALD) outperforms the baseline methods on these tasks. Of note, the performance ρBALD improves as the dimensionality of the covariates increases. In contrast, the performance of the propensity score-based $\mu\pi$BALD worsens as the dimensionality of the covariates increases. Propensity score estimation is known to be a problem in high-dimensions [DAm+21]. We see that both μBALD and $\mu\rho$BALD perform consistently as dimensionality increases, with $\mu\rho$BALD showing a statistically significant improvement over μBALD on two of the three tasks. These improvements indicate that $\mu\rho$BALD is more robust for data with high-dimensional covariates than $\mu\pi$BALD ; moreover, $\mu\rho$BALD does not need an additional propensity score model.

[2] $\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{N}\sum_x (\hat{\tau}(x) - \tau(x))^2}$

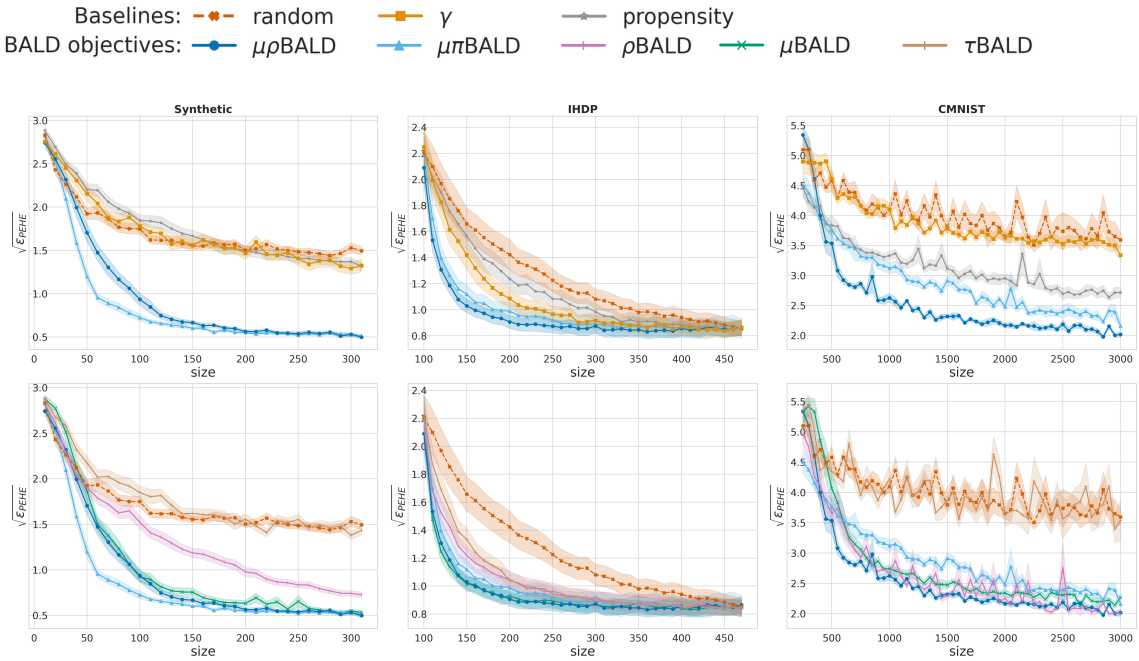
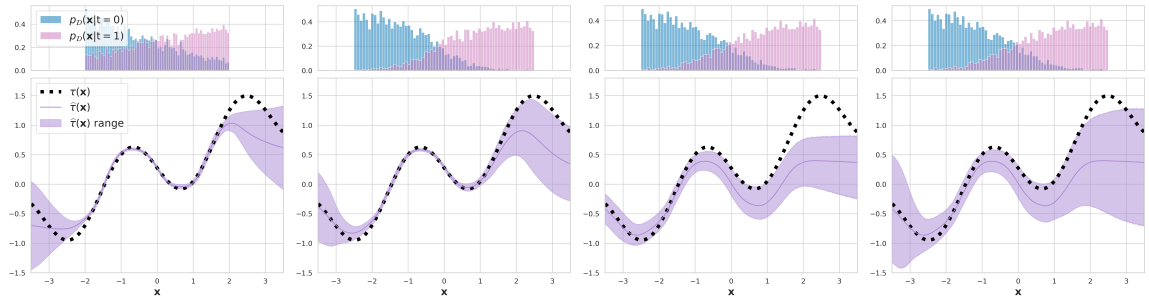


Figure 4.6: $\sqrt{\epsilon_{PEHE}}$ performance (shaded standard error) for DUE models. **(left to right) synthetic** (40 seeds), and **IHDP** (200 seeds). We observe that BALD objectives outperform the **random**, $\boldsymbol{\gamma}$ and **propensity** acquisition functions significantly, suggesting that epistemic uncertainty aware methods that target reducible uncertainty can be more sample efficient.

Chapter 5

Scalable Structural Uncertainty for Causal Machine Learning

In Chapter 4, we focused on **statistical uncertainty**. Statistical uncertainty is informative of when units described by $\mathbf{x}$ are not represented in the data, or when they are not represented in a given treatment arm (violations of Assumption 2.4 (positivity)). In this chapter, we will focus on scalable methods to communicate **structural uncertainty**.



(a) representation (b) positivity (c) ignorability (d) unified

Figure 5.1: The purple shaded areas in the lower panes depict regions of ignorance about the response to intervention for units summarized by the covariate value, $\mathbf{X} = \mathbf{x}$. The training data density for the untreated and treated groups are shown in the upper panes. (Figure 5.1a) For ignorance due to measurements $\mathbf{x}$ without **representation** in the observed data, the region should get wider as the distance between $\mathbf{x}$ and the training data increases. (Figure 5.1b) For ignorance without **positivity**, the region should get wider as $P(\mathrm{T} = 0 \mid \mathbf{x})$ or $P(\mathrm{T} = 1 \mid \mathbf{x}) \to 1$. (Figure 5.1c) For ignorance without **ignorability**, the CATE estimator, $\delta(\mathbf{x}, \boldsymbol{\theta})$, can be arbitrarily biased, hence the discrepancy between the blue solid line and the black dashed line. Therefore, the ignorance region should include the true CATE $\delta(\mathbf{x})$ on the training data manifold where overlap is satisfied. (Figure 5.1d) All sources of ignorance jointly.

In Figure 5.1, we see that in contrast to statistical uncertainty (Figures 5.1a and 5.1b), structural uncertainty (Figure 5.1c) does not go away with more data, and it can bias the estimate of the effect arbitrarily. This can be catastrophic for policy decisions based on treatment effects derived from observational data, and should give us all pause.

A common source of structural uncertainty in treatment effect estimation from observational data is hidden confounding. A confounding variable is a causal parent of both the intervention, T, and the outcome, Y. When we introduced our discussion on statistical uncertainty Section 2.6.1, we treated the observed covariates, $\mathbf{X}$, as confounders or proxies for confounders. We depict a hidden confounder, U, graphically in Figure 5.2. In contrast to the observed confounder, X, the hidden confounder, U, is either unknown, cannot be observed, or is not included in the observational dataset, $\mathcal{D} = \{\mathbf{x}, \mathrm{t}, \mathrm{y}\}_{i=1}^{n}$. A hidden confounder results in a violation of Assumption 2.3 (ignorability). That is, while the conditional independence relationship, $\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathrm{X}, \mathrm{U}$, may hold, the relationship, $\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathrm{X}$, does not. As such, we build on the work of Tan [Tan06] and Kallus, Mao, and Zhou [KMZ19] presented in Section 3.3, which quantifies the bias in treatment effect estimation when, $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t})$ do not equate to one another.

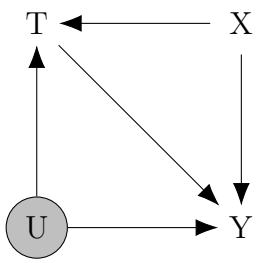


Figure 5.2: Graphical representation of hidden confounding. The causal parent, U, of T and Y exists, but is either not observable or unknown.

The work we present in this chapter unifies the following two publications:

1. Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. "Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding." *ICML*. (2021).

2. Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. "Scalable Sensitivity and Uncertainty Analysis for Causal-Effect Estimates of Continuous-Valued Interventions." *NeurIPS*. (2022).

In Section 5.1 we present discrete and continuous generalizations of the marginal sensitivity model. In Section 5.2 we derive bounds on the CAPO for binary, discrete, and continuous-intervention marginal sensitivity models. Along with bounds on the CATE for the binary MSM. In Section 5.3 we derive estimators for the CAPO and CATE bounds. In Section 5.4 we incorporate statistical uncertainty into our estimators. In Section 5.5 we show how these methods are incorporated into scalable machine learning. In Section 5.6 we provide intuition for interpreting the sensitivity model parameter, Λ. Finally, in Section 5.7 we present experimental results for several applications.

Direct follow up work that will not be presented in this thesis include the following two workshop papers:

1. Andrew Jesson*, Peter Manshausen*, Alyson Douglas*, Duncan Watson-Parris, Yarin Gal, and Philip Stier. "Using Non-Linear Causal Models to Study Aerosol-Cloud Interactions in the Southeast Pacific." *Causal Inference and Machine Learning: Why now? Workshop at NeurIPS*. (2021).

2. Maëlys Solal, Andrew Jesson, Yarin Gal, Alyson Douglas. "Using uncertainty-aware machine learning models to study aerosol-cloud interactions." *Tackling Climate Change with Machine Learning: workshop at NeurIPS*. (2022).

Further publications on the topic of causal sensitivity analysis that the author has served an advisory role on during their DPhil are:

1. Myrl G Marmarelis, Elizabeth Haddad, Andrew Jesson, Neda Jahanshad, Aram Galstyan, Greg Ver Steeg. "Partial identification of dose responses with hidden confounders." *UAI*. (2023).

2. Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, Uri Shalit. "B-Learner: Quasi-Oracle Bounds on Heterogeneous Causal Effects Under Hidden Confounding." *ICML*. (2023).

5.1 Marginal Sensitivity Models

In Section 3.3.1, we described the MSM of Tan [Tan06], $\mathcal{P}_{\mathrm{B}}(\Lambda)$, for binary treatments. We now present a generalization of the MSM for arbitrary discrete valued treatments (Section 5.1.1), and a marginal sensitivity model for continuous valued treatments (Section 5.1.2).

5.1.1 Discrete-Valued Interventions

Where Tan [Tan06] introduces the MSM for binary-valued treatments, we formalized the MSM for arbitrary discrete-valued treatments in [Jes+22]. For discrete-valued treatments, $\mathcal{T}_D = \{\mathrm{t}_i\}_{i=1}^{n_d}$, the (nominal) generalized propensity score [HI04], $p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})$, states how the treatment status, t, depends on the covariates, $\mathbf{x}$, and is identifiable from observational data. The potential outcomes, $\{\mathrm{Y_t} : \mathrm{t} \in \mathcal{T}_D\}$, conditioned on the covariates, $\mathbf{x}$, are distributed as $\{P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) : \mathrm{t} \in \mathcal{T}_D\}$. As for binary-valued treatments, each of these conditional distributions can be written as mixtures with weights based on the generalized propensity score, yielding the following set of mixture distributions,

$$\left\{ P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = \sum_{\mathrm{t}' \in \mathcal{T}_D} p(\mathrm{t}' \mid \mathbf{X} = \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}) \right\}. \tag{5.1}$$

Each conditional distribution of the potential outcome given the observed treatment, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, is identifiable from observational data, but each conditional distribution of the potential outcome given the counterfactual treatment, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$, and therefore each mixture, $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$, is not. Under the ignorability assumption, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$ for all $\mathrm{t}' \in \mathcal{T}_D$.

In order to recover the form of the binary treatment MSM, we can postulate a relationship between the unidentifiable $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and the identifiable $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. Under the assumption that $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ is absolutely continuous with respect to $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$,

we define the Radon-Nikodym derivative:

$$
\begin{aligned}
\lambda_D(\mathrm{y_t}, \mathbf{x}; \mathrm{t}) &= \frac{d(P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}))}{d(1 - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, \\
&= \frac{1}{1 - p(\mathrm{t} \mid \mathbf{x})} \left(\frac{dP(\mathrm{y_t} \mid \mathbf{x})}{dP(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})} - \frac{p(\mathrm{t} \mid \mathbf{x}) dP(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})}{dP(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})} \right), \\
&= \frac{1}{1 - p(\mathrm{t} \mid \mathbf{x})} \left(\frac{\sum_{\mathrm{t}' \in \mathcal{T}_D} p(\mathrm{t}' \mid \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{t}', \mathbf{x})}{dP(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})} - \frac{p(\mathrm{t} \mid \mathbf{x}) dP(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})}{dP(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})} \right), \\
&= \frac{1}{1 - p(\mathrm{t} \mid \mathbf{x})} \left(\frac{\sum_{\mathrm{t}' \in \mathcal{T}_D} p(\mathrm{t}' \mid \mathbf{x}) p(\mathrm{y_t} \mid \mathrm{t}', \mathbf{x})}{p(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})} - \frac{p(\mathrm{t} \mid \mathbf{x}) p(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})}{p(\mathrm{y_t} \mid \mathbf{x}, \mathrm{t})} \right), \\
&= \frac{1}{1 - p(\mathrm{t} \mid \mathbf{x})} \left(\frac{\sum_{\mathrm{t}' \in \mathcal{T}_D} \cancel{p(\mathrm{t}' \mid \mathbf{x})} \frac{p(\mathrm{t}' \mid \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{\cancel{p(\mathrm{t}' \mid \mathbf{x})}}}{\frac{p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t}) \cancel{p(\mathrm{y_t})}}{p(\mathrm{t} \mid \mathbf{x})}} - \frac{p(\mathrm{t} \mid \mathbf{x}) \frac{p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t}) \cancel{p(\mathrm{y_t})}}{\cancel{p(\mathrm{t} \mid \mathbf{x})}}}{\frac{p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t}) \cancel{p(\mathrm{y_t})}}{\cancel{p(\mathrm{t} \mid \mathbf{x})}}} \right), \\
&= \frac{p(\mathrm{t} \mid \mathbf{x})}{1 - p(\mathrm{t} \mid \mathbf{x})} \frac{1 - p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t})}{p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t})}, \\
&= \frac{p(\mathrm{t} \mid \mathbf{x})}{1 - p(\mathrm{t} \mid \mathbf{x})} \Big/ \frac{p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t})}{1 - p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t})},
\end{aligned}
\tag{5.2}
$$

where, $p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t}) \equiv p(\mathrm{t} \mid \mathbf{x}, \mathrm{y_t})$ is the unidentifiable complete propensity density for treatment.

Finally, the discrete treatment MSM further postulates that the odds of receiving the treatment $\mathrm{T} = \mathrm{t}$ for subjects with covariates $\mathbf{X} = \mathbf{x}$ can only differ from $p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})/(1 - p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}))$ by at most a factor of Λ,

$$
\Lambda^{-1} \leq \lambda_D(\mathrm{y_t}, \mathbf{x}; t) \leq \Lambda. \tag{5.3}
$$

Which leads to the following set of functions, $\mathcal{P}_{\mathrm{D}}(\Lambda)$, defining the Discrete Marginal Sensitivity Model (DMSM).

Definition 5.1. *Discrete Marginal Sensitivity Model (DMSM)*

$$
\mathcal{P}_D(\Lambda) \coloneqq \left\{ w(\mathrm{y_t}, \mathbf{x}) \coloneqq \frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} : \alpha_D(\mathbf{x}, \mathrm{t}, \Lambda) \leq w(\mathrm{y_t}, \mathbf{x}) \leq \beta_D(\mathbf{x}, \mathrm{t}, \Lambda) \right\},
$$

$\forall \mathrm{y} \in \mathcal{Y}_t, \forall \mathbf{x} \in \mathcal{X}$. *Where,*

$$
\alpha_D(\mathbf{x}, \mathrm{t}, \Lambda) = \frac{1}{\Lambda p(\mathrm{t} \mid \mathbf{x})} + 1 - \frac{1}{\Lambda}, \quad \textit{and} \quad \beta_D(\mathbf{x}, \mathrm{t}, \Lambda) = \frac{\Lambda}{p(\mathrm{t} \mid \mathbf{x})} + 1 - \Lambda.
$$

5.1.2 Continuous-Valued Interventions

In Jesson et al. [Jes+22], we introduced the Continuous Marginal Sensitivity Model (CMSM), a Marginal Sensitivity Model for continuous treatment variables. The

set of potential outcome distributions, $\{P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) : \mathrm{t} \in \mathcal{T}\}$, conditioned on assigned treatments, $\mathrm{T} = \mathrm{t}$, and observed covariates, $\mathbf{X} = \mathbf{x}$, are identifiable from observed data, $\mathcal{D}$. But, the set of marginal potential outcome distributions, $\{P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) : \mathrm{t} \in \mathcal{T}\}$, each given as a continuous mixture,

$$P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = \int_{\mathcal{T}} P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}) dP(\mathrm{t}' \mid \mathbf{X} = \mathbf{x}),$$

are not. This is because the component distributions of the mixture, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$, are not identifiable without further assumptions as the potential outcome, $\mathrm{Y_t}$, is not observable for units exposed to treatment $\mathrm{T} = \mathrm{t}'$ for $\mathrm{t}' \neq \mathrm{t}$: the well-known "fundamental problem of causal inference" [Hol86]. Yet, under weak-ignorability, $\mathrm{Y_t} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$, the assumed conditional independence [Daw79] between potential outcomes $\mathrm{Y_t}$ and and assigned treatments, T, given observed covariates, $\mathbf{X}$, implies that $P(\mathrm{Y_t} \in A \mid \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \in A \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ are identical for all $A \subseteq \mathcal{Y}$, $\mathbf{x} \in \mathcal{X}$, and $\mathrm{t} \in \mathcal{T}$. Thus, even though such divergence is not identifiable using observational data alone, any divergence between these two distributions would be indicative of hidden confounding.

The CMSM supposes a degree of divergence between the unidentifiable $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ and the identifiable $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ by assuming that the rate of change of $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ with respect to $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ is bounded by some value greater than or equal to 1. This supposition is formalized by the Radon-Nikodym derivative, $\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) = \frac{dP(\mathrm{y_t} \mid \mathbf{X}=\mathbf{x})}{dP(\mathrm{y_t} \mid \mathrm{T}=\mathrm{t}, \mathbf{X}=\mathbf{x})}$, under the following assumption.

Assumption 5.1. *$P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ is absolutely continuous with respect to $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, and $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and the Lebesgue measure are mutually absolutely continuous.*

Theorem 5.1. *Given Assumption 5.1, the Radon-Nikodym derivative above is equal to the ratio between the unidentifiable "complete" propensity density for treatment $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$ and the identifiable "nominal" propensity density for treatment $p(\mathrm{t} \mid \mathbf{x})$,*

$$\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) = \frac{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathrm{Y_t} = \mathrm{y_t}, \mathbf{X} = \mathbf{x})}, \tag{5.4}$$

Proof. Let the range of $\mathrm{Y_t}$ be the measurable space $(\mathcal{Y}, \mathcal{A})$, and $\nu(A)$ denote the

Lebesgue measure for any measurable $A \in \mathcal{A}$. Then,

$$
\begin{aligned}
\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) &= \frac{dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, && \text{R-N derivative,} \\
&= \frac{dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{d\nu} \frac{d\nu}{dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, && \text{Assumption 5.1,} \\
&= \frac{dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{d\nu} \left(\frac{dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{d\nu} \right)^{-1}, && \text{Assumption 5.1,} \\
&= \frac{d}{d\nu} \int_A p(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x}) d\nu \left(\frac{d}{d\nu} \int_A p(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) d\nu \right)^{-1}, && \text{R-N theorem,} \\
&= \frac{p(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, && \text{fund. th. calculus,} \\
&= \frac{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathrm{Y_t} = \mathrm{y_t}, \mathbf{X} = \mathbf{x})}, && \text{Bayes' Rule.}
\end{aligned}
$$

□

The value of the ratio, $\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t})$, cannot be identified from the observational data alone; therefore, the merit of the CMSM is that it enables a domain expert to express their belief in what is a plausible degree of hidden confounding through the parameter $\Lambda \geq 1$. Where,

$$\Lambda^{-1} \leq \frac{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathrm{Y_t} = \mathrm{y}, \mathbf{X} = \mathbf{x})} \leq \Lambda,$$

reflects a hypothesis that the "complete", unidentifiable propensity density for subjects with covariates $\mathbf{X} = \mathbf{x}$ can be different from the identifiable "nominal" propensity density by at most a factor of Λ. These inequalities allow for the specification of user hypothesized complete propensity density functions, $p(\mathrm{t} \mid \mathrm{Y_t} = \mathrm{y_t}, \mathbf{X} = \mathbf{x})$, and we define the CMSM as the set of such functions that agree with the inequalities:

Definition 5.2. *Continuous Marginal Sensitivity Model (CMSM)*

$$\mathcal{P}_C(\Lambda) \coloneqq \left\{ w(\mathrm{y_t}, \mathbf{x}) \coloneqq \frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} : \alpha_C(\mathbf{x}, \mathrm{t}, \Lambda) \leq w(\mathrm{y_t}, \mathbf{x}) \leq \beta_C(\mathbf{x}, \mathrm{t}, \Lambda) \right\},$$

$\forall \mathrm{y} \in \mathcal{Y}_t, \forall \mathbf{x} \in \mathcal{X}$. *Where,*

$$\alpha_C(\mathbf{x}, \mathrm{t}, \Lambda) \coloneqq \frac{1}{\Lambda p(\mathrm{t} \mid \mathbf{x})}, \quad \textit{and} \quad \beta_C(\mathbf{x}, \mathrm{t}, \Lambda) \coloneqq \frac{\Lambda}{p(\mathrm{t} \mid \mathbf{x})}.$$

5.2 Causal-Effect Bounds Without Ignorability

The CAPO and CATE functions cannot be point identified from observational data without ignorability, but under the MSM, DMSM, and CMSM we can identify a set of

functions that are consistent with both the observational data $\mathcal{D}$ and the assumptions encoded by the parameter Λ. All of the functions in this set are possible from the point of view of the observational data alone. So to cover the range of all possible functional values, we seek an interval function that maps covariate values, $\mathbf{X} = \mathbf{x}$, to the upper and lower bounds of this set for every treatment value, $\mathrm{t} \in \mathcal{T}$. In this section, we explain how to translate the set of inverse complete propensity functions assumed by a Marginal Sensitivity Model, $\mathcal{P}_{(\cdot)}(\Lambda)$, into bounds on causal-effects. Specifically, we look at bounds on the Conditional Average Potential Outcome (CAPO), $f_{\mathrm{t}}(\mathbf{x})$; and Conditional Average Treatment Effect (CATE), $\delta(\mathbf{x})$.

Whether we use the MSM for binary, $\mathcal{P}_B(\Lambda)$; discrete, $\mathcal{P}_D(\Lambda)$; or continuous treatments, $\mathcal{P}_C(\Lambda)$: we start with the expression for the CAPO given by Kallus, Mao, and Zhou [KMZ19],

$$f_{\mathrm{t}}(\mathbf{x}) = \frac{\int_{\mathcal{Y}} \mathrm{y}_{\mathrm{t}} w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) dP(\mathrm{y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) dP(\mathrm{y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})},$$

which, as we stated in Section 3.3.2, elegantly expresses the unbiased conditional expectation of the potential outcome in terms of the unidentifiable inverse complete propensity $w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) = 1/e(\mathrm{y}_{\mathrm{t}}, \mathbf{x})$ and the conditional probability or density $p(\mathrm{y}_{\mathrm{t}} \mid \mathbf{x}, \mathrm{t})$ of the outcome. We add and subtract the possibly biased statistical CAPO, $f(\mathbf{x}, \mathrm{t})$, from Section 5.2, and proceed from,

$$f_{\mathrm{t}}(\mathbf{x}) = f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \big(\mathrm{y}_{\mathrm{t}} - f(\mathbf{x}, \mathrm{t})\big) w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) dP(\mathrm{y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) dP(\mathrm{y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})},$$

for the sake of cleaner notation in the following.

Binary and discrete interventions. For the cases of binary and discrete interventions, we propose the following equivalent expression for $f_{\mathrm{t}}(\mathbf{x})$:

$$f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_D(\Lambda)\big) = f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} (\mathrm{y} - f(\mathbf{x}, \mathrm{t})) w(\mathrm{y}, \mathbf{x}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}{\alpha'_D(\mathbf{x}, \mathrm{t}, \Lambda) + \int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}, \tag{5.6}$$

where the function, $w(\mathrm{y}, \mathbf{x})$, has range, $[0, 1]$, such that,

$$w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) = \alpha_D(\mathbf{x}, \mathrm{t}, \Lambda) + w(\mathrm{y}, \mathbf{x})\big(\beta_D(\mathbf{x}, \mathrm{t}, \Lambda) - \alpha_D(\mathbf{x}, \mathrm{t}, \Lambda)\big),$$

and,

$$\alpha'_D(\mathbf{x}, \mathrm{t}, \Lambda) \coloneqq \frac{\alpha_D(\mathbf{x}, \mathrm{t}, \Lambda)}{\beta_D(\mathbf{x}, \mathrm{t}, \Lambda) - \alpha_D(\mathbf{x}, \mathrm{t}, \Lambda)}.$$

Proof for the equality of $f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_D(\Lambda)\big)$ and $f_{\mathrm{t}}(\mathbf{x})$ given the true inverse propensity, $w(\mathrm{y}_{\mathrm{t}}, \mathbf{x})$ is given in Lemma D.11.

Continuous interventions. For continuous interventions, we propose the following equivalent expression for $f_{\mathrm{t}}(\mathbf{x})$,

$$f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_C(\Lambda)\big) = f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \big(\mathrm{y} - f(\mathbf{x}, \mathrm{t})\big) w(\mathrm{y}, \mathbf{x}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}{(\lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}. \tag{5.7}$$

where again the function, $w(\mathrm{y}, \mathbf{x})$, has range, $[0, 1]$, such that,

$$w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) = \alpha_C(\mathbf{x}, \mathrm{t}, \Lambda) + w(\mathrm{y}, \mathbf{x})\big(\beta_C(\mathbf{x}, \mathrm{t}, \Lambda) - \alpha_C(\mathbf{x}, \mathrm{t}, \Lambda)\big).$$

Note, that from the definitions of $\alpha_C(\mathbf{x}, \mathrm{t}, \Lambda)$ and $\beta_C(\mathbf{x}, \mathrm{t}, \Lambda)$ in Definition 5.2, we have

$$\begin{aligned} \frac{\alpha_C(\mathbf{x}, \mathrm{t}, \Lambda)}{\beta_C(\mathbf{x}, \mathrm{t}, \Lambda) - \alpha_C(\mathbf{x}, \mathrm{t}, \Lambda)} &= \frac{\frac{1}{\Lambda p(\mathrm{t}|\mathbf{x})}}{\frac{\Lambda}{p(\mathrm{t}|\mathbf{x})} - \frac{1}{p(\mathrm{t}|x)}}, \\ &= (\lambda^2 - 1)^{-1}, \end{aligned}$$

which shows that in contrast to the MSM and DMSM, the dependency on $p(\mathrm{t} \mid \mathbf{x})$ factors out in the CMSM. Proof for the equality of $f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_C(\Lambda)\big)$ and $f_{\mathrm{t}}(\mathbf{x})$ given the true inverse propensity, $w(\mathrm{y}_{\mathrm{t}}, \mathbf{x})$, is also given in Lemma D.11.

The uncertainty sets that include all possible values of $w(\mathrm{y}, \mathbf{x})$ that agree with a given MSM, $\mathcal{P}_D(\Lambda)$; i.e., the sets of functions that violate ignorability by no more than Λ, are given in Definitions 3.2, 5.1 and 5.2. With such a set of functions, we can now define the CAPO and CATE bounds.

Definition 5.3. *The lower, $\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big)$, and upper, $\overline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big)$, bounds on the CAPO, $f_{\mathrm{t}}(\mathbf{x})$, under an assumed Marginal Sensitivity Model, $\mathcal{P}_{(\cdot)}(\Lambda))$ are,*

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big) \coloneqq \inf_{w(\mathrm{y}, \mathbf{x}) \in \mathcal{P}_{(\cdot)}(\Lambda)} f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big) = \inf_{w(\mathrm{y}) \in \mathcal{W}^H_{ni}} f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big), \tag{5.8a}$$

$$\overline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big) \coloneqq \sup_{w(\mathrm{y}, \mathbf{x}) \in \mathcal{P}_{(\cdot)}(\Lambda)} f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big) = \sup_{w(\mathrm{y}) \in \mathcal{W}^H_{nd}} f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big). \tag{5.8b}$$

Where the sets, $\mathcal{W}^H_{ni}$ and $\mathcal{W}^H_{nd}$, are defined as,

$$\mathcal{W}^H_{ni} \coloneqq \{w : w(y) = H(\mathrm{y}_H - \mathrm{y})\}_{\mathrm{y}_H \in \mathcal{Y}}, \tag{5.9a}$$

$$\mathcal{W}^H_{nd} \coloneqq \{w : w(y) = H(\mathrm{y} - \mathrm{y}_H)\}_{\mathrm{y}_H \in \mathcal{Y}}, \tag{5.9b}$$

and, $H(\cdot)$, is the Heaviside step function. The equalities in Equations (5.8a) and (5.8b) follow from Lemma 1 in Kallus, Mao, and Zhou [KMZ19] and Lemma D.11. Lemma D.13 reassures us that the result of Lemma 1 in Kallus, Mao, and Zhou [KMZ19] holds for the CMSM.

The bounds on the CAPO in Equation (5.8) under the binary MSM, $\mathcal{P}_B(\Lambda)$, completely define the bounds on the CATE.

Definition 5.4. *The lower, $\underline{\delta}(\mathbf{x}, \mathcal{P}_B(\Lambda))$, and upper, $\overline{\delta}(\mathbf{x}, \mathcal{P}_B(\Lambda))$, bounds on the CATE, $\delta(\mathbf{x})$, under an assumed binary Marginal Sensitivity Model, $\mathcal{P}_B(\Lambda))$ are,*

$$\underline{\delta}(\mathbf{x}, \mathcal{P}_B(\Lambda))) \coloneqq \underline{f}_1(\mathbf{x}, \mathcal{P}_B(\Lambda))) - \overline{f}_0(\mathbf{x}, \mathcal{P}_B(\Lambda))), \tag{5.10a}$$

$$\overline{\delta}(\mathbf{x}, \mathcal{P}_B(\Lambda))) \coloneqq \overline{f}_1(\mathbf{x}, \mathcal{P}_B(\Lambda))) - \underline{f}_0(\mathbf{x}, \mathcal{P}_B(\Lambda))). \tag{5.10b}$$

Equation (5.10b) is proved in Lemma D.13 for the MSM, $\mathcal{P}_B(\Lambda)$. Equation (5.10a) can be proved analogously to Equation (5.10b).

5.3 CAPO and CATE Interval Estimators

We now need to provide estimators for the bounds presented in Section 5.2. Before addressing scale in Section 5.5, we will present our most general result, which holds asymptotically for convergent density estimators of bounded outcomes, $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$, and consistent estimators of the statistical CAPO, $f(\mathbf{x}, \mathrm{t}) = \mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$ and nominal propensity, $p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})$. Note, the nominal propensity is only needed for the MSM and DMSM. While the result holds for both parametric and non-parametric estimators, we will frame the discussion in terms of parametric estimators of aforementioned estimands: $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$, $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$, and $p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\theta}_n)$, which facilitates our later discussion on scale. Here, $\boldsymbol{\theta}_n$, will stand for the parameters of the model fit to dataset, $\mathcal{D}_n = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$.

We also need estimators of the integrals in Equations (5.6) and (5.7) for the term, $f_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big)$. We present results using Monte-Carlo (MC) integration to estimate the expectation of arbitrary functions $h(y)$ with respect to the density estimator, $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$,

$$S_m(h(\mathrm{y})) \coloneqq \frac{1}{m} \sum_{i=1}^{m} h(\mathrm{y}_i), \quad \mathrm{y}_i \sim p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n).$$

We outline how the Gauss-Hermite quadrature rule is an alternate estimator of these expectations in Appendix D.2.3.

Definition 5.5. *The CAPO lower, $\underline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big)$, and upper, $\overline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big)$, bound estimators under an MSM, $P_{(\cdot)}(\Lambda) \in \{\mathcal{P}_B(\Lambda), \mathcal{P}_D(\Lambda), \mathcal{P}_C(\Lambda)\}$, are*

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big) \coloneqq \inf_{w \in \mathcal{W}_{ni}^H} f_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big), \tag{5.11}$$

$$\overline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big) \coloneqq \sup_{w \in \mathcal{W}_{nd}^H} f_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big), \tag{5.12}$$

where,

$$f_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big) \equiv f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) + \frac{S_m\left(w(\mathrm{y})(\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n))\right)}{(\Lambda^2 - 1)^{-1} + S_m\left(w(\mathrm{y})\right)}.$$

Definition 5.6. *The CATE lower,* $\underline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n, S_m\big)$*, and upper,* $\overline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n, S_m\big)$*, bound estimators under the binary MSM,* $\mathcal{P}_B(\Lambda)$*, are,*

$$\underline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n, S_m\big) \coloneqq \underline{f}_1\big(\mathbf{x}, P_B(\Lambda), \boldsymbol{\theta}_n, S_m\big) - \overline{f}_0\big(\mathbf{x}, P_B(\Lambda), \boldsymbol{\theta}_n, S_m\big), \tag{5.13}$$

$$\overline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n, S_m\big) \coloneqq \overline{f}_1\big(\mathbf{x}, P_B(\Lambda), \boldsymbol{\theta}_n, S_m\big) - \underline{f}_0\big(\mathbf{x}, P_B(\Lambda), \boldsymbol{\theta}_n, S_m\big). \tag{5.14}$$

This brings us to the main result of this section.

Theorem 5.2. *In the limit of data* ($n \to \infty$) *and MC samples* ($m \to \infty$)*. For observed* $(\mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}) \in \mathcal{D}_n = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$ *and bounded* Y*. We assume that* $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ *converges in measure to* $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$*,* $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ *is a consistent estimator of* $f(\mathbf{x}, \mathrm{t})$*,* $p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\theta}_n)$ *is consistent estimator of* $p(\mathrm{t} \mid \mathbf{x})$*, and* $p(\mathrm{t} \mid \mathrm{y}_\mathrm{t}, \mathbf{x})$ *is bounded away from 0 uniformly for all* $\mathrm{y}_\mathrm{t} \in \mathcal{Y}$*. Then, for a marginal sensitivity model,* $P_{(\cdot)}(\Lambda) \in \{\mathcal{P}_B(\Lambda), \mathcal{P}_D(\Lambda), \mathcal{P}_C(\Lambda)\}$*,*

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big) \xrightarrow{p} \underline{f}_{\mathrm{t}}(\mathbf{x}, P_{(\cdot)}(\Lambda)),$$

and

$$\overline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m\big) \xrightarrow{p} \overline{f}_{\mathrm{t}}(\mathbf{x}, P_{(\cdot)}(\Lambda)).$$

Further, for the binary marginal sensitivity model, $\mathcal{P}_B(\Lambda)$*,*

$$\underline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n, S_m\big) \xrightarrow{p} \underline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda)\big),$$

and

$$\overline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n, S_m\big) \xrightarrow{p} \overline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda)\big).$$

Proof in Appendix D.2.1.

Now that we have estimators for our target bounds, we need to solve the optimization problems for the infinimum and supremum in Definition 5.5.

5.3.1 Solving For w

We are interested in a scalable algorithm to compute the intervals on the CAPO (Eqs. (5.11) and (5.12)) and the CATE (Eqs. (5.13) and (5.14)) functions given a belief in the degree of divergence from Assumption 2.3 (ignorability), Λ. The need for scalability stems not only from dataset size. The intervals may also need to be evaluated for arbitrarily many values of the treatment variable, t, and the sensitivity parameter Λ.

The bounds on the CAPO and CATE functions can be calculated independently for each instance $\mathbf{x}$. The upper and lower bounds of the CAPO function under treatment, t, and sensitivity parameter, Λ, can be estimated for any observed covariate value, $\mathbf{x}$, as

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big) \coloneqq f_{\mathrm{t}}\Big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m^{H(\underline{\mathrm{y}}-\mathrm{y})}\Big),$$
$$\overline{f}_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big) \coloneqq f_{\mathrm{t}}\Big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m^{H(\mathrm{y}-\overline{\mathrm{y}})}\Big),$$

where the superscript, $H(\cdot)$, descibres the function, $w(\mathrm{y})$, in the integral estimator, S_m, and the solutions, $\underline{\mathrm{y}}$ and $\overline{\mathrm{y}}$, are found using the grid search Algorithm 1. We choose the grid search method because it is very amenable to parallelization and thus easily scalable. We provide alternative optimization methods in Algorithm 2 and Appendix D.2.3.3.

Algorithm 1 Grid Search Interval Optimizer

Require: $\mathbf{x}$ is an instance of $\mathbf{X}$, t is a treatment level to evaluate, Λ is a belief in the amount of hidden confounding, $\boldsymbol{\theta}$ are optimized model parameters, $\widehat{\mathcal{Y}}$ is a set of unique values $\{\mathrm{y} \sim p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)\}$.

function GridSearch($\mathbf{x}$, t, Λ, $\boldsymbol{\theta}$, $\widehat{\mathcal{Y}}$)
  $\overline{f} \leftarrow -\infty$, $\overline{\mathrm{y}} \leftarrow 0$
  $\underline{f} \leftarrow \infty$, $\underline{\mathrm{y}} \leftarrow 0$
  for $\mathrm{y}_H \in \widehat{\mathcal{Y}}$ **do**
    $\overline{\kappa} \leftarrow f_{\mathrm{t}}\Big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m^{H(\mathrm{y}-\mathrm{y}_H)}\Big)$
    $\underline{\kappa} \leftarrow f_{\mathrm{t}}\Big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n, S_m^{H(\mathrm{y}_H-\mathrm{y})}\Big)$
    if $\overline{\kappa} > \overline{f}$ **then**
      $\overline{f} \leftarrow \overline{\kappa}$, $\overline{\mathrm{y}} \leftarrow \mathrm{y}_H$
    if $\underline{\kappa} < \underline{f}$ **then**
      $\underline{f} \leftarrow \underline{\kappa}$, $\underline{\mathrm{y}} \leftarrow \mathrm{y}_H$
  return $\underline{\mathrm{y}}, \overline{\mathrm{y}}$

The corresponding optimized CATE bounds for a binary MSM, $P_B(\Lambda)$, are,

$$\underline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n\big) = \underline{f}_1\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big) - \overline{f}_0\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big),$$

and

$$\overline{\delta}\big(\mathbf{x}, \mathcal{P}_B(\Lambda), \boldsymbol{\theta}_n\big) = \overline{f}_1\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big) - \underline{f}_0\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big).$$

5.4 Finite-Sample Causal-Effect Bounds Without Ignorability

The above bounds are valid and sharp in the limit as $n \to \infty$. We now take statistical uncertainty into account and provide scalable finite-sample bounds on the CAPO and CATE estimates when the ignorability assumption is relaxed according to a given marginal sensitivity model, $P_{(\cdot)}(\Lambda)$.

Following Zhao, Small, and Bhattacharya [ZSB19], Dorn and Guo [DG23], and Chernozhukov et al. [Che+21], we construct $(1-\alpha)$ statistical confidence intervals for the upper and lower bounds under an MSM, $P_{(\cdot)}(\Lambda)$, using the percentile bootstrap estimator. We have shown in [Jes+20] and [Jes+21b] that statistical uncertainty is appropriately high for regions with poor overlap. Let $P_{\mathcal{D}}$ be the empirical distribution of the observed data sample, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n = \{\mathbf{S}_i\}_{i=1}^n$. Let $\widehat{P}_{\mathcal{D}} = \{\widehat{\mathcal{D}}_k\}_{k=1}^{n_b}$ be the bootstrap distribution over n_b datasets, $\widehat{\mathcal{D}}_k = \{\widehat{\mathbf{S}}_i\}_{i=1}^n$, sampled with replacement from the empirical distribution, $P_{\mathcal{D}}$. Let Q_α be the α-quantile of $f_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big)$ in the bootstrap resampling distribution: $Q_\alpha \coloneqq \inf_{f^*}\left\{\widehat{P}_{\mathcal{D}}(f_{\mathrm{t}}\big(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n\big) \leq f^*) \geq \alpha\right\}$. Then, letting $\boldsymbol{\theta}_n^k$ be the parameters of the model of the k-th bootstrap sample of the data, we can define our statistical uncertainty aware ignorance intervals about the CAPO, $\mathrm{f_t}(\mathbf{x})$, and CATE, $\delta(\mathbf{x})$, functions.

Definition 5.7. *The bootstrap $1-\alpha$ confidence interval of the upper and lower bounds of the CAPO function under a marginal sensitivity model, $P_{(\cdot)}(\Lambda)$, is*

$$\mathrm{CI}_b\left(f_{\mathrm{t}}(\mathbf{x}, P_{(\cdot)}(\Lambda), \alpha)\right) \coloneqq \left[\underline{f}_{\mathrm{t}}^b(\mathbf{x}, P_{(\cdot)}(\Lambda), \alpha), \overline{f}_{\mathrm{t}}^b(\mathbf{x}, P_{(\cdot)}(\Lambda), \alpha)\right],$$

where,

$$\underline{f}_{\mathrm{t}}^b(\mathbf{x}, P_{(\cdot)}(\Lambda), \alpha) = Q_{\alpha/2}\left(\left\{\underline{f}_{\mathrm{t}}(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n^k)\right\}_{k=1}^b\right),$$

and

$$\overline{f}_{\mathrm{t}}^b(\mathbf{x}, P_{(\cdot)}(\Lambda), \alpha) = Q_{1-\alpha/2}\left(\left\{\overline{f}_{\mathrm{t}}(\mathbf{x}, P_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n^k)\right\}_{k=1}^b\right).$$

Definition 5.8. *Under the binary MSM, $P_B(\Lambda)$, the bootstrap $1-\alpha$ confidence interval of the upper and lower bounds of the CATE function is*

$$\mathrm{CI}_b\left(\delta(\mathbf{x}, P_B(\Lambda), \alpha)\right) \coloneqq \left[\underline{\delta}^b(\mathbf{x}, P_B(\Lambda), \alpha), \overline{\delta}^b(\mathbf{x}, P_B(\Lambda), \alpha)\right],$$

where,

$$\underline{\delta}^b(\mathbf{x}, P_B(\Lambda), \alpha) = \underline{f}_1^b(\mathbf{x}, P_B(\Lambda), \alpha) - \overline{f}_0^b(\mathbf{x}, P_B(\Lambda), \alpha),$$

and

$$\overline{\delta}^b(\mathbf{x}, P_B(\Lambda), \alpha) = \overline{f}_1^b(\mathbf{x}, P_B(\Lambda), \alpha) - \underline{f}_0^b(\mathbf{x}, P_B(\Lambda), \alpha).$$

We finally offer definitions for naive statistical uncertainty aware ignorance intervals on the APO and and ATE functions.

Definition 5.9. *The naive bootstrap $1-\alpha$ confidence interval of the upper and lower bounds of the APO function under a marginal sensitivity model, $P_{(\cdot)}(\Lambda)$, is*

$$\mathrm{CI}_b\left(f_{\mathrm{t}}(P_{(\cdot)}(\Lambda), \alpha)\right) \coloneqq \left[\underline{f}_{\mathrm{t}}^b(P_{(\cdot)}(\Lambda), \alpha), \overline{f}_{\mathrm{t}}^b(P_{(\cdot)}(\Lambda), \alpha)\right],$$

where,

$$\underline{f}_{\mathrm{t}}^b(P_{(\cdot)}(\Lambda), \alpha) = \frac{1}{n}\sum_{i=1}^{n} \underline{f}_{\mathrm{t}}^b(\mathbf{x}_i, P_{(\cdot)}(\Lambda), \alpha),$$

and

$$\overline{f}_{\mathrm{t}}^b(P_{(\cdot)}(\Lambda), \alpha) = \frac{1}{n}\sum_{i=1}^{n} \overline{f}_{\mathrm{t}}^b(\mathbf{x}_i, P_{(\cdot)}(\Lambda), \alpha).$$

Definition 5.10. *Under the binary MSM, $P_B(\Lambda)$, the naive bootstrap $1-\alpha$ confidence interval of the upper and lower bounds of the ATE function is given by:*

$$\mathrm{CI}_b\left(\delta(P_B(\Lambda), \alpha)\right) \coloneqq \left[\underline{\delta}^b(P_B(\Lambda), \alpha), \overline{\delta}^b(P_B(\Lambda), \alpha)\right],$$

where,

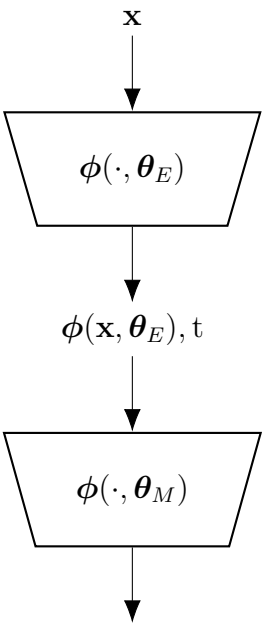
$$\underline{\delta}^b(P_B(\Lambda), \alpha) = \frac{1}{n}\sum_{i=1}^{n} \underline{\delta}^b(\mathbf{x}_i, P_B(\Lambda), \alpha),$$

and

$$\overline{\delta}^b(P_B(\Lambda), \alpha) = \frac{1}{n}\sum_{i=1}^{n} \overline{\delta}^b(\mathbf{x}_i, P_B(\Lambda), \alpha).$$

5.5 Scalable Treatment Effect Estimation Under Structural Uncertainty

The neural network motif we presented in Section 3.1 is readily adapted to the setting of scalable structural uncertainty quantification. We re-illustrate this motif in Figure 5.3 for convenience. The encoder network, $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E)$, architecture can be modified to accommodate a variety of input data modalities. In Section 5.7.1.1 we demonstrate this for images using convolutional neural networks (CNNs) [LeC+98] and in Section 5.7.3 for satellite remote sensing data using Transformers [Vas+17]. Violations in ignorability can induce multi-modal distributions over Y; therefore, we need flexible density estimators to account for this.



$$\boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM}) \coloneqq \boldsymbol{\phi}(\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E), \mathrm{t}, \boldsymbol{\theta}_M)$$

$$f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}) \coloneqq \boldsymbol{\theta}_f^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM})$$

Figure 5.3: Neural network design motif for causal effect estimation

Continuous Outcomes. For continuous outcomes, we model $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta})$ with a Gaussian mixture density over j mixture components, noting that with a sufficient number of mixture components it can approximate any continuous distribution [Tit+85]. The density is expressed mathematically as,

$$p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) = \sum_{j=1}^{n_j} \pi_j\big(\boldsymbol{\theta}_\pi^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM})\big) \mathcal{N}\big(\mathrm{y} \mid \boldsymbol{\theta}_{f_j}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM}), \boldsymbol{\theta}_{\sigma_j}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM})\big). \tag{5.15}$$

And the CAPO estimator is given by,

$$f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) = \sum_{j=1}^{n_\mathrm{y}} \pi_j\big(\boldsymbol{\theta}_\pi^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_E)\big) \boldsymbol{\theta}_{f_j}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM}). \tag{5.16}$$

Discrete Outcomes. For discrete outcomes, we use a standard Categorical distribuition (or Bernoulli for binary outcomes). The density is expressed mathematically as,

$$p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) = \mathrm{Categorical}\Big(\mathrm{y} \mid \mathrm{softmax}\big(\boldsymbol{\theta}_f^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM})\big)\Big). \tag{5.17}$$

And the CAPO estimator is given by,

$$f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) = \mathrm{softmax}\big(\boldsymbol{\theta}_f^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_{EM})\big). \tag{5.18}$$

Propensity Model For the binary or discrete marginal sensitivity models ($\mathcal{P}_B(\Lambda)$ or $\mathcal{P}_B(\Lambda)$), we need an estimator for the propensity score, $p(\mathrm{t} \mid \mathbf{x})$. Following the Dragonnet architecture of Shi, Blei, and Veitch [SBV19], we append a prediction head after the encoder newtork, $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E)$,

$$p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\theta}_n) = \mathrm{softmax}\big(\boldsymbol{\theta}_{\mathrm{t}}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_E)\big). \tag{5.19}$$

Models are optimized by maximizing the log-likelihood of $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ and $p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\theta}_n)$ over the observational data, $\mathcal{D}_n = (\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)_{i=1}^n$.

5.6 Interpreting Λ

Before deriving semi-parametric interval estimators for CAPO and APO, it is worth pausing here and breaking down Equation (5.7) to get an intuitive sense of how the specification of Λ in the CMSM affects the bounds on the causal estimands. Note that the CMSM is defined in terms of a *density ratio*, $p(\mathrm{t} \mid \mathbf{x})/p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})$, whereas the MSM for binary-valued treatments is defined in terms of an *odds ratio*, $\frac{P(\mathrm{t}|\mathbf{x})}{(1-P(\mathrm{t}|\mathbf{x}))} / \frac{P(\mathrm{t}|\mathrm{y}_{\mathrm{t}},\mathbf{x})}{(1-P(\mathrm{t}|\mathrm{y}_{\mathrm{t}},\mathbf{x}))}$. Importantly, naively substituting densities into the MSM for binary-treatments would violate the condition that $\lambda > 0$ as the densities $p(\mathrm{t} \mid \mathbf{x})$ or $p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})$ can each be greater than one, which would result in a negative $1 - p(\mathrm{t} \mid \cdot)$. The odds ratio is familiar to practitioners, but the density ratio is less so.

Λ as A Proportion of Outcome Error. When $\Lambda \to 1$, then the $(\Lambda^2 - 1)^{-1}$ term (and thus the denominator) in Equation (5.7) tends to infinity. As a result, the CAPO under Λ converges to the empirical estimate of the CAPO — $f(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda \to 1) \to f(\mathbf{x}, \mathrm{t})$ — as expected. Thus, the supremum and infimum in Equations (5.8) and (5.8a) become independent of w, and the ignorance intervals concentrate on point estimates. Next, consider complete relaxation of the ignorability assumption, $\Lambda \to \infty$. Then, the $(\Lambda^2 - 1)^{-1}$ term tends to zero, and we are left with,

$$\begin{aligned} f(w; \cdot, \Lambda \to \infty) &\to f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - f(\mathbf{x}, \mathrm{t}))dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}{\int_{\mathcal{Y}} w(\mathrm{y})dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}, \\ &= f(\mathbf{x}, \mathrm{t}) + \int (\mathrm{Y} - f(\mathbf{x}, \mathrm{t}))dP(w(\mathrm{y}) \mid \mathbf{x}, \mathrm{t}), \end{aligned}$$

where, $p(w(\mathrm{y}) \mid \mathbf{x}, \mathrm{t}) \coloneqq \frac{w(\mathrm{y})p(\mathrm{y}|\mathbf{x},\mathrm{t})}{\int_{\mathcal{Y}} w(\mathrm{y}')dP(\mathrm{y}'|\mathbf{x},\mathrm{t})}$, a distribution over Y given $\mathbf{X} = x$ and $\mathrm{T} = \mathrm{t}$. Thus, when we *relax* the ignorability assumption entirely, the CAPO can be anywhere in the range of Y.

The parameter Λ, therefore relates to the proportion of unexplained range in Y assumed to come from unobserved confounders after observing $\mathbf{x}$ and t. When a user sets Λ to 1, they assume that the entire unexplained range of Y comes from unknown mechanisms independent of T. As the user increases Λ, they attribute some of the unexplained range of Y to mechanisms causally connected to T. For bounded $\mathrm{Y_t}$, this proportion can be calculated as,

$$\rho(\mathbf{x}, \mathrm{t}; \Lambda) \coloneqq \frac{\overline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda) - \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)}{\overline{f}(\mathbf{x}, \mathrm{t}; \Lambda \to \infty) - \underline{f}(\mathbf{x}, \mathrm{t}; \Lambda \to \infty)} = \frac{\overline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda) - \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)}{\mathrm{y_{max}} - \mathrm{y_{min}} \mid \mathbf{X} = x, \mathrm{T} = \mathrm{t}}.$$

The user can sweep over a set of Λ values and report the bounds corresponding to a ρ value they deem tolerable (e.g., $\rho = 0.5$ yields bounds for the assumption that half the unexplained range in Y is due to unobserved confounders). For unbounded outcomes, the limits can be estimated empirically by increasing Λ to a large value. Figure 5.4 compares ρ and Λ.

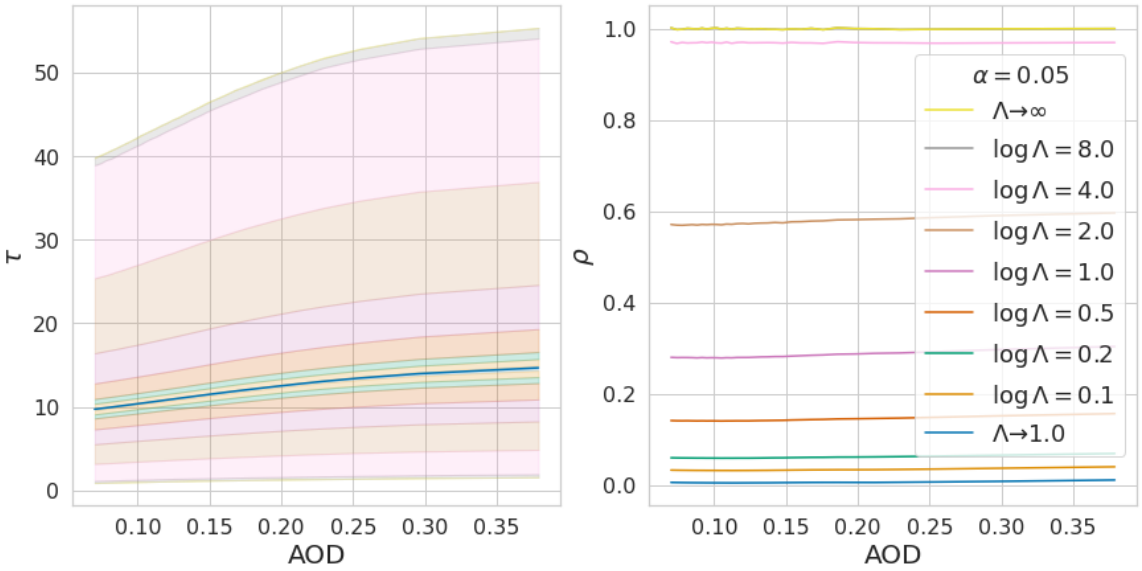


Figure 5.4: Interpreting Λ as a proportion (ρ) of the unexplained range of $\mathrm{Y_t}$ attributed to unobserved confounding variables. Here, AOD is the intervention variable, which will be defined in Section 5.7.2.1

Λ as An Upper Bound on A KL-Divergence. The bounds on the density ratio can also be expressed as bounds on the Kullback–Leibler (KL) divergence between

$P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$. Specifically,

$$
\begin{aligned}
\Lambda^{-1} &\leq \frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} \leq \Lambda, \\
\log\left(\Lambda^{-1}\right) &\leq \log\left(\frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}\right) \leq \log\left(\Lambda\right), \\
\mathbb{E}_{f(\mathrm{y_t}|\mathbf{x},\mathrm{t})} \log\left(\Lambda^{-1}\right) &\leq \mathbb{E}_{f(\mathrm{y_t}|\mathbf{x},\mathrm{t})} \log\left(\frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}\right) \leq \mathbb{E}_{f(\mathrm{y_t}|\mathbf{x},\mathrm{t})} \log\left(\Lambda\right), \\
\log\left(\Lambda^{-1}\right) &\leq \mathbb{E}_{f(\mathrm{y_t}|\mathbf{x},\mathrm{t})} \log\left(\frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}\right) \leq \log\left(\Lambda\right), \\
\log\left(\Lambda^{-1}\right) &\leq \int_{\mathcal{Y}} \log\left(\frac{dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}\right) dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) \leq \log\left(\Lambda\right), \\
\log\left(\Lambda^{-1}\right) &\leq -D_{\mathrm{KL}}\left(P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) \| P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})\right) \leq \log\left(\Lambda\right), \\
\left|\log\left(\Lambda\right)\right| &\geq D_{\mathrm{KL}}\left(P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) \| P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})\right).
\end{aligned}
$$

That is, the KL-divergence between $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ is upper-bounded by the absolute value of $\log \Lambda$.

5.7 Applications

5.7.1 Scalable Sensitivity Analysis for Conditional Average Treatment Effects

5.7.1.1 Experiments

In this section we evaluate our methods using synthetic and semi-synthetic datasets. To assess our method on high-dimensional data, we introduce a new benchmark dataset, HC-MNIST. To illustrate how our uncertainty aware bounds can be used for deferring treatment, we introduce a hidden confounding variant of the IHDP dataset [Hil11]. Details about the data generating processes including dataset links, code links, and validation splitting procedures are given in Appendices B.1.1, B.2.2 and B.4.2.

The sampling procedures outlined in Section 5.3.1 and Section 5.4 for the estimators in Definitions 5.7 and 5.8 requires models for $p(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ and the nominal propensity $e(\mathbf{x})$. We use a mixture density network for $p(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ and a standard neural network with categorical likelihood for $\widehat{e}(\mathbf{x}; \boldsymbol{\theta}_n)$. Deep Ensembles [LPB17] are used to approximate sampling $\boldsymbol{\theta} \sim p(\boldsymbol{\Theta} \mid \mathcal{D})$. In general, modelling $p(\boldsymbol{\Theta} \mid \mathcal{D})$ is a choice to be made by the practitioner, for example, by using Bayesian Neural Networks or simpler Bayesian models for $p(\mathrm{Y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$. Details for each experiment, including

architectures, hyper-parameter tuning, training procedures, and compute infrastructure are given by Jesson et al. [Jes+21b] with code to reproduce experiments available at `https://github.com/OATML/quince`.

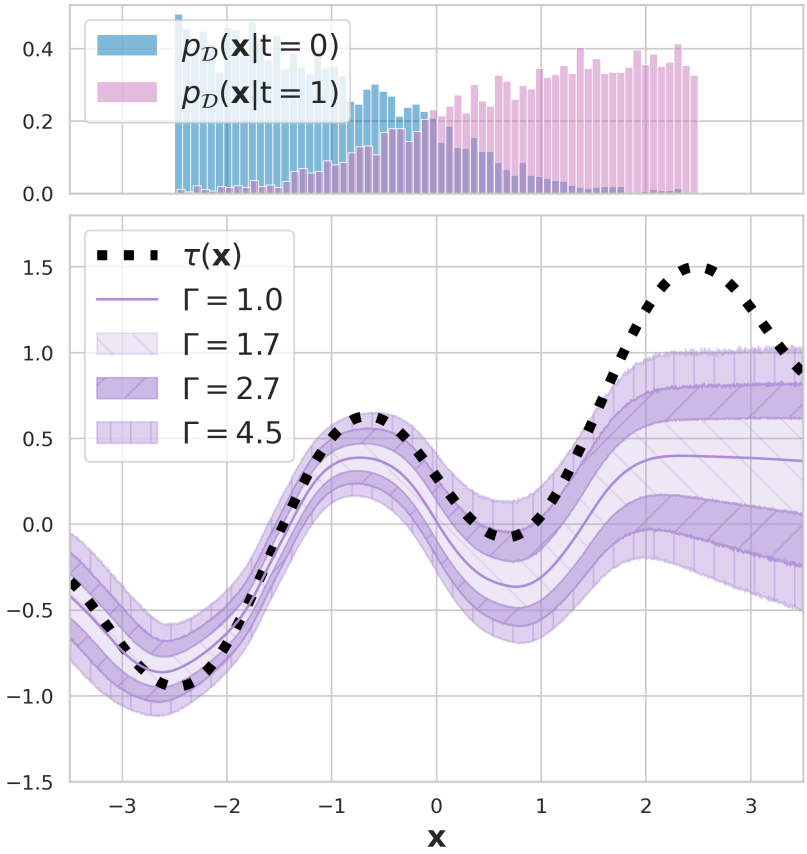


Figure 5.5: Varying Γ for Marginal Sensitivity Model. Ground truth $\Gamma^* = 2.7$. While the bounds follow the true CATE $\tau(x)$ on the support of $p_{\mathcal{D}}(\mathbf{x})$, they become nonsensical for out-of-distribution data ($\mathbf{x} < -2.5$ and $\mathbf{x} > 2.5$) and when there is a lack of overlap.

Simulated Data We first consider the one-dimensional example introduced by Kallus, Mao, and Zhou [KMZ19] Appendix B.1.1. Figure 5.5, generated with $n = 10000$ and $\log \Gamma^* = 1$, illustrates the nonlinear CATE function of these data. This is a useful example because both the CATE and the bias induced by hidden confounding are heterogeneous in $\mathbf{x}$. Further, Figure 5.5 shows that our estimator converges to tight bounds on the CATE interval for varying choices of Γ, achieving coverage when the assumed Γ matches the true value Γ^* used to generate the data. For this experiment and the next we assume that the outcomes correspond to costs, so that we aim to treat when $\tau(\mathbf{x}) \leq 0$.

For a quantitative evaluation, we use the same minimax-optimal policy as Kallus, Mao, and Zhou [KMZ19], namely, $\pi^*(\mathbf{x};\Gamma) = \mathbb{I}(\overline{\tau}(\mathbf{x};\Gamma) \leq 0) + \pi_0(\mathbf{x})\mathbb{I}(\underline{\tau}(\mathbf{x};\Gamma) < 0 < \overline{\tau}(\mathbf{x};\Gamma))$. This says that the optimal policy always treats when $\overline{\tau}(\mathbf{x};\Gamma) \leq 0$ and otherwise reverts to the default policy $\pi_0(\mathbf{x})$. Setting $\pi_0(\mathbf{x}) = 0$, *do not treat*, our approximation to the optimal policy is given by $\widehat{\overline{\pi}}(\mathbf{x};\Gamma) = \mathbb{I}(\widehat{\overline{\tau}}(\mathbf{x};\Gamma) \leq 0)$. The risk associated with a given policy is defined as $V(\pi;\tau) = \mathbb{E}[\pi(\mathbf{x})\mathrm{Y}_1 + (1-\pi(\mathbf{x}))\mathrm{Y}_0]$. Intuitively, policy risk will be minimized when $\widehat{\overline{\tau}}(\mathbf{x})$ is aligned exactly with the true CATE $\tau(\mathbf{x})$, and any deviations between $\widehat{\overline{\tau}}(\mathbf{x})$ and $\tau(\mathbf{x})$ will result in a higher policy risk score. To compare different methods on a finite sample, we report the *Policy Risk Error* as the mean squared error between the risk of the optimal treatment policy $\mathbb{I}(\tau(x) < 0)$, and the policy risk of a given policy π.

Table 5.1: **Simulated Data:** Policy risk errors for various policies under data generating processes with different Γ^*. Average test-set policy risk errors and 95% confidence intervals over 50 randomly generated datasets are reported. Statistically significant improvements for "well-specified" $\Gamma = \Gamma^*$, as determined by a paired t-test (1% threshold), shown in green. Policy risk errors are multiplied by 100 for readability.

Method ($n = 1000$)	$\log \Gamma^*$	$\hat{\pi}(\mathbf{x};\exp(0.5))$	$\hat{\pi}(\mathbf{x};\exp(1.0))$	$\hat{\pi}(\mathbf{x};\exp(1.5))$
Proposed	0.5	$\mathbf{0.07 \pm 0.03}$	0.28 ± 0.03	0.38 ± 0.04
	1.0	0.71 ± 0.20	$\mathbf{0.10 \pm 0.04}$	0.31 ± 0.03
	1.5	3.99 ± 0.59	0.75 ± 0.18	$\mathbf{0.13 \pm 0.04}$
[KMZ19]	0.5	$\mathbf{0.10 \pm 0.04}$	0.44 ± 0.09	0.99 ± 0.38
	1.0	0.48 ± 0.19	$\mathbf{0.25 \pm 0.11}$	0.81 ± 0.39
	1.5	3.33 ± 0.61	0.52 ± 0.19	$\mathbf{0.52 \pm 0.40}$

In Table 5.1, we compare the Policy Risk Error of our method to the one proposed by Kallus, Mao, and Zhou [KMZ19]. The average and 95% confidence intervals over 50 random realizations of training ($n = 1000$), validation ($n = 100$), and test ($n = 1000$) datasets are reported. On the diagonals we assess each policy and method with a "well-specified" $\Gamma = \Gamma^*$. These results show empirical evidence for the tightness of our interval estimator's bounds, and improved accuracy w.r.t. Kallus, Mao, and Zhou [KMZ19] on this low-dimension problem.

HC-MNIST: Hidden Confounding and High-dimensional Data

For this experiment, we adopt the one-dimensional simulated setting into a high-dimensional setting Appendix B.4.2. Specifically, we assign to each image of the MNIST dataset [LeC98] a latent feature $\phi \in [-2, 2]$ as follows: all images of the digits 0 are assigned a $\phi \in [-2, -1.6]$, all images 1 have $\phi \in [-1.6, -1.2]$, and so on

Table 5.2: **HC-MNIST:** Policy risk for various policies under data generating processes with different Γ^*. The proposed method approaches the ideal policy value of -1.41 under optimal policy given the true CATE. Average test-set policy risk errors and 95% confidence intervals over 20 randomly generated datasets are reported. This shows that our method scales well to large-sample, high-dimensional datasets.

	Proposed Method		
$\log \Gamma^*$	$\hat{\pi}(\mathbf{x}; \exp(0.5))$	$\hat{\pi}(\mathbf{x}; \exp(1.0))$	$\hat{\pi}(\mathbf{x}; \exp(1.5))$
0.5	**-1.40 ± 0.01**	-1.36 ± 0.01	-1.35 ± 0.01
1.0	-1.32 ± 0.02	**-1.40 ± 0.01**	-1.36 ± 0.01
1.5	-1.98 ± 0.02	-1.30 ± 0.02	**-1.38 ± 0.01**

up to the digit 9. The images of every digit are sorted by brightness and ordered equally within the interval of ϕ values assigned to images of that digit. Finally, these one-dimensional hidden values ϕ are used as the inputs to the same model of hidden confounding introduced by Kallus, Mao, and Zhou [KMZ19] and used in the simulated data experiments above. We report the results of our method in Table 5.2, showing it achieves near optimal policy risk under the true level of hidden confounding. We do not to report results for Kallus, Mao, and Zhou [KMZ19] here as their kernel based method did not scale well to the full dataset size of MNIST, and it did not give sensible results when training only on a subset of the dataset.

IHDP Hidden Confounding

In this section we demonstrate how our uncertainty-aware interval estimator can be used to inform deferral policies for treatment recommendations. To this end we use the IHDP dataset [Hil11] as Jesson et al. [Jes+20] show that low overlap and/or similarity are problems for IHDP. For insufficient context, we induce hidden confounding by hiding covariate x_9 during model training and CATE estimation; however, it is still used for the generation of synthetic observed outcomes as per the response surface B described by Hill [Hil11] Appendix B.2.2.

In contrast to the above experiments, treatment $T = 1$ is recommended if and only if $\tau(\mathbf{x}) > 0$; we propose a deferral policy that simulates deferral to an expert and withholds a recommendation if the predicted CATE interval intersects 0. We select Γ_s such that the uncertainty aware CATE interval $[\widehat{\underline{\tau}}(\mathbf{x}; \Gamma_s), \widehat{\overline{\tau}}(\mathbf{x}; \Gamma_s)]$ crosses 0. We then defer predictions with the lowest Γ_s value; these are predictions the model is least sure about. We compare using the same policy for the Kallus, Mao, and Zhou [KMZ19] method, and to the epistemic uncertainty based method proposed by Jesson et al.

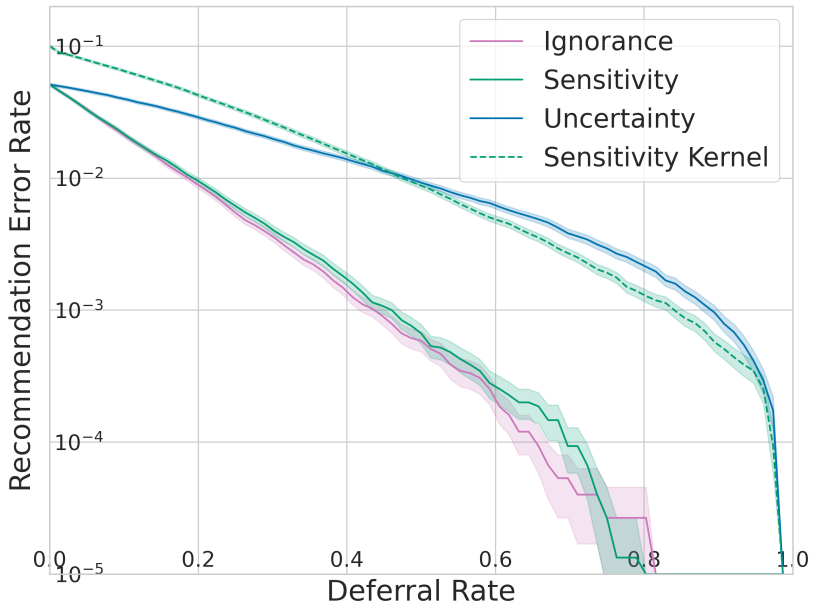


Figure 5.6: **IHDP Hidden Confounding:** Error rate as we sweep over the percentage of deferred points. We propose that recommendations should be deferred when there is ignorance. On the x-axis we vary the share of recommendations deferred, simulating various levels of practitioner caution. *Ignorance* (ours) accounts for all lack of knowledge. *Uncertainty* [Jes+20] accounts only for insufficient similarity and overlap. *Sensitivity* only accounts for hidden confounding, without accounting for insufficient similarity and overlap. *Sensitivity Kernel* is the kernel method of Kallus, Mao, and Zhou [KMZ19], which does not account for other sources of ignorance. Results show that all sources of ignorance are important on IHDP with one hidden confounder.

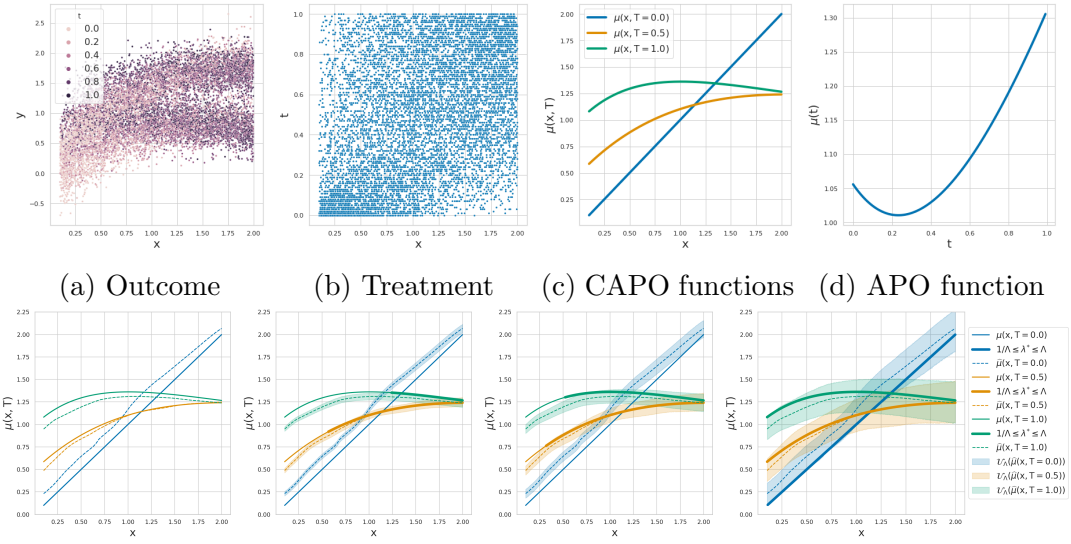
[Jes+20]. We report the error rate between recommendations given by $\mathbb{I}(\tau(\mathbf{x}) > 0)$ and $\mathbb{I}(\widehat{\tau}(\mathbf{x}) > 0)$ on the remaining recommendations that were not deferred.

In Figure 5.6, we see that the epistemic *uncertainty* policy (blue solid line) has a moderate decrease in error rate as the rate of deferral increases. The green solid *sensitivity* line shows that the error rate decreases as we defer recommendations based only on levels of hidden confounding. We should see the same behavior for the sensitivity method (green dashed line) proposed by Kallus, Mao, and Zhou [KMZ19], but it appears to struggle for higher dimensional covariates. The purple solid *ignorance* line shows that using the uncertainty aware CATE interval further improves results, showing that our method can account for all sources of ignorance discussed.

5.7.2 Scalable Sensitivity Analysis for CAPOs.

5.7.2.1 Experiments

Here we empirically validate our method for scalable sensitivity and uncertainty analysis for continuous valued interventions. First, we consider a synthetic structural causal model (SCM) to demonstrate the validity of our method. Next, we show the scalability of our methods by applying them to a real-world climate-science-inspired problem. Details for each experiment, including architectures, hyper-parameter tuning, training procedures, and compute infrastructure are given by Jesson et al. [Jes+22] with code to reproduce experiments available at `https://github.com/OATML/overcast`.



(a) Outcome (b) Treatment (c) CAPO functions (d) APO function

(e) Λ=1.0 (f) Λ=1.1 (g) Λ=1.2 (h) Λ=1.6

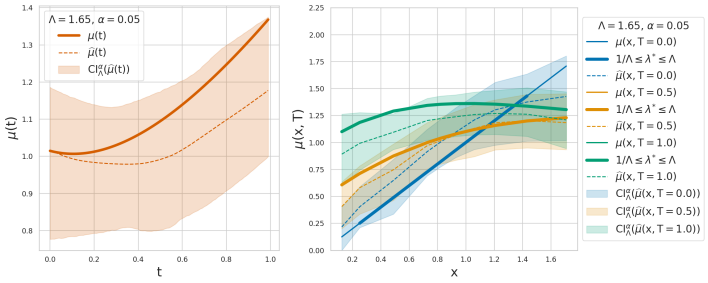
Figure 5.7: Figures 5.7a to 5.7d: Synthetic data and ground truth functions. Figures 5.7e to 5.7h Causal uncertainty under hypothesized Λ values. Solid lines are ground truth; thick solid lines where the true λ^* is within the range of hypothesized Λ, thin solid lines otherwise. The dotted lines are the estimated CAPO. Shaded regions are estimated CMSM intervals.

Synthetic Figure 5.7 presents the synthetic dataset (additional details about the SCM are given in Appendix B.1.2). Figure 5.7a plots the observed outcomes, y, against the observed confounding covariate, x. Each datapoint is colored by the magnitude of the observed treatment, t. The binary unobserved confounder, u, induces a bi-modal distribution in the outcome variable, y, at each measured value, x. Figure 5.7b plots the assigned treatment, t, against the observed confounding covariate, x. We can see that the coverage of observed treatments, t, varies for each value of x. For example, there is uniform coverage at X = 1, but low coverage for high treatment

values at X = 0.1, and low coverage for low treatment values at X = 2.0. Figure 5.7c plots the true CAPO function over the domain of observed confounding variable, X, for several values of treatment (T = 0.0, T = 0.5, and T = 1.0). For lower magnitude treatments, t, the CAPO function becomes more linear, and for higher values, we see more effect heterogeneity and attenuation of the effect size as seen from the slope of the CAPO curve for T = 0.5 and T = 1.0. Figure 5.7d plots the the APO function over the domain of the treatment variable T.

Structural Uncertainty We want to show that in the limit of large samples (we set n to $100k$), the bounds on the CAPO and APO functions under the CMSM include the ground truth when the CMSM is correctly specified. That is, when $1/\Lambda \leq \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) \leq \Lambda$, for user specified parameter Λ, the estimated intervals should cover the true CAPO or APO. This is somewhat challenging to demonstrate as the true density ratio $\lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u})$ (Eq. (B.4)), varies with t, x, and u. Figures 5.7e to 5.7h work towards communicating this. In Figure 5.7e, we see that each predicted CAPO function (dashed lines) is biased away from the true CAPO functions (solid lines). We use thick solid lines to indicate cases where $1/\Lambda \leq \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) \leq \Lambda$, and thin solid lines otherwise. Therefore thick solid lines indicate areas where we expect the causal intervals to cover the true functions. Under the erroneous assumption of ignorability ($\Lambda = 1$), the CMSM bounds have no width. In Figure 5.7f, we see that as we relax our ignorability assumption ($\Lambda = 1.1$) the intervals (shaded regions) start to grow. Note the thicker orange line: this indicates that for observed data described by X > 0.5 and T $= 0.5$, the actual density ratio is in the bounds of the CMSM with parameter $\Lambda = 0.5$. We see that our predicted bounds cover the actual CAPO function for these values. We see our bounds grow again in Figure 5.7g when we increase Λ to 1.2. We see that more data points have λ^* values that lie in the CMSM range and that our bounds cover the actual CAPO function for these values. In Figure 5.7h we again increase the parameter of the CMSM. We see that the bounds grow again, and cover the true CAPO functions for all of the data that satisfy $1/\Lambda \leq \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) \leq \Lambda$.

Statistical Uncertainty Now we relax the infinite data assumption and set $n = 1000$. This decrease in data will increase the estimator error for the CAPO and APO functions. So the estimated functions will not only be biased due to hidden confounding, but they may also be erroneous due to finite sample variance. We show this in Figure 5.8b where the blue dashed line deviates from the actual blue solid line as $\mathbf{x}$ increases beyond 1.0. However, Figure 5.8b shows that under the correct CMSM, the uncertainty aware confidence intervals, Section 2.6, cover the actual CAPO functions



(a) APO Function (b) CAPO Functions

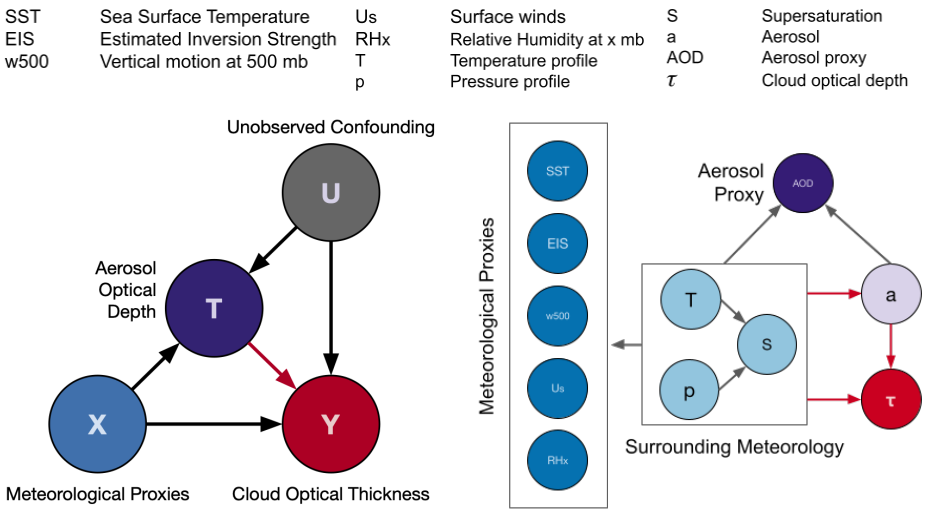
Figure 5.8: Statistical and causal uncertainty, α is statistical significance level for the bootstrap. see Figure 5.7 for other details.

for the range of treatments considered. Figure 5.8a demonstrates that this holds for the APO function as well.

5.7.3 Estimating Aerosol-Cloud-Climate Effects from Satellite Data

Background The development of the model above, and the inclusion of treatment as a continuous variable with multiple, unknown confounders, is motivated by a real-life use case for a prime topic in climate science. Aerosol-cloud interactions (ACI) occur when anthropogenic emissions in the form of aerosol enter a cloud and act as cloud condensation nuclei (CCN). An increase in the number of CCN results in a shift in the cloud droplets to smaller sizes which increases the brightness of a cloud and delays precipitation, increasing the cloud's lifetime, extent, and possibly thickness [Two77; Alb89; Tol+17]. However, the magnitude and sign of these effects are heavily dependent on the environmental conditions surrounding the cloud [DL20]. Clouds remain the largest source of uncertainty in our future climate projections [Mas+21]; it is pivotal to understand how human emissions may be altering their ability to cool. Our current climate models fail to accurately emulate ACI, leading to uncertainty bounds that could offset global warming completely or double the effects of rising CO_2 [Bou+13].

Defining the Causal Relationships Clouds are integral to multiple components of the climate system, as they produce precipitation, reflect incoming sunlight, and can trap outgoing heat [SF09]. Unfortunately, their interconnectedness often leads to hidden sources of confounding when trying to address how anthropogenic emissions alter cloud properties.



(a) (b)

Figure 5.9: Causal diagrams. Figure 5.9a, a simplified causal diagram representing what we are reporting within; aerosol optical depth (AOD, regarded as the treatment T) modulates cloud optical depth (τ, Y), which itself is affected by hidden confounders (U) and the meteorological proxies (X). Figure 5.9b, an expanded causal diagram of ACI. The aerosol (a) and aerosol proxy (AOD), the true confounders (light blue), their proxies (dark blue), and the cloud optical depth (red).

Ideally, we would like to understand the effect of aerosols (a) on the cloud optical thickness, denoted τ. However, this is currently impossible. Aerosols come in varying concentrations, chemical compositions, and sizes [Sch+16] and we cannot measure these variables directly. Therefore, we use aerosol optical depth (AOD) as a continuous, 1-dimensional proxy for aerosols. Figure 5.9b accounts for the known fact that AOD is an imperfect proxy impacted by its surrounding meteorological environment [Chr+17]. The meteorological environment is also a confounder that impacts cloud thickness τ and aerosol concentration a. Additionally, we depend on simulations of the current environment in the form of reanalysis to serve as its proxy.

Here we report AOD as a continuous treatment and the environmental variables as covariates. However, aerosol is the actual treatment, and AOD is only a confounded, imperfect proxy (Figure 5.9a). This model cannot accurately capture all causal effects and uncertainty due to known and unknown confounding variables. We use this simplified model as a test-bed for the methods developed within this paper and as a demonstration that they can scale to the underlying problem. Future work will tackle the more challenging and realistic causal model shown in Figure 5.9b, noting that the treatment of interest a is multi-dimensional and cannot be measured directly.

Model We use daily observed $1^\circ \times 1^\circ$ means of clouds, aerosol, and the environment from sources shown in Table B.4 of Appendix B.5. To model the spatial correlations between the covariates on a given day, we use multi-headed attention [Vas+17] to define a transformer-based feature extractor. Modeling the spatial dependencies between meteorological variables is motivated by confounding that may be latent in the relationships between neighboring variables. These dependencies are unobserved from the perspective of a single location. This architectural change respects both the assumed causal graph (Fig. 5.9a) and some of the underlying physical causal structure. We see in Figure 5.10 (Left) that modeling *context* with the transformer architecture significantly increases the predictive accuracy of the model when compared to a simple feed-forward neural network (*no context*). **Discussion & Results** The results for the APO of cloud optical depth (τ) as the "treatment", AOD, increases are shown in Figure 5.10. As the assumed strength of confounding increases ($\Lambda > 1$), the range of uncertainty in the treatment outcome increases. Within this range of confounding, the modeled outcomes agree with two conflicting hypotheses. First, that aerosol acts to invigorate the cloud, inducing a large response that would follow a maximum curve within this uncertainty range [CS11; DL21]. And second, that aerosol has little impact on cloud depth, and the actual response is a minimal, flat line [Gry+19]. We further find the reported dose-response curves in agreement with multiple estimates of aerosol-cloud interactions using satellite observations [BTG02; Myh+07; Tol+19]. The upper bound for $\log \Lambda = .2$ agrees with measurements of the in-cloud environment and aerosol-cloud interactions from aircraft-mounted sensors [PZ13], this may indicate the need for additional control variables when using satellite data.

The resolution of the satellite observations ($1^\circ \times 1^\circ$ daily means) could be averaging various cloud types and obscuring the signal. Future work will investigate how higher resolution (20km $\times$ 20km) data with constraints on cloud type may resolve some confounding influences. However, even our more detailed causal model (Figure 5.9b) cannot account for all confounders; we expected, and have seen, imperfections in our model of this complex effect. The model's results require further expert validation to interpret the outcomes and uncertainty.

Societal Impact Geoengineering of clouds by aerosol seeding could offset some amount of warming due to climate change, but also have disastrous global impacts on weather patterns [Dia+22]. Given the uncertainties involved in understanding aerosol-cloud interactions, it is paramount that policy makers are presented with projected outcomes if a proposals assumptions are wrong or relaxed.

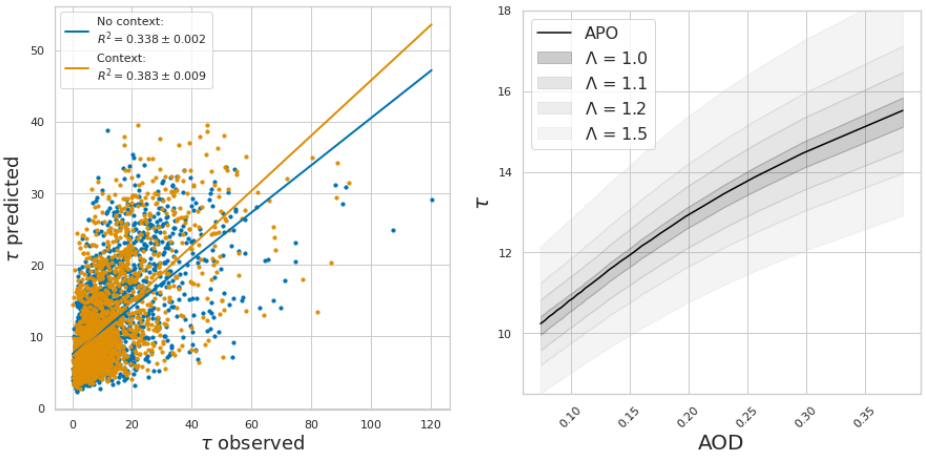


Figure 5.10: Left: The values of the observed, true τ against the modeled τ. Right: The curve for continuous treatment outcome of our aerosol proxy (AOD) on cloud optical depth (τ). The darkest shaded region ($\Lambda = 1$) represents the uncertainty in the treatment outcome from the ensemble due to finite data. As the strength of confounders increases ($\Lambda > 1.0$), the range of uncertainty in the treatment outcome increases.

Chapter 6

Conclusions

This thesis presents new methodologies for quantifying statistical and structural uncertainty for conditional causal-effect inference that scale to large observational datasets with complex input modalities. Our research demonstrates progress in both the development of novel methodologies and the application of scalable statistical and structural uncertainty-aware causal machine learning.

Scalable statistical uncertainty quantification has seen significant progress since the publication of the works discussed here. De Brouwer, Gonzalez, and Hyland [DGH22] and Hess et al. [Hes+23] have introduced new approaches to quantify statistical uncertainty using causal neural differential equation models, while Wen et al. [Wen+23] proposed a similar method tailored for graph neural networks. Additionally, Kivlichan et al. [Kiv+21] applied statistical uncertainty-aware machine learning methodologies for decision support in content moderation, highlighting the broad applicability of these techniques. And Durso-Finley et al. [Dur+23] leverage scalable uncertainty-aware causal ML methods to predict the progression of multiple sclerosis lesions from patient-level MRI scans. They demonstrate the potential of such methodology in medical applications. For future work, we believe that the Bayesian perspective on in-context learning [Bro+20; Mül+21; Xie+21; FHW23; Lee+23] presents an opportunity for new methodologies in scalable statistical uncertainty quantification for causal-effect estimation using pre-trained conditional generative models.

Moving beyond the observational setting, using statistical uncertainty to inform experimental design is also an active area of study. Aglietti et al. [Agl+20; Agl+21; Agl+23] and Zhang et al. [Zha+23] propose Bayesian optimization frameworks to learn causal effects when the structural causal model is known. Tigas et al. [Tig+22], Toth et al. [Tot+22], Tigas et al. [Tig+23], Branchini et al. [Bra+23], and Deleu et al.

[Del+24] have developed Bayesian methods to inform which intervention to apply to efficiently learn both the functional and structural relationships of a causal model given a directed acyclic graph skeleton. Interesting future work along this trajectory would be using experimental design to learn the latent causal factors when they are not given *a priori*.

The development of methodologies for quantifying structural uncertainty via sensitivity analysis for conditional average treatment effects has been a particularly active area of research [Pad+22; SGC22; DG23; DY23; Fra+23a; Opr+23; Fra+23b; Mar+23; Yin+24; FMF24; MFF24; Mar+24; DGK24]. Noteworthy is the effort by Marmarelis et al. [Mar+24] to account for both structural and statistical uncertainty in their methodology.

This work has focused on the three problems of causal-effect inference, decision-making, and model updating in isolation. The domain of sequential decision-making (reinforcement learning) combines these three aspects into a single process. Jesson et al. [Jes+24] is our exploration into incorporating what we have learned from the projects presented in this thesis into sequential decision-making, albeit in a non-causal setting. We think that causal reinforcement learning [ZB20] is an exciting avenue of ongoing research for scalable statistical and structural uncertainty in causal machine learning.

Appendix A

Publications and Pre-prints

A.1 Published Works

This thesis presents work from the following four publications:

1. Andrew Jesson*, Sören Mindermann*, Uri Shalit, and Yarin Gal. "Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models." *NeurIPS*. (2020).

2. Andrew Jesson*, Panagiotis Tigas*, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. "Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data." *NeurIPS*. (2021).

3. Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. "Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding." *ICML*. (2021).

4. Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. "Scalable Sensitivity and Uncertainty Analysis for Causal-Effect Estimates of Continuous-Valued Interventions." *NeurIPS*. (2022).

Additional works the author has contributed to on the topic of scalable statistical uncertainty are:

1. Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. "GeneDisco: A Benchmark for Experimental Design in Drug Discovery." *ICLR*. (2022).

2. Panagiotis Tigas*, Yashas Annadani*, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. "Interventions, Where and How? Experimental Design for Causal Models at Scale." *NeurIPS*. (2022).

3. Clare Lyle*, Arash Mehrjou*, Pascal Notin*, Andrew Jesson, Stefan Bauer, Yarin Gal, Patrick Schwab. "DiscoBAX-Discovery of optimal intervention sets in genomic experiment design." *ICML*. (2023).

4. Yashas Annadani*, Panagiotis Tigas*, Desi R Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, Stefan Bauer. "Differentiable Multi-Target Causal Bayesian Experimental Design." *ICML*. (2023).

Additional works the author has contributed to on the topic of scalable structural uncertainty are:

1. Andrew Jesson*, Peter Manshausen*, Alyson Douglas*, Duncan Watson-Parris, Yarin Gal, and Philip Stier. "Using Non-Linear Causal Models to Study Aerosol-Cloud Interactions in the Southeast Pacific." *Causal Inference and Machine Learning: Why now? Workshop at NeurIPS*. (2021).

2. Maëlys Solal, Andrew Jesson, Yarin Gal, Alyson Douglas. "Using uncertainty-aware machine learning models to study aerosol-cloud interactions." *Tackling Climate Change with Machine Learning: workshop at NeurIPS*. (2022).

3. Myrl G Marmarelis, Elizabeth Haddad, Andrew Jesson, Neda Jahanshad, Aram Galstyan, Greg Ver Steeg. "Partial identification of dose responses with hidden confounders." *UAI*. (2023).

4. Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, Uri Shalit. "B-Learner: Quasi-Oracle Bounds on Heterogeneous Causal Effects Under Hidden Confounding." *ICML*. (2023).

Appendix B

Datasets

B.1 Simulated Datasets

B.1.1 Binary Treatment, Continuous Outcome

The simulated dataset presented by Kallus, Mao, and Zhou [KMZ19] is described by the following structural causal model (SCM):

$$\mathrm{u} \coloneqq N_{\mathrm{u}}, \tag{B.1a}$$

$$\mathrm{x} \coloneqq N_{\mathrm{x}}, \tag{B.1b}$$

$$\mathrm{t} \coloneqq N_{\mathrm{t}}, \tag{B.1c}$$

$$\mathrm{y} \coloneqq (2\mathrm{t}-1)\mathrm{x} + (2\mathrm{t}-1) - 2\sin(2(2\mathrm{t}-1)\mathrm{x}) - 2(2\mathrm{u}-1)(1+0.5\mathrm{x}) + N_{\mathrm{y}}, \tag{B.1d}$$

where $N_{\mathrm{u}} \sim \mathrm{Bern}(0.5)$, $N_{\mathrm{x}} \sim \mathrm{Unif}[-2,2]$, $N_{\mathrm{u}} \perp\!\!\!\perp N_{\mathrm{x}}$, $N_{\mathrm{t}} \sim \mathrm{Bern}(e(\mathrm{x},\mathrm{u}))$, $e(\mathrm{x},\mathrm{u}) = \frac{\mathrm{u}}{\alpha_{\mathrm{t}}(\mathrm{x};\Gamma^*)} + \frac{1-\mathrm{u}}{\beta_{\mathrm{t}}(\mathrm{x};\Gamma^*)}$, $e(\mathrm{x}) = \mathrm{sigmoid}(0.75\mathrm{x}+0.5)$, and $N_{\mathrm{y}} \sim \mathcal{N}(0,1)$.

Remember that only x, t, and y are observed. So the bias induced in the CATE estimate by hidden confounding at x is given by

$$\tilde{\delta}(\mathrm{x}) - \delta(\mathrm{x}) = 2(2+\mathrm{x})\left(P(\mathrm{u}=1 \mid \mathrm{T}=1, \mathrm{X}=\mathrm{x}) - P(\mathrm{u}=1 \mid \mathrm{T}=0, \mathrm{X}=\mathrm{x})\right), \tag{B.2}$$

where $\tilde{\delta}(\mathrm{x})$ is the confounded CATE estimate.

Each random realization of the simulated dataset generates 1000 training examples, 100 validation examples, and 1000 test examples. In the experiments we report results over 50 random realizations. The seeds for the random number generators are i, $i+1$, and $i+2$; $\{i \in [0, 1, \ldots, 49]\}$, for the training, validation, and test sets, respectively. Code is available in file /library/datasets/synthetic.py on github at `https://github.com/anndvision/quince`.

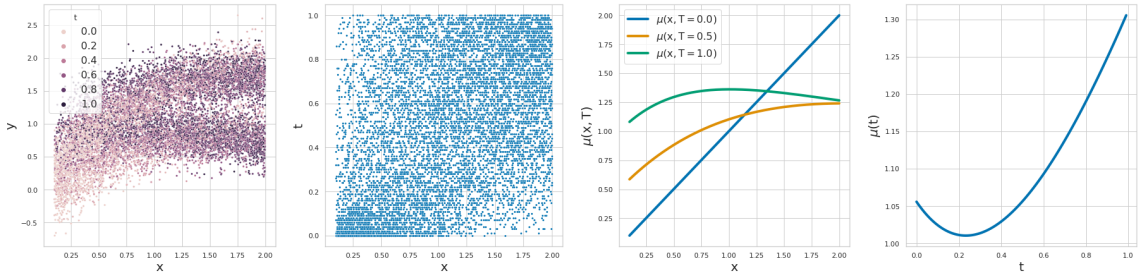
B.1.2 Continuous Treatment, Continuous Outcome

$$\begin{aligned} \mathrm{u} &:= N_{\mathrm{u}}, \\ \mathrm{x} &:= N_{\mathrm{x}}, \\ \mathrm{t} &:= N_{\mathrm{t}}, \\ \mathrm{y_t} &:= \mathrm{t} + \mathbf{x}\exp(-\mathrm{tx}) - \gamma_{\mathrm{y}}(\mathrm{u} - 0.5) * (0.5 * \mathrm{x} + 1) + N_{\mathrm{y}}, \end{aligned} \tag{B.3}$$

where, $N_{\mathrm{u}} \sim p(\mathrm{u}) := \mathrm{Bern}(\mathrm{u} \mid 0.5)$, $N_{\mathrm{x}} \sim p(\mathrm{x}) := \mathrm{Unif}[\mathrm{x} \mid 0.1, 2.0]$, $N_{\mathrm{t}} \sim p(\mathrm{t} \mid \mathrm{x}, \mathrm{u}) := \text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}, \beta = 1)$, and $N_{\mathrm{y}} \sim \mathcal{N}(0, 0.04)$. For the results in this paper $\gamma_t = 0.3$ and $\gamma_y = 0.5$.

The ground truth ratio, $\lambda = \frac{p(\mathrm{t}|\mathrm{x})}{p(\mathrm{t}|\mathrm{x},\mathrm{u})}$, is then given by,

$$\begin{aligned} \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) &= \frac{\mathbb{E}_{p(\mathrm{u})}[p(\mathrm{t} \mid \mathrm{x}, \mathrm{u})]}{p(\mathrm{t} \mid \mathrm{x}, \mathrm{u})} \\ &= \frac{\sum_{\mathrm{u}'=0}^{1} 0.5 * \text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}', \beta = 1)}{\text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}, \beta = 1)} \end{aligned} \tag{B.4}$$



(a) Outcome (b) Intervention (c) CAPO function (d) APO function

Figure B.1: Synthetic data with hidden confounding

B.2 IHDP

Hill [Hil11] use the The Infant Health Development Program (IHDP), a randomized experiment that assessed the impact of specialist home visits on children's performance in cognitive tests, to introduce a causal inference dataset. The treatment group reveives "intensive high-quality child care and home visits from a trained provider." Real covariates include: "measurements on the child–birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status–as well as behaviors engaged in during pregnancy–smoked cigarettes, drank alcohol, took drugs–and measurements on the mother at the time she gave birth–age, marital status, educational attainment (did not graduate from high school, graduated

from high school, attended some college but did not graduate, graduated from college), whether she worked during pregnancy, whether she received prenatal care–and the site (8 total) in which the family resided at the start of the intervention. There are 6 continuous covariates and 19 binary covariates." However, outcomes are simulated based on covariates and treatment, making this dataset semi-synthetic. Covariates were made different between the treatment and control groups by excluding "a non-random portion of the treatment group: all children with nonwhite mothers." There are 747 units in the dataset (139 treated, 608 control), with 25 covariates related to the children and their mothers. The IHDP dataset is available for download at `https://www.fredjo.com/`.

Table B.1: **IHDP Continuous Covariates**

Covariate	Description
x_1	birthweight
x_2	head circumference
x_3	number of weeks pre-term
x_4	birth order
x_5	"neo-natal health index"
x_6	mom's age

Following Shalit, Johansson, and Sontag [SJS17] and Hill [Hil11], we use the simulated outcome implemented as setting "B" in the NPCI package [Dor16] and we use the noiseless/expected outcome to compute the ground truth CATE. Response surface B, designed by Hill [Hil11], is described by the following SCM:

$$\mathbf{x} \coloneqq N_{\mathbf{x}}, \tag{B.5a}$$

$$\mathrm{t} \coloneqq N_{\mathrm{t}}, \tag{B.5b}$$

$$\mathrm{y} \coloneqq (\mathrm{t}-1)\left(\exp(\beta_{\mathbf{x}}(\mathbf{x}+\mathbf{w}))+N_{\mathrm{Y}_0}\right)+\mathrm{t}\left(\beta_{\mathbf{x}}\mathbf{x}-\omega^s+N_{\mathrm{Y}_1}\right), \tag{B.5c}$$

where $(N_{\mathbf{x}}, N_{\mathrm{t}}) \sim p_{\mathcal{D}}(\{\mathrm{x}_1, \ldots \mathrm{x}_{25}\}, \mathrm{t})$, $N_{\mathrm{Y}_0} \sim \mathcal{N}(0,1)$, and $N_{\mathrm{Y}_1} \sim \mathcal{N}(0,1)$. The coefficients $\beta_{\mathbf{x}}$ are a vector of randomly sampled values $(0.0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$. Hill [Hil11] describes ω^s as follows: "For the sth simulation, $[\omega^s]$ is chosen in the overlap setting, where we estimate the effect of the treatment on the treated [(CATT)], such that CATT equals 4; similarly it was chosen in the incomplete setting, where we estimate the effect of the treatment on the controls [(CATC)], so that CATC equals 4." An offset vector $\mathbf{w}$, equal in dimension to $\mathbf{x}$, with every value set to 0.5, is added to $\mathbf{x}$.

Table B.2: **IHDP Binary Covariates** Binary covariates $x_9 - \mathrm{x}_{18}$ are attributes of the mother. Mother's education level "College" indicated by covariates $\mathrm{x}_{10} - \mathrm{x}_{12}$ all zero. Site 8 indicated by covariates $\mathrm{x}_{19} - \mathrm{x}_{25}$ all zero. We show the frequency of occurrence for each binary covariate $p(\mathrm{x} = 1)$, as well as the adjusted mutual information $I(\mathrm{x}; \mathrm{t})$ between the binary covariate and the treatment variable.

Covariate	Description	$I(\mathrm{x}; \mathrm{t})$	$p(\mathrm{x} = 1)$
x_7	child's gender (female=1)	0.00	0.51
x_8	is child a twin	0.00	0.09
x_9	married when child born	**0.02**	**0.52**
x_{10}	left High School	0.00	0.36
x_{11}	completed High School	0.00	0.27
x_{12}	some College	0.00	0.22
x_{13}	child is first born	0.00	0.36
x_{14}	smoked cigarettes when pregnant	**0.01**	**0.48**
x_{15}	consumed alcohol when pregnant	0.00	0.14
x_{16}	used drugs when pregnant	0.00	0.96
x_{17}	worked during pregnancy	**0.01**	**0.59**
x_{18}	received any prenatal care	**0.01**	0.96
x_{19}	site 1	0.00	0.14
x_{20}	site 2	**0.01**	0.14
x_{21}	site 3	0.00	0.16
x_{22}	site 4	**0.01**	0.08
x_{23}	site 5	**0.02**	0.07
x_{24}	site 6	**0.01**	0.13
x_{25}	site 7	**0.02**	0.16

We run IHDP experiments according to the protocol described in [SJS17]: we run 1000 repetitions of the experiment, where each test set has 75 points and the remaining 672 available points are split 70% to 30% for training and validation. The ground truth outcomes are normalized to a mean of 0 and standard deviation of 1 over the training set. For evaluation, each model's predictions are unnormalized to calculate the PEHE.

B.2.1 IHDP Covariate Shift.

As previously mentioned, we selected a variable (marital status of mother) and exclude datapoints where the mother was unmarried from training (while leaving the test set unaltered). We selected this feature for three reasons: it is active in roughly 50% of data points, the distributions of the remaining covariates were distinct based on a T-SNE visualization [MH08], and the feature is only marginally correlated with

treatment (which ensures that we study the impact of covariate shift, not unobserved confounding). The feature is hidden to the models to make the detection of covariate shift non-trivial, and to induce a more realistic scenario where latent factors are often unaccounted for in observational data.

Marital status may be considered a sensitive socio-economic factor. We do not intend the experiment to be politically insensitive, rather that it emphasizes the problem of demographic exclusion in observational data due to issues such as historical bias, along with the danger of making confident but uninformed predictions when demographic exclusion is latent. Omitting these variables can lead to subpar model performance – particularly for members of a socio-economic minority.

B.2.2 IHDP Hidden Confounding

To induce hidden confounding, we need to select a variable u that is associated with the treatment that will be hidden from the CATE interval estimator, and design a response surface where the outcome will always be affected by u. In Table B.2, we list 3 potential candidates for u: x_9, x_{14}, and x_{17}. Each of these variable have a non-negligible association with the treatment, as indicated by the adjusted mutual information score $I(\mathrm{x};\mathrm{t})$, and have a frequency of taking the value 1 at around 0.5 (increasing the chances that we will have both positive and negative examples in each of the training, validation, and testing splits). Here we select x_9 and define the following SCM:

$$\mathrm{u} \coloneqq N_{\mathrm{u}}, \tag{B.6a}$$

$$\mathbf{x} \coloneqq N_{\mathbf{x}}, \tag{B.6b}$$

$$\mathrm{t} \coloneqq N_{\mathrm{t}}, \tag{B.6c}$$

$$\mathrm{y} \coloneqq (\mathrm{t}-1)(\exp(\beta_{\mathbf{x}}(\mathbf{x}+\mathbf{w}) + \beta_{\mathrm{u}}(\mathrm{u}+0.5)) + N_{\mathrm{Y}_0}) + \mathrm{t}(\beta_{\mathbf{x}}\mathbf{x} + \beta_{\mathrm{u}}\mathrm{u} - \omega^s + N_{\mathrm{Y}_1})), \tag{B.6d}$$

where $(N_{\mathrm{u}}, N_{\mathbf{x}}, N_{\mathrm{t}}) \sim p_{\mathcal{D}}(\mathrm{x}_9, \{\mathrm{x}_1, \ldots, \mathrm{x}_8, \mathrm{x}_{10}, \ldots, \mathrm{x}_{25}\}, \mathrm{t})$, $N_{\mathrm{Y}_0} \sim \mathcal{N}(0,1)$, and $N_{\mathrm{Y}_1} \sim \mathcal{N}(0,1)$. The coefficient β_{u} is randomly sampled from $(0.1, 0.2, 0.3, 0.4, 0.5)$ with probabilities $(0.2, 0.2, 0.2, 0.2, 0.2)$. The remaining parameters–$\beta_{\mathbf{x}}$, ω^s, and $\boldsymbol{\omega}$–are given as above, taking into account u.

For each random realization of the dataset, the IHDP data is split into training ($n = 470$), validation ($n = 202$) and test ($n = 75$) subsets using the scikit-learn

function train_test_split(). The random seeds for both splitting and outcome generation are $\{i \in [0, 1, \ldots, 999]\}$ for the 1000 realizations generated. Code to generate this dataset is available in file /library/datasets/ihdp.py on github at `https://github.com/anndvision/quince`.

B.3 ACIC 2016

Dorie et al. [Dor+19] introduced a dataset named after the 2016 Atlantic Causal Inference Conference (ACIC) where it was used for a competition. ACIC is a collection of semi-synthetic datasets whose covariates are taken from a large study conducted on pregnant women and their children to identifying causal factors leading to developmental disorders [Nis72]. There are 4802 observations and 58 covariates. Outcomes and treatments are simulated, as in IHDP, according to different data-generating process for each dataset. We chose this dataset instead of the 2018 ACIC challenge [Shi+18] because the latter is aimed at only ATE estimation and the CATE is equal for each observation in most datasets.

B.4 MNIST

B.4.1 CEMNIST Overlap

Table B.3: **CEMNIST-Overlap** Details of "Causal effect MNIST" dataset.

Digit(s)	Number of train samples	Number treated	Y_0	Y_1	$\delta(\mathbf{x})$
9	6000	≈ 666	1	0	-1
2	≈ 666	≈ 666	0	1	1
other odds	≈ 666 each	≈ 333 each	1	0	-1
other evens	≈ 666 each	≈ 333 each	0	1	1

The original MNIST [Den12] image dataset contains a training set of size 60000 and a test set of size 10000, where each digit class 0-9 represents 10% of points. We use a subset of the training data, shown in Table B.3. Similarly, we use a subset of the test set, with the same proportion for each digit as in the training set (and the same proportion of treated points). The variables Y^1, Y^0 are deterministic as shown in Table B.3. Some numbers in Table B.3 are approximate because they are generated according to the probabilities in Table 4.1.

The dataset serves two purposes. First, it illustrates why the standard practice of rejecting points with propensity scores close to 0 or 1 can be worse than rejecting

randomly. The digit 9 has the most data making it easy to predict the CATE, $\delta(\mathbf{x})$, but its propensity score is only 0.1, so that 9s will be rejected early. It might be a common situation in practice that a sub-population represents the majority of the data and therefore its CATE, $\delta(\mathbf{x})$, is easy to estimate. Second, the digit 2 suffers from strict non-overlap (propensity score of 1). It should be the first digit class to be rejected by any method since its CATE, $\delta(\mathbf{x})$, cannot be estimated. When increasing the rejected proportion, digits other than 9 should subsequently be rejected as only 334 and 333 examples are observed for their treatment and control groups respectively. However, propensity-based rejection is likely to retain these sub-populations because their propensity score is 0.5.

We repeated the CEMNIST experiment 20 times, each time generating a new dataset with a different random initialization for each model. Note that this is a single dataset, unlike other causal inference benchmarks, so it is only suited for CATE estimation, not ATE estimation.

B.4.2 HC-MNIST

HC-MNIST is an extension of the discrete treatment synthetic dataset in Appendix B.1.1 with high-dimensional covariates, $\mathbf{x}$. Specifically, the covariates are MNIST digits. HC-MNIST is described by the following SCM:

$$\mathrm{u} \coloneqq N_{\mathrm{u}}, \tag{B.7a}$$

$$\mathbf{x} \coloneqq N_{\mathbf{x}}, \tag{B.7b}$$

$$\phi \coloneqq \left(\mathrm{clip}\left(\frac{\mu_{N_{\mathbf{x}}} - \mu_{\mathrm{c}}}{\sigma_{\mathrm{c}}}; -1.4, 1.4\right) - \mathrm{Min}_{\mathrm{c}}\right) \frac{\mathrm{Max}_{\mathrm{c}} - \mathrm{Min}_{\mathrm{c}}}{1.4 - \text{-}1.4} \tag{B.7c}$$

$$\mathrm{t} \coloneqq N_{\mathrm{t}}, \tag{B.7d}$$

$$\mathrm{y} \coloneqq (2\mathrm{t} - 1)\phi + (2\mathrm{t} - 1) - 2\sin(2(2\mathrm{t} - 1)\phi) - 2(2\mathrm{u} - 1)(1 + 0.5\phi) + N_{\mathrm{y}}, \tag{B.7e}$$

where N_{u}, N_{t} (swapping x for ϕ), and N_{y} are as described in Appendix B.1.1. $N_{\mathbf{x}}$ is a sample of an MNIST image. The sampled image has a corresponding label $c \in [0, \ldots, 9]$. $\mu_{N_{\mathbf{x}}}$ is the average intensity of the sampled image. μ_{c} and σ_{c} are the mean and standard deviation of the average image intensities over all images with label c in the MNIST training set. In other words, $\mu_{\mathrm{c}} = \mathbb{E}[\mu_{N_{\mathbf{x}}} \mid \mathrm{c}]$ and $\sigma_{\mathrm{c}}^2 = \mathrm{Var}[\mu_{N_{\mathbf{x}}} \mid \mathrm{c}]$. To map the high dimensional images $\mathbf{x}$ onto a one-dimensional manifold ϕ with the same domain as $\mathrm{x} \in [-2, 2]$ above, we first clip the standardized average image intensity on the range $(-1.4, 1.4)$. Each digit class has its own domain in ϕ, so there

is a linear transformation of the clipped value onto the range $[\text{Min}_\text{c}, \text{Max}_\text{c}]$. Finally, $\text{Min}_\text{c} = -2 + \frac{4}{10}\text{c}$, and $\text{Max}_\text{c} = -2 + \frac{4}{10}(\text{c}+1)$.

For each random realization of the dataset, the MNIST training set is split into training ($n = 35000$) and validation ($n = 15000$) subsets using the scikit-learn function train_test_split(). The test set is generated using the MNIST test set ($n = 10000$). The random seeds are $\{i \in [0, 1, \ldots, 19]\}$ for the 20 random realizations generated. Code to generate this dataset is available in file /library/datasets/hcmnist.py on github at `https://github.com/anndvision/quince`.

B.4.3 CMNIST

Following the setup from [Jes+21b], we use a simulated dataset based on MNIST [LeC98]. CMNIST is described by the following SCM:

$$\mathbf{x} \coloneqq N_\mathbf{x}, \tag{B.8a}$$

$$\phi \coloneqq \left(\text{clip}\left(\frac{\mu_{N_\mathbf{x}} - \mu_\text{c}}{\sigma_\text{c}}; -1.4, 1.4\right) - \text{Min}_\text{c}\right)\frac{\text{Max}_\text{c} - \text{Min}_\text{c}}{1.4 - \text{-}1.4} \tag{B.8b}$$

$$\text{t} \coloneqq N_\text{t}, \tag{B.8c}$$

$$\text{y} \coloneqq (2\text{t}-1)\phi + (2\text{t}-1) - 2\sin(2(2\text{t}-1)\phi) + 2(1 + 0.5\phi) + N_\text{y}, \tag{B.8d}$$

where N_t (swapping x for ϕ), and N_y are as described in Appendix B.1. $N_\mathbf{x}$ is a sample of an MNIST image. The sampled image has a corresponding label $c \in [0, \ldots, 9]$. $\mu_{N_\mathbf{x}}$ is the average intensity of the sampled image. μ_c and σ_c are the mean and standard deviation of the average image intensities over all images with label c in the MNIST training set. In other words, $\mu_\text{c} = \mathbb{E}[\mu_{N_\mathbf{x}} \mid \text{c}]$ and $\sigma_\text{c}^2 = \text{Var}[\mu_{N_\mathbf{x}} \mid \text{c}]$. To map the high dimensional images $\mathbf{x}$ onto a one-dimensional manifold ϕ with domain $[-3, 3]$ above, we first clip the standardized average image intensity on the range $(-1.4, 1.4)$. Each digit class has its own domain in ϕ, so there is a linear transformation of the clipped value onto the range $[\text{Min}_\text{c}, \text{Max}_\text{c}]$. Finally, $\text{Min}_\text{c} = -2 + \frac{4}{10}\text{c}$, and $\text{Max}_\text{c} = -2 + \frac{4}{10}(\text{c}+1)$.

For each random realization of the dataset, the MNIST training set is split into training ($n = 35000$) and validation ($n = 15000$) subsets using the scikit-learn function train_test_split(). The test set is generated using the MNIST test set ($n = 10000$). The random seeds are $\{i \in [0, 1, \ldots, 19]\}$ for the 10 random realizations generated.

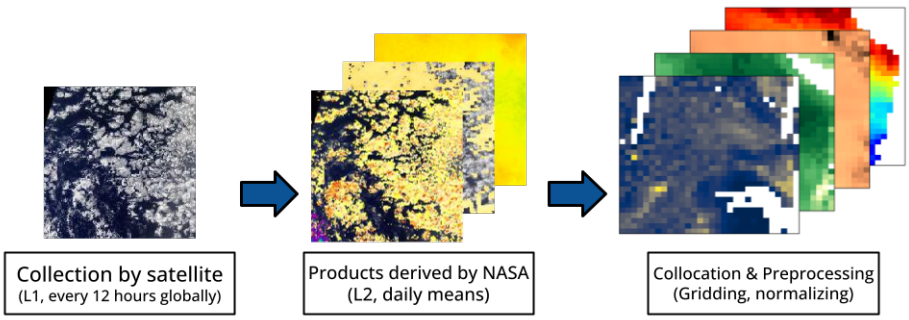


Figure B.2: Workflow of observed clouds from satellite to ingestion by model.

Table B.4: Sources of satellite observations.

Product name	Description
Cloud optical depth τ	MODIS (1.6, 2.1, 3.7 μm)
Precipitation	NOAA CMORPH
Sea Surface Temperature	NOAA WHOI
Vertical Motion	MERRA-2
Estimated Inversion Strength	MERRA-2
Relative Humidity	MERRA-2
Aerosol Optical Depth	MERRA-2

B.5 Satellite Climate Observations

The Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Aqua satellite observes the Earth twice daily at ∼1 km x 1 km resolution native resolution (Level 1) [BP06]. We used the daily mean, 1° x 1° gridded version (Level 2) in order to somewhat homogenize our observations of clouds and the atmosphere confined to a region off the coast of South America in the Pacific basin. MODIS observations are fed into the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) real-time model in order to emulate the atmosphere and it's components, such as aerosol [Gel+17]. Aerosol optical depth at 550nm from MERRA-2 is derived from MODIS observations of aerosol from multiple satellites (Terra, Aqua, Suomi-NPP), with corrections for sun glint and near-cloud optical effects [Bos+15]. We collocated all gridded observations of clouds and reanalysis aerosol with our meteorological proxies of the environment (EIS, SST, w500, RH700, RH850), then normalized our features before feeding them into the model.

Appendix C

Implementation Details

C.1 Deferral of Recommendations to An Expert

We evaluate our methods by considering *treatment recommendations*. A simplified recommendation strategy for an individual-level treatment of a unit with covariates x_i is to recommend $t = 1$ if the predicted $\widehat{\delta}(x_i)$ is positive, and $t = 0$ if negative. However, if there is insufficient knowledge about the CATE an individual, and a high cost associated with making errors, it may be preferable to withhold the recommendation, and e.g. refer the case for further scrutiny. It is therefore important to have an informed *rejection policy* for a treatment assigned based on a given CATE estimator.

To evaluate a rejection policy for a CATE estimator we assign a cost of 1 to making incorrect predictions and a cost of 0 for making a correct recommendation. At a fixed number of rejections, the utility of a policy can be defined as the inverse of the total number of erroneous recommendations made, i.e., if a policy can correctly identify the model's mistakes and refer such patients to a human expert then it should have a higher utility.

Rejection policies We introduce two rejection policies based on the epistemic and predictive uncertainty estimates of an uncertainty aware CATE estimator. Both policies opt to reject if the uncertainty estimate is greater than a threshold that rejects a given proportion of the training data r_{reject}. The training data is used since there may not be a large enough test set in practice. For all policies, we determine thresholds on the training set to simulate a real-world individual-level recommendation scenario. The *epistemic uncertainty* policy uses a sample-based estimator of the uncertainty in CATE (second *r.h.s.* term in Equation (4.5)) where M Monte Carlo samples are taken from each of $q(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1|\mathcal{D})$. Note that, for the T-Learner, this posterior factorizes into

two independent distributions $q(\boldsymbol{\theta}_0|\mathcal{D}), q(\boldsymbol{\theta}_1|\mathcal{D})$ because there are separate models for the outcome given treatment and no treatment. Furthermore, other models share parameters for $f_0(\cdot), f_1(\cdot)$ so the individual parameters in $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ may overlap. The *predictive* uncertainty policy uses an estimator of $\widehat{Var}_{pred}[Y^1 - Y^0|\mathbf{x}_i]$, which has the same functional form as Equation (4.5), but instead of being over the difference in expected values of the output distribution it is over samples of the output distribution.

We compare the utility of these policies to a random rejection baseline and two policies based on propensity scores. The first propensity policy (*propensity quantiles*) finds a two sided threshold on the distribution of estimated propensity scores such that a proportion $(1 - r_{\text{reject}})$ of the training data is retained. The second policy (*propensity trimming*) implements a trimming algorithm following the guidelines proposed by Caliendo and Kopeinig [CK08].

C.1.1 Models

We evaluate and compare each rejection policy using several uncertainty-aware CATE estimators. The estimators are the Bayesian versions of CEVAE [Lou+17], TARNet, CFR-MMD [SJS17], Dragonnet [SBV19], and a deep version of the T-Learner [SJS17]. Each model is augmented by introducing Bayesian parameter uncertainty and by predicting a distribution over model outputs. For image data, two convolutional bottom layers are added to each model.

Each model is augmented with Bayesian parameter uncertainty by adding dropout with a probability of 0.1 after each layer (0.5 for layers before the output layer), and setting weight decay penalties to be inversely proportional to the number of examples in the training dataset. At test time, uncertainty estimates are calculated over 100 MC samples.

For the Bayesian T-Learner we use two BNNs, each having 5 dense, 200 neuron, layers. Dropout is added after each dense layers, followed by ELU activation functions. A linear output layer is added to each network, with a sigmoid activation function if the target is binary. For image data, we add a 2-layer convolutional neural network module, with 32 and 64 filters per layer. Spatial dropout [Tom+15], and ELU activations follow each convolutional layer, and the output is flattened before being passed to the rest of the network. For image data, the Bayesian CEVAE decoder is modified by using a transposed convolution block for the part of the decoder that models $p(\mathbf{x}|\mathrm{z})$. For the propensity policies, we use a propensity model that has the

same form as a single branch of the Bayesian T-learner. The propensity model's L2 regularization is tuned for calibration as this is important for propensity models. We also experimented with a logistic regression model which performed worse.

Adam optimization [KB14] is used with a learning rate of 0.001 (On CEMNIST the learning rate for the BCEVAE is reduced to 0.0002), and we train each model for a maximum of 2000 epochs, using early stopping with a patience of 50.

Aside from these changes, model architectures, optimization strategies and loss weighting follow what is reported in their respective papers. More details can be seen in the github repository `https://github.com/OATML/ucate`.

Appendix D

Theory

D.1 Mathematical Background

Lemma D.1. *Law of total variance [WHH06].*

$$\mathrm{Var}(\mathrm{Y}) = \mathbb{E}\left[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F})\right] + \mathrm{Var}[\mathbb{E}[\mathrm{Y} \mid \mathrm{F}]]$$

Lemma D.2. *Entropy of a random variable* Y *[Sha48].*

$$\mathrm{H}(\mathrm{Y}) = -\int_{\mathcal{Y}} \log(\mathrm{y}) dP(\mathrm{y})$$

When Y *is a continuous random variable, this is known as the differential entropy.*

Lemma D.3. *Mutual information between random variables* Y *and* F *[Sha48].*

$$\begin{aligned} \mathrm{I}(\mathrm{Y};\mathrm{F}) &= \mathrm{H}(\mathrm{Y}) - \mathbb{E}\big[\mathrm{H}[\mathrm{Y} \mid \mathrm{F}]\big] \\ &= \mathrm{H}(\mathrm{F}) - \mathbb{E}[\mathrm{H}[\mathrm{F} \mid \mathrm{Y}] \end{aligned} \tag{D.1}$$

Lemma D.4. *Jensen's inequality [Jen06] states that for a random variable,* $\mathbf{X}$*, and convex function,* ψ*,*

$$\psi(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[\psi(\mathbf{X})] \tag{D.2}$$

Lemma D.5. *Let* $\mathrm{Y} \sim \mathcal{N}(f^*, \sigma_{\mathrm{y}}^2)$*, let* $\int p(\mathrm{y} \mid \mathrm{f}) dP(f) \sim \mathcal{N}(\int f dP(f), \sigma_{\mathrm{y}}^2)$*, and* σ_{y}^2 *is known. Then,*

$$\mathrm{I}(\mathrm{Y};\mathrm{F}) = \frac{1}{2}\log\Big(1 + \sigma_{\mathrm{y}}^{-2}\,\mathrm{Var}(\mathrm{F})\Big).$$

[Sri+09]

Proof.

$$
\begin{aligned}
\mathrm{I}(\mathrm{Y};\mathrm{F}) &= \mathrm{H}(\mathrm{Y}) - \mathbb{E}\big[\mathrm{H}[\mathrm{Y} \mid \mathrm{F}]\big] \\
&= \frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y})) - \mathbb{E}\left[\frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y} \mid \mathrm{F}))\right] \\
&= \frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y})) - \mathbb{E}\left[\frac{1}{2}\log(2\pi\sigma_{\mathrm{y}}^2)\right] \\
&= \frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y})) - \frac{1}{2}\log(2\pi\sigma_{\mathrm{y}}^2) \\
&= \frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y})) - \frac{1}{2}\log(2\pi\sigma_{\mathrm{y}}^2) \\
&= \frac{1}{2}\log\left(\frac{\mathrm{Var}(\mathrm{Y})}{\sigma_{\mathrm{y}}^2}\right) \\
&= \frac{1}{2}\log\left(\frac{\sigma_{\mathrm{y}}^2 + \mathrm{Var}[\mathbb{E}[\mathrm{Y} \mid \mathrm{F}]]}{\sigma_{\mathrm{y}}^2}\right) && \text{Lemma D.1} \\
&= \frac{1}{2}\log\left(1 + \sigma_{\mathrm{y}}^{-2}\,\mathrm{Var}\left(\mathbb{E}[\mathrm{Y} \mid \mathrm{F}]\right)\right) \\
&= \frac{1}{2}\log\left(1 + \sigma_{\mathrm{y}}^{-2}\,\mathrm{Var}(\mathrm{F})\right)
\end{aligned}
$$

□

Lemma D.6. *Let* $\mathrm{Y} \sim \mathcal{N}(f^*, \sigma_{\mathrm{y}}^2)$*, let* $\int p(\mathrm{y} \mid \mathrm{f})dP(f) \sim \mathcal{N}(\int f dP(f), \sigma_{\mathrm{y}}^2)$*. Then,*

$$\mathrm{I}(\mathrm{Y};\mathrm{F}) \geq \frac{1}{2}\log\left(1 + \overline{\sigma}_{\mathrm{y}}^{-2}\,\mathrm{Var}(\mathrm{F})\right),$$

where,

$$\overline{\sigma}_{\mathrm{y}}^{-2} = \iint (\mathrm{y} - \mathrm{f})^2 dP(\mathrm{y} \mid \mathrm{f}) dP(\mathrm{f})$$

[Sri+09]

Proof.

$$
\begin{aligned}
\mathrm{I}(\mathrm{Y};\mathrm{F}) &= \mathrm{H}(\mathrm{Y}) - \mathbb{E}\big[\mathrm{H}[\mathrm{Y} \mid \mathrm{F}]\big] \\
&= \frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y})) - \mathbb{E}\left[\frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y} \mid \mathrm{F}))\right] \\
&\geq \frac{1}{2}\log(2\pi\,\mathrm{Var}(\mathrm{Y})) - \frac{1}{2}\log\left(2\pi\mathbb{E}[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F})]\right) \\
&= \frac{1}{2}\log\left(\frac{\mathrm{Var}(\mathrm{Y})}{\mathbb{E}[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F})]}\right) && \text{Lemma D.4} \\
&= \frac{1}{2}\log\left(\frac{\mathbb{E}[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F}) + \mathrm{Var}[\mathbb{E}[\mathrm{Y} \mid \mathrm{F}]]}{\mathbb{E}[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F})}\right) && \text{Lemma D.1} \\
&= \frac{1}{2}\log\left(1 + \frac{\mathrm{Var}[\mathbb{E}[\mathrm{Y} \mid \mathrm{F}]]}{\mathbb{E}[\mathrm{Var}(\mathrm{Y} \mid \mathrm{F})}\right) \\
&= \frac{1}{2}\log\left(1 + \overline{\sigma}_{\mathrm{y}}^{-2}\,\mathrm{Var}\left(\mathbb{E}[\mathrm{Y} \mid \mathrm{F}]\right)\right) \\
&= \frac{1}{2}\log\left(1 + \overline{\sigma}_{\mathrm{y}}^{-2}\,\mathrm{Var}(\mathrm{F})\right)
\end{aligned}
$$

□

Lemma D.7. *For sets A and B*

$$\sup(A + B) = \sup(A) + \sup(B)$$

Zakon [Zak04].

Lemma D.8. *If A and B are non-empty sets of positive real numbers then*

$$\sup(AB) = \sup(A)\sup(B)$$

Zakon [Zak04].

Lemma D.9. *For the Heaviside step function, $H(\cdot)$,*

$$\int_{-\infty}^{\mathrm{y}^*} f(\mathrm{y})d\mathrm{y} = \int H(\mathrm{y}^* - \mathrm{y})f(\mathrm{y})d\mathrm{y}$$

$$\int_{\mathrm{y}^*}^{-\infty} f(\mathrm{y})d\mathrm{y} = \int H((\mathrm{y} - \mathrm{y}^*)f(\mathrm{y})d\mathrm{y}$$

D.2 Marginal Sensitivity Models

Lemma D.10.

$$\mathrm{f}_{\mathrm{t}}(\mathbf{x}) = \mathbb{E}[\mathrm{Y}_{\mathrm{t}} \mid \mathbf{X} = \mathbf{x}] = \frac{\int_{\mathcal{Y}} \mathrm{y}_{\mathrm{t}} w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) dP(\mathrm{y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y}_{\mathrm{t}}, \mathbf{x}) dP(\mathrm{y}_{\mathrm{t}} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})} \tag{D.5}$$

$$w(\mathrm{y_t}, \mathbf{x}) = \frac{1}{p(\mathrm{t} \mid \mathrm{Y_t} = \mathrm{y_t}, \mathbf{X} = \mathbf{x})} \tag{D.6}$$

Proof.

$$\begin{aligned}
\mathrm{f_t}(\mathbf{x}) &= \int_{\mathcal{Y}} \mathrm{y_t} dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x}), \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}, && \int_{\mathcal{Y}} dP(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x}) = 1 \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{dP(\mathrm{t}, \mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})}}{\int_{\mathcal{Y}} \frac{dP(\mathrm{t}, \mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})}}, && P(B) = \frac{P(A, B)}{P(A \mid B)} \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})} dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} \frac{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})} dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, && P(A, B) = P(B) P(A \mid B) \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{1}{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})} dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{\int_{\mathcal{Y}} \frac{1}{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})} dP(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}, && \text{cancelling terms}
\end{aligned}$$

□

Lemma D.11. *For a marginal sensitivity model, $\mathcal{P}_{(\cdot)}(\Lambda)$, given in Definitions 3.2, 5.1 and 5.2, let,*

$$w(\mathrm{y_t}, \mathbf{x}) = \alpha_{(\cdot)}(\mathbf{x}, \mathrm{t}, \Lambda) + w(\mathrm{y}, \mathbf{x})(\beta_{(\cdot)}(\mathbf{x}, \mathrm{t}, \Lambda) - \alpha_{(\cdot)})$$

, for a function $w(\mathrm{y}, \mathrm{t})$ with range [0, 1]. Then,

$$f_{\mathrm{t}}(\mathbf{x}) = f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \big(\mathrm{y} - f(\mathbf{x}, \mathrm{t})\big) w(\mathrm{y}, \mathbf{x}) dP(\mathrm{y} \mid \mathrm{t}, \mathbf{x})}{\frac{\alpha_{(\cdot)}(\mathbf{x}, \mathrm{t}, \Lambda)}{\beta_{(\cdot)}(\mathbf{x}, \mathrm{t}, \Lambda) - \alpha_{(\cdot)}(\mathbf{x}, \mathrm{t}, \Lambda)} + \int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x}) dP(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} \tag{D.8}$$

Proof.

$$\begin{aligned}
\mathrm{f_t}(\mathbf{x}) &= \frac{\int_{\mathcal{Y}} \mathrm{y_t} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})} && \text{Lemma D.10} \\
&= f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \big(\mathrm{y_t} - f(\mathbf{x}, \mathrm{t})\big) w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y_t}, \mathbf{x}) dP(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})} \\
&= f + \frac{\alpha \int \big(\mathrm{y} - f\big) w(\mathrm{y}) dP(\mathrm{y}) + (\beta - \alpha) \int \big(\mathrm{y} - f)\big) w(\mathrm{y}) dP(\mathrm{y})}{\alpha \int dP(\mathrm{y}) + (\beta - \alpha) \int w(\mathrm{y}) dP(\mathrm{y})} \\
&= f + \frac{(\beta - \alpha) \int \big(\mathrm{y} - f)\big) w(\mathrm{y}) dP(\mathrm{y})}{\alpha + (\beta - \alpha) \int w(\mathrm{y}) dP(\mathrm{y})} \\
&= f + \frac{\int \big(\mathrm{y} - f)\big) w(\mathrm{y}) dP(\mathrm{y})}{\frac{\alpha}{\beta - \alpha} + \int w(\mathrm{y}) dP(\mathrm{y})}
\end{aligned}$$

□

D.2.1 CATE Interval Estimator

Theorem D.1. *For a given marginal sensitivity model,* $P_{(\cdot)} \in \{\mathcal{P}_B(\Lambda), \mathcal{P}_D(\Lambda), \mathcal{P}_C(\Lambda)\}$,

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda), \boldsymbol{\theta}\big) \xrightarrow{p} \underline{f}_{\mathrm{t}}(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)),$$

and

$$\overline{f}_{\mathrm{t}}\big(\mathbf{x}; \mathcal{P}_{(\cdot)}(\Lambda) \xrightarrow{p} \overline{f}_{\mathrm{t}}(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)),$$

under the following assumptions:

Assumption D.1. $n \to \infty$.

Assumption D.2. $m \to \infty$.

Assumption D.3. $(\mathbf{x}, \mathrm{t}) \in \mathcal{D}_n = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n$.

Assumption D.4. Y *is a bounded random variable.*

Assumption D.5. $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ *converges in measure to* $p(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$. *Specifically,* $\lim_{n\to\infty} P(\{\mathrm{y} \in \mathcal{Y} : |p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) - p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)| \geq \epsilon\}) = 0$, *for every* $\epsilon \geq 0$, *where* $\mathcal{D}_n$ *is a dataset of size* n. *Convergence in measure is a generalization of convergence in probability.*

Assumption D.6. $p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\theta}_n)$ *is a consistent estimator of* $\mathbb{E}[\mathrm{T} \mid \mathbf{X} = \mathbf{x}]$. *(*$\mathcal{P}_B(\Lambda)$ *and* $\mathcal{P}_D(\Lambda)$ *only).*

Assumption D.7. $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ *is a consistent estimator of and* $\mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$.

Assumption D.8. $p(\mathrm{t} \mid \mathbf{x}, \mathrm{y}_\mathrm{t})$ *is bounded away from 0 and 1 uniformly for all* $\mathbf{x} \in \mathcal{X}$, $\mathrm{y} \in \mathcal{Y}$, *and* $\mathrm{t} \in \mathcal{T}$.

Proof. Here we prove that

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n\big) \xrightarrow{p} \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda),$$

from which

$$\overline{f}_{\mathrm{t}}\big(\mathbf{x}; \Lambda, \boldsymbol{\theta}_n\big) \xrightarrow{p} \overline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)$$

can be proved analogously. Where,

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda, \boldsymbol{\theta}\big) \coloneqq \underline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda), \boldsymbol{\theta}\big), \quad \underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda\big) \coloneqq \underline{f}_{\mathrm{t}}\big(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)\big),$$

for compactness. Note that $\xrightarrow{p}$ indicates convergence in probability.

As a reminder,

$$\underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda) = \inf_{w(\mathrm{y}) \in \mathcal{W}_{\mathrm{ni}}^{H}} f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} (\mathrm{y} - f(\mathbf{x}, \mathrm{t})) w(\mathrm{y}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}{\alpha'_{\mathrm{t}}(\mathbf{x}, \Lambda) + \int_{\mathcal{Y}} w(\mathrm{y}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t})}. \tag{D.10}$$

and

$$\underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n, S_m) = \inf_{w(\mathrm{y}) \in \mathcal{W}_{\mathrm{ni}}^{H}} f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) + \frac{S_m\big((\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)) w(\mathrm{y})\big)}{\alpha'_{\mathrm{t}}(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n) + S_m\big(w(\mathrm{y})\big)}, \tag{D.11}$$

with

$$\mathcal{W}_{\mathrm{ni}}^{H} \coloneqq \{w : w(y) = H(\mathrm{y}_H - \mathrm{y})\}_{\mathrm{y}_H \in \mathcal{Y}}.$$

First, we look at the Monte-Carlo integral estimators, S_m,

$$S_m\big((\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)) w(\mathrm{y})\big) = \frac{1}{m} \sum_{i=1}^{m} (\mathrm{y}_i - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) w(\mathrm{y}_i),$$

$$S_m\big(w(\mathrm{y})\big) = \frac{1}{m} \sum_{i=1}^{m} w(\mathrm{y}_i) : \quad \mathrm{y}_i \sim p(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n).$$

By Assumption D.2 and the law of large numbers we have,

$$\lim_{m \to \infty} S_m\big((\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)) w(\mathrm{y})\big) = \int (\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)) w(\mathrm{y}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n),$$

and

$$\lim_{m \to \infty} S_m\big(w(\mathrm{y})\big) = \int w(\mathrm{y}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n).$$

Therefore,

$$\begin{aligned} \lim_{m \to \infty} \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n, S_m) &= \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n) \\ &= \inf_{w(\mathrm{y}) \in \mathcal{W}_{\mathrm{ni}}^{H}} f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) + \frac{\int_{\mathcal{Y}} (\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)) w(\mathrm{y}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)}{\alpha'_{\mathrm{t}}(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n) + \int_{\mathcal{Y}} w(\mathrm{y}) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)} \end{aligned}$$

We need to show that

$$\lim_{n \to \infty} P(|\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n\big) - \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)| \geq \epsilon) = 0,$$

for all $\epsilon > 0$, where the parameters $\boldsymbol{\theta}_n$ are dependent on the size of the dataset n. First, making use of Lemma D.9, we define the following quantities:

$$\begin{aligned} \kappa_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}, n) &\coloneqq \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n), \\ \kappa^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}, n) &\coloneqq \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n), \\ I_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}) &\coloneqq \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - f(\mathbf{x}, \mathrm{t})) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}), \\ I^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}) &\coloneqq \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}), \end{aligned}$$

so that,

$$\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n\big) = f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) + \inf_{\mathrm{y}^* \in \mathcal{Y}} \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}, n)}{\alpha'_{\mathrm{t}}(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n) + \kappa^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}, n)},$$
$$\underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda) = f(\mathbf{x}, \mathrm{t}) + \inf_{\mathrm{y}^* \in \mathcal{Y}} \frac{I_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t})}{\alpha'_{\mathrm{t}}(\mathbf{x}, \Lambda) + I^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t})}.$$

For compactness, we use the following shorthand notation:

$$\kappa_{\mathrm{y}}^{\mathrm{y}^*} \coloneqq \kappa_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}, n), \quad \kappa^{\mathrm{y}^*} \coloneqq \kappa^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}, n),$$
$$I_{\mathrm{y}}^{\mathrm{y}^*} \coloneqq I_{\mathrm{y}}^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}), \quad I^{\mathrm{y}^*} \coloneqq I^{\mathrm{y}^*}(\mathbf{x}, \mathrm{t}),$$
$$\alpha'_n \coloneqq \alpha'_{\mathrm{t}}(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda), \boldsymbol{\theta}_n), \quad \alpha' \coloneqq \alpha'_{\mathrm{t}}(\mathbf{x}, \mathcal{P}_{(\cdot)}(\Lambda)),$$
$$f_n = f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n), \quad f = f(\mathbf{x}, \mathrm{t}).$$

Then, we need to express $|\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n\big) - \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)|$ as a sum of products of the following 4 terms: $\Delta^1(n) = |f_n - f|$, $\Delta^2(n) = |\alpha' - \alpha'_n|$, $\Delta^3(n) = \sup_{\mathrm{y}^* \in \mathcal{Y}} \left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right|$, and $\Delta^4(n) = \sup_{\mathrm{y}^* \in \mathcal{Y}} \left|I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*}\right|$:

$$|\underline{f}_{\mathrm{t}}\big(\mathbf{x}, \Lambda, \boldsymbol{\theta}_n\big) - \underline{f}_{\mathrm{t}}(\mathbf{x}, \Lambda)| \leq \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| f_n - f + \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha'_n + \kappa^{\mathrm{y}^*}} - \frac{I_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}} \right|, \tag{D.15a}$$
$$= |f_n - f| + \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha'_n + \kappa^{\mathrm{y}^*}} - \frac{I_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}} \right|, \tag{D.15b}$$
$$\leq |f_n - f| \tag{D.15c}$$
$$+ \sup_{\mathrm{y}^* \in \mathcal{Y}} \left\{ \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\alpha' - \alpha'_n|}{|\alpha'_n + \kappa^{\mathrm{y}^*}||\alpha' + I^{\mathrm{y}^*}|} + \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*}|}{|\alpha'_n + \kappa^{\mathrm{y}^*}||\alpha' + I^{\mathrm{y}^*}|} + \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha' + I^{\mathrm{y}^*}|} \right\}, \tag{D.15d}$$
$$= |f_n - f| \tag{D.15e}$$
$$+ \sup_{\mathrm{y}^* \in \mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||\alpha' - \alpha'_n|}{|\alpha'_n + \kappa^{\mathrm{y}^*}||\alpha' + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^* \in \mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}||I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*}|}{|\alpha'_n + \kappa^{\mathrm{y}^*}||\alpha' + I^{\mathrm{y}^*}|} + \sup_{\mathrm{y}^* \in \mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha' + I^{\mathrm{y}^*}|}, \tag{D.15f}$$
$$= \Delta^1(n) + \Delta^2(n) \sup_{\mathrm{y}^* \in \mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'_n + \kappa^{\mathrm{y}^*}||\alpha' + I^{\mathrm{y}^*}|} \tag{D.15g}$$
$$+ \Delta^4(n) \sup_{\mathrm{y}^* \in \mathcal{Y}} \frac{|\kappa_{\mathrm{y}}^{\mathrm{y}^*}|}{|\alpha'_n + \kappa^{\mathrm{y}^*}||\alpha' + I^{\mathrm{y}^*}|} + \Delta^3(n) \sup_{\mathrm{y}^* \in \mathcal{Y}} \frac{1}{|\alpha' + I^{\mathrm{y}^*}|}. \tag{D.15h}$$

Line (D.15a) by Lemma 3 in Kallus, Mao, and Zhou [KMZ19]. Lines (D.15a) - (D.15b) by Lemma D.7. Lines (D.15b) - (D.15d) by Lemma D.12 below. Lines (D.15d) - (D.15f) by Lemma D.7. Lines (D.15f) - (D.15h) by Lemma D.8.

So, we now need only prove that $\Delta^1(n) \xrightarrow{p} 0$, $\Delta^2(n) \xrightarrow{p} 0$, $\Delta^3(n) \xrightarrow{p} 0$, and $\Delta^4(n) \xrightarrow{p} 0$, when $n \to \infty$. Note that both $\Delta^1(n) \xrightarrow{p} 0$ and $\Delta^2(n) \xrightarrow{p} 0$ are covered by Assumptions D.6 and D.7; namely, $p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\theta}_n)$ *and* $f(\mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n)$ *are consistent estimators of* $\mathbb{E}[\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}]$ *and* $\mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$.

First, we prove that $\Delta^4(n) \xrightarrow{p} 0$.

Prove that $\sup_{\mathrm{y}^* \in \mathcal{Y}} \left| I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*} \right| \xrightarrow{p} 0$

$$\begin{aligned}
\sup_{\mathrm{y}^* \in \mathcal{Y}} \left| I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*} \right| &= \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) - \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) \right| \\
&= \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| P(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) - P(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) \right|
\end{aligned}$$

Convergence in probability implies convergence in distribution ($\lim_{n \to \infty} P_n(\mathbf{X} \leq \mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$), so by Assumption D.5,

$$\begin{aligned}
\lim_{n \to \infty} P\left(\left| \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*} \right| \right| \geq \epsilon \right) &= \lim_{n \to \infty} P\left(\sup_{\mathrm{y}^* \in \mathcal{Y}} \left| P_{\boldsymbol{\theta}}(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) - P(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) \right| \geq \epsilon \right) \\
&= P\left(\sup_{\mathrm{y}^* \in \mathcal{Y}} \left| P(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) - P(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) \right| \geq \epsilon \right) \\
&= P\left(\sup_{\mathrm{y}^* \in \mathcal{Y}} |0| \geq \epsilon \right) \\
&= P\left(0 \geq \epsilon \right) \\
&= 0.
\end{aligned}$$

Finally, we prove $\Delta^3(n) \xrightarrow{p} 0$.

Prove that $\sup_{\mathrm{y}^* \in \mathcal{Y}} \left| \kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*} \right| \xrightarrow{p} 0$

$$
\begin{aligned}
\sup_{\mathrm{y}^* \in \mathcal{Y}} |\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}| &= \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - f_n) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) - \int_{-\infty}^{\mathrm{y}^*} (\mathrm{y} - f) dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) \right|, \\
&= \sup_{\mathrm{y}^* \in \mathcal{Y}} \Bigg| \\
&\qquad \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) \\
&\qquad + f \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) - f_n \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y} \mid \mathbf{x}, \mathrm{t}, \boldsymbol{\theta}_n) \\
&\qquad \Bigg|, \\
&= \sup_{\mathrm{y}^* \in \mathcal{Y}} \left| \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP_{\boldsymbol{\theta}}(\mathrm{y}) - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP(\mathrm{y}) + f \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y}) - f_n \int_{-\infty}^{\mathrm{y}^*} dP_{\boldsymbol{\theta}}(\mathrm{y}) \right|, \\
&= \sup_{\mathrm{y}^* \in \mathcal{Y}} \Bigg| \\
&\qquad \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP_{\boldsymbol{\theta}}(\mathrm{y}) - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP(\mathrm{y}) + (f - f_n + f_n) \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y}) \\
&\qquad - f_n \int_{-\infty}^{\mathrm{y}^*} (p_{\boldsymbol{\theta}}(\mathrm{y}) - p(\mathrm{y}) + p(\mathrm{y})) dy \\
&\qquad \Bigg|, \\
&= \sup_{\mathrm{y}^* \in \mathcal{Y}} \Bigg| \\
&\qquad \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP_{\boldsymbol{\theta}}(\mathrm{y}) - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP(\mathrm{y}) + (f - f_n) \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y}) \\
&\qquad - f_n \int_{-\infty}^{\mathrm{y}^*} (p_{\boldsymbol{\theta}}(\mathrm{y}) - p(\mathrm{y})) \, dy \\
&\qquad \Bigg|, \\
&= \sup_{\mathrm{y}^* \in \mathcal{Y}} \Bigg| \\
&\qquad \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP_{\boldsymbol{\theta}}(\mathrm{y}) - \int_{-\infty}^{\mathrm{y}^*} \mathrm{y} dP(\mathrm{y}) + (f - f_n) \int_{-\infty}^{\mathrm{y}^*} dP(\mathrm{y}) \\
&\qquad - f_n \left(P_{\boldsymbol{\theta}}(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) - P(\mathrm{y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t}) \right) \\
&\qquad \Bigg|.
\end{aligned}
$$

As a first step, we can use the result for $\Delta^4(n)$ to remove the green term from the

supremum and now we need to show that

$$\lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP_{\boldsymbol{\theta}}(\mathrm{y})-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})+(f-f_n)\int_{-\infty}^{\mathrm{y}^*}dP(\mathrm{y})\right|\geq\epsilon.\right)=0$$

Next, under Assumption D.7 we have $f-f_n\xrightarrow{p}0$, and we are left finally to show that

$$\lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP_{\boldsymbol{\theta}}(\mathrm{y})-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})\right|\geq\epsilon.\right)=0$$

Assumption D.4 states that Y is a bounded random variable. As such, there exists a $g(\mathrm{y})$ such that $|\mathrm{y}p_{\boldsymbol{\theta}}(\mathrm{y})|\leq g(\mathrm{y})$ for all n and $y\in\mathcal{Y}$. Therefore, in conjunction with Assumption D.5, by Lebesgue's dominated convergence theorem we have $\lim_{n\to\infty}\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP_{\boldsymbol{\theta}}(\mathrm{y})=\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})$,

$$\begin{aligned}
&\lim_{n\to\infty} P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP_{\boldsymbol{\theta}}(\mathrm{y})-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})\right|\geq\epsilon.\right)\\
&\quad= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\lim_{n\to\infty}\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP_{\boldsymbol{\theta}}(\mathrm{y})-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})\right|\geq\epsilon.\right)\\
&\quad= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}\left|\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})-\int_{-\infty}^{\mathrm{y}^*}\mathrm{y}dP(\mathrm{y})\right|\geq\epsilon.\right)\\
&\quad= P\left(\sup_{\mathrm{y}^*\in\mathcal{Y}}|0|\geq\epsilon.\right)\\
&\quad= 0.
\end{aligned}$$

Therefore, $\underline{f}_{\mathrm{t}}\big(\mathbf{x},\mathcal{P}_{(\cdot)}(\Lambda),\boldsymbol{\theta}_n\big)\xrightarrow{p}\underline{f}_{\mathrm{t}}(\mathbf{x},\mathcal{P}_{(\cdot)}(\Lambda))$, and $\overline{f}_{\mathrm{t}}\big(\mathbf{x};\mathcal{P}_{(\cdot)}(\Lambda),\boldsymbol{\theta}_n\big)\xrightarrow{p}\overline{f}_{\mathrm{t}}(\mathbf{x},\mathcal{P}_{(\cdot)}(\Lambda))$ can be proved analogously. Finally, under the binary MSM, $\mathcal{P}_B(\Lambda)$, because the CATE bounds are completely determined by the CAPO bounds, we have shown: $\underline{\delta}(\mathbf{x},\mathcal{P}_B(\Lambda),\boldsymbol{\theta}_n)\xrightarrow{p}\underline{\delta}(\mathbf{x},\mathcal{P}_B(\Lambda))$, and $\overline{\delta}(\mathbf{x},\mathcal{P}_B(\Lambda),\boldsymbol{\theta}_n)\xrightarrow{p}\overline{\delta}(\mathbf{x},\mathcal{P}_B(\Lambda))$. □

Lemma D.12. *Let $\kappa_{\mathrm{y}}^{\mathrm{y}^*}$, κ^{y^*}, $I_{\mathrm{y}}^{\mathrm{y}^*}$, I^{y^*}, α_n', and α' take real values. Further, let $\alpha_n'+\kappa^{\mathrm{y}^*}>0$ and $\alpha'+I^{\mathrm{y}^*}>0$. Then,*

$$\left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha_n'+\kappa^{\mathrm{y}^*}}-\frac{I_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha'+I^{\mathrm{y}^*}}\right|\leq\frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*}\right||\alpha'-\alpha_n'|}{|\alpha_n'+\kappa^{\mathrm{y}^*}||\alpha'+I^{\mathrm{y}^*}|}+\frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*}\right|\left|I^{\mathrm{y}^*}-\kappa^{\mathrm{y}^*}\right|}{|\alpha_n'+\kappa^{\mathrm{y}^*}||\alpha'+I^{\mathrm{y}^*}|}+\frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*}-I_{\mathrm{y}}^{\mathrm{y}^*}\right|}{|\alpha'+I^{\mathrm{y}^*}|} \tag{D.18}$$

Proof.

$$\left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha_n' + \kappa^{\mathrm{y}^*}} - \frac{I_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}}\right| \leq \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha_n' + \kappa^{\mathrm{y}^*}} - \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}}\right| + \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}} - \frac{I_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}}\right| \tag{D.19a}$$

$$= \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha_n' + \kappa^{\mathrm{y}^*}} - \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}}{\alpha' + I^{\mathrm{y}^*}}\right| + \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right|}{\left|\alpha' + I^{\mathrm{y}^*}\right|} \tag{D.19b}$$

$$= \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\alpha' + I^{\mathrm{y}^*}) - \kappa_{\mathrm{y}}^{\mathrm{y}^*}(\alpha_n' + \kappa^{\mathrm{y}^*})}{(\alpha_n' + \kappa^{\mathrm{y}^*})(\alpha' + I^{\mathrm{y}^*})}\right| + \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right|}{\left|\alpha' + I^{\mathrm{y}^*}\right|} \tag{D.19c}$$

$$= \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\alpha' - \alpha_n')}{(\alpha_n' + \kappa^{\mathrm{y}^*})(\alpha' + I^{\mathrm{y}^*})} + \frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*})}{(\alpha_n' + \kappa^{\mathrm{y}^*})(\alpha' + I^{\mathrm{y}^*})}\right| + \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right|}{\left|\alpha' + I^{\mathrm{y}^*}\right|} \tag{D.19d}$$

$$\leq \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(\alpha' - \alpha_n')}{(\alpha_n' + \kappa^{\mathrm{y}^*})(\alpha' + I^{\mathrm{y}^*})}\right| + \left|\frac{\kappa_{\mathrm{y}}^{\mathrm{y}^*}(I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*})}{(\alpha_n' + \kappa^{\mathrm{y}^*})(\alpha' + I^{\mathrm{y}^*})}\right| + \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right|}{\left|\alpha' + I^{\mathrm{y}^*}\right|} \tag{D.19e}$$

$$= \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*}\right|\left|\alpha' - \alpha_n'\right|}{\left|\alpha_n' + \kappa^{\mathrm{y}^*}\right|\left|\alpha' + I^{\mathrm{y}^*}\right|} + \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*}\right|\left|I^{\mathrm{y}^*} - \kappa^{\mathrm{y}^*}\right|}{\left|\alpha_n' + \kappa^{\mathrm{y}^*}\right|\left|\alpha' + I^{\mathrm{y}^*}\right|} + \frac{\left|\kappa_{\mathrm{y}}^{\mathrm{y}^*} - I_{\mathrm{y}}^{\mathrm{y}^*}\right|}{\left|\alpha' + I^{\mathrm{y}^*}\right|} \tag{D.19f}$$

Line (D.19a) by the triangle inequality for absolute values: $|a - b| \leq |a - c| + |c - b|$. Lines (D.19a) - (D.19b) by the right-distributive property for division and preservation of division property for absolute values: $|\frac{a}{b}| = \frac{|a|}{|b|}$. Lines (D.19b) - (D.19c) by cross multiplication. Lines (D.19c) - (D.19d) by successive application of the distributive property for multiplication and the right-distributive property for division. Lines (D.19d) - (D.19e) by the subadditivity property of absolute values. Lines (D.19e) - (D.19f) by successive applications of the multiplicativity ($|ab| = |a||b|$) and preservation of division properties for absolute values. □

D.2.2 CMSM is OK

Lemma D.13. *The sensitivity bounds given in Equations (5.8a) and (5.8b) have the following equivalent expressions:*

$$\overline{f}(\mathbf{x}, \mathrm{t}; \Lambda) = \sup_{w(\mathrm{y}) \in \mathcal{W}_{nd}^H} f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - f(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}},$$

$$\underline{f}(\mathbf{x}, \mathrm{t}; \Lambda) = \inf_{w(\mathrm{y}) \in \mathcal{W}_{ni}^H} f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - f(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}},$$

where $f(\mathbf{x}, \mathrm{t}) = \mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$, $\mathcal{W}_{nd}^H = \{w : H(\mathrm{y} - \mathrm{y}_H)\}_{\mathrm{y}_H \in \mathcal{Y}}$, $\mathcal{W}_{ni}^H = \{w : H(\mathrm{y}_H - \mathrm{y})\}_{\mathrm{y}_H \in \mathcal{Y}}$, *and*

$$H(\mathrm{z}) \coloneqq \begin{cases} 1, \mathrm{z} \geq 0 \\ 0, \mathrm{z} < 0 \end{cases},$$

Proof. We follow the argument of [KMZ19] and show that our alternative formulations of $\alpha(\cdot, \Lambda)$ and $\beta(\cdot, \Lambda)$ do not change the conclusions of their linear program solution. Starting from $f(\mathbf{x}, \mathrm{t}) = \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}$, and applying a one-to-one change of variables, $\frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} = \alpha(\mathbf{x}; \mathrm{t}, \Lambda) + w(\mathrm{y})(\beta(\mathbf{x}; \mathrm{t}, \Lambda) - \alpha(\mathbf{x}; \mathrm{t}, \Lambda))$ with $w : \mathcal{Y} \to [0, 1]$, $\alpha(\mathbf{x}; \mathrm{t}, \Lambda) = 1/\Lambda p(\mathrm{t} \mid \mathbf{x})$, $\beta(\mathbf{x}; \mathrm{t}, \Lambda) = \Lambda / p(\mathrm{t} \mid \mathbf{x})$, we arrive at:

$$\overline{f}(\mathbf{x}, \mathrm{t}; \Lambda) = \sup_{w: \mathcal{Y} \to [0,1]} \frac{\int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, \tag{D.20}$$

and

$$\underline{f}(\mathbf{x}, \mathrm{t}; \Lambda) = \inf_{w: \mathcal{Y} \to [0,1]} \frac{\int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, \tag{D.21}$$

after some cancellations. Duality can be used to prove that the $w^*(\mathrm{y})$ which achieves the supremum in Equation (D.20) belongs to the set of step functions $\mathcal{W}^H_{\mathrm{nd}}$. An analogous proof for Equation (D.21) would show that the $w^*(\mathrm{y})$ which achieves the infimum in Equation (D.21) belongs to the set of step functions $\mathcal{W}^H_{\mathrm{ni}}$.

The optimization problem in Equation (D.20) can be rewritten as a linear-fractional program:

$$\begin{aligned} \text{maximize} \quad & \frac{a\langle \mathrm{y}, w(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + c}{b\langle 1, w(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + d} && \text{(D.22a)} \\ \text{subject to} \quad & 0 \leq w(\mathrm{y}) \leq 1 : \forall \mathrm{y} \in \mathcal{Y}, && \text{(D.22b)} \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})}$ is the inner product with respect to $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})$, $a = b = \lambda^2 - 1$, $c = \int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}$, and $d = \int_{\mathcal{Y}} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}$.

The linear-fractional program of Equation (D.22) is equivalent to the following linear program:

$$\begin{aligned} \text{maximize} \quad & a\langle \mathrm{y}, \widetilde{w}(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + c\widetilde{v}(\mathbf{x}) && \text{(D.23a)} \\ \text{subject to} \quad & \widetilde{w}(\mathrm{y}) \leq \widetilde{v}(\mathbf{x}) : \forall \mathrm{y} \in \mathcal{Y} && \text{(D.23b)} \\ & -\widetilde{w}(\mathrm{y}) \leq 0 : \forall \mathrm{y} \in \mathcal{Y} && \text{(D.23c)} \\ & b\langle 1, \widetilde{w}(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + d\widetilde{v}(\mathbf{x}) = 1 && \text{(D.23d)} \\ & \widetilde{v}(\mathbf{x}) \geq 0, && \text{(D.23e)} \end{aligned}$$

where

$$\widetilde{w}(\mathrm{y}) = \frac{w(\mathrm{y})}{b\langle 1, w(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + d} \quad \text{and} \quad \widetilde{v}(\mathbf{x}) = \frac{1}{b\langle 1, w(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + d}$$

by the Charnes-Cooper transformation.

Let the dual function $\rho(\mathrm{y})$ be associated with the primal constraint Eq. (D.23b), the dual function $\eta(\mathrm{y})$ be associated with the primal constraint Eq. (D.23c), and γ be the dual variable associated with the primal constraint Eq. (D.23d). The dual program is then:

$$\begin{aligned}
\text{minimize} \quad & \gamma && \text{(D.24a)}\\
\text{subject to} \quad & \rho(\mathrm{y}) - \eta(\mathrm{y}) + \gamma b p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) = a\mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) : \forall \mathrm{y} \in \mathcal{Y} && \text{(D.24b)}\\
& -\langle 1, \rho(\mathrm{y})\rangle + \gamma d \geq c && \text{(D.24c)}\\
& \rho(\mathrm{y}) \in \mathbb{R}_+, \eta(\mathrm{y}) \in \mathbb{R}_+, \gamma \in \mathbb{R} && \text{(D.24d)}
\end{aligned}$$

At most one of $\rho(\mathrm{y})$ or $\eta(\mathrm{y})$ is non-zero by complementary slackness; therefore, condition Eq. (D.24b) implies that

$$\begin{aligned}
\rho(\mathrm{y}) &= (\lambda^2 - 1) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) \max\{\mathrm{y} - \gamma, 0\} \text{ when } \eta = 0,\\
\eta(\mathrm{y}) &= (\lambda^2 - 1) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) \max\{\gamma - \mathrm{y}, 0\} \text{ when } \rho = 0.
\end{aligned}$$

[KMZ19] argue that constraint Eq. (D.24c) ought to be tight (an equivalence) at optimality, otherwise there would exist a smaller, feasible γ that satisfies the linear program. Therefore,

$$\begin{aligned}
-\langle 1, \rho(\mathrm{y})\rangle + \gamma d &= c,\\
-\int_{\mathcal{Y}} (\lambda^2 - 1) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) \max\{\mathrm{y} - \gamma, 0\} d\mathrm{y} + \gamma \int_{\mathcal{Y}} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} &= \int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y},\\
(\lambda^2 - 1) \int_{\mathcal{Y}} \max\{\mathrm{y} - \gamma, 0\} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} &= \int_{\mathcal{Y}} (\gamma - \mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}.
\end{aligned} \tag{D.25}$$

Letting $C_{\mathrm{Y}} > 0$ such that $|\mathcal{Y}| \leq C_{\mathrm{Y}}$, it is impossible that either $\gamma > C_{\mathrm{Y}}$ (the r.h.s. would be 0 and the l.h.s. would be > 0) or $\gamma < -C_{\mathrm{Y}}$ (the r.h.s. would be > 0 and the l.h.s. would be < 0). Thus, $\exists \mathrm{y}^* \in [-C_{\mathrm{Y}}, C_{\mathrm{Y}}]$ such that when $\mathrm{y} < \mathrm{y}^*$, $\eta > 0$ so $w = 0$ and when $\mathrm{y} \geq \mathrm{y}^*$, $\rho > 0$ so $w = 1$. Therefore, the optimal $w^*(\mathrm{y})$ that achieves the supremum in Equation (D.20) is in $\mathcal{W}^H_{\mathrm{nd}}$.

This result holds under

$$\begin{aligned}
f(\mathbf{x}, \mathrm{t}) &= \frac{\int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, && \text{(D.26a)}\\
&= \frac{\int_{\mathcal{Y}} \mathrm{y}_{\mathrm{t}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}, && \text{(D.26b)}\\
&= f(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - f(\mathbf{x}, \mathrm{t})) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, && \text{(D.26c)}
\end{aligned}$$

thus concluding the proof. □

D.2.3 Alternative Estimators and Optimizers

D.2.3.1 Approximating integrals using Gauss-Hermite quadrature

Gauss-Hermite quadrature is a numerical method to approximate indefinite integrals of the following form: $\int_{-\infty}^{\infty} \exp(-\mathrm{y}^2) p(\mathrm{y}) d\mathrm{y}$. In this case,

$$\int_{-\infty}^{\infty} \exp(-\mathrm{y}^2) p(\mathrm{y}) d\mathrm{y} \approx \sum_{i=1}^{m} g_i p(\mathrm{y}),$$

where m is the number of samples drawn. The y_i are the roots of the physicists Hermite polynomial $H_m^*(\mathrm{y})(i = 1, 2, \ldots, m)$ and the weights are given by

$$g_i = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 [H_{m-1}^*(\mathrm{y}_k)]^2}$$

This method can be used to calculate the expectation of a function, $h(y)$, with respect to a Gaussian distributed outcome $p(\mathrm{y}) = \mathcal{N}(\mathrm{y} \mid \mu, \sigma^2)$ through a change of variables, such that,

$$\begin{aligned} \mathbb{E}_{p(\mathrm{y})}[h(\mathrm{y})] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\mathrm{y}^2\right) h\left(\sqrt{2}\sigma \mathrm{y} + \mu\right) d\mathrm{y} \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{m} g_i h\left(\sqrt{2}\sigma \mathrm{y} + \mu\right). \end{aligned} \tag{D.27}$$

Definition D.1. Gauss-Hermite quadrature integral estimator when $p(\mathrm{y}|\mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$ is a parametric Gaussian density estimator, $\mathcal{N}(\mathrm{y} \mid f(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}), \widetilde{\sigma}^2(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}))$*:*

$$S_G(h(\mathrm{y})) \coloneqq \frac{1}{\sqrt{\pi}} \sum_{i=1}^{m} g_i h\left(\sqrt{2}\widetilde{\sigma}^2(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\mathrm{y} + f(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\right)$$

Alternatively, when the density of the outcome is modelled using a n_{y} component Gaussian mixture, $p(\mathrm{y}) = \sum_{j=1}^{n_{\mathrm{y}}} \pi_j \mathcal{N}(\mathrm{y} \mid \mu_j, \sigma_j^2)$

$$\begin{aligned} \mathbb{E}_{p(\mathrm{y})}[h(\mathrm{y})] &= \frac{1}{\sqrt{\pi}} \sum_{j=1}^{n_{\mathrm{y}}} \pi_j \int_{-\infty}^{\infty} \exp\left(-\mathrm{y}^2\right) h\left(\sqrt{2}\sigma_j \mathrm{y} + \mu_j\right) d\mathrm{y}, \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^{n_{\mathrm{y}}} \pi_j \sum_{i=1}^{m} g_i h\left(\sqrt{2}\sigma_j \mathrm{y} + \mu_j\right). \end{aligned}$$

Definition D.2. Gauss-Hermite quadrature integral estimator for expectations when $p(\mathrm{y}|\mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$ is a parametric Gaussian Mixture Density,

$$\sum_{j=1}^{n_\mathrm{y}} \widetilde{\pi}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) \mathcal{N}\left(\mathrm{y} \mid \widetilde{\mu}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}), \widetilde{\sigma}_j^2(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\right)$$

:

$$S_{GM}(h(\mathrm{y})) \coloneqq \frac{1}{\sqrt{\pi}} \sum_{j=1}^{n_\mathrm{t}} \widetilde{\pi}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) \sum_{i=1}^{m} g_i h\left(\sqrt{2}\widetilde{\sigma}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\mathrm{y} + \widetilde{\mu}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\right)$$

D.2.3.2 Line search interval optimization

Algorithm 2 Line Search Interval Optimizer

Require: $\mathbf{x}^*$ is an instance of $\mathbf{X}$, t^* is a treatment level to evaluate, Λ is a belief in the amount of hidden confounding, $\boldsymbol{\theta}$ are optimized model parameters, $\widehat{\mathcal{Y}}$ is a set of unique values $\mathrm{y} \in \mathcal{Y}$ *sorted in ascending order*.

function LineSearch($\mathbf{x}^*$, t^*, Λ, $\boldsymbol{\theta}$, $\widehat{\mathcal{Y}}$)
  $\overline{f} \leftarrow -\infty$, $\overline{\kappa} \leftarrow \infty$
  $\underline{f} \leftarrow \infty$, $\underline{\kappa} \leftarrow -\infty$
  $\underline{\delta} \leftarrow$ True, $\overline{\delta} \leftarrow$ True
  while $\underline{\delta}$ **do**
    $\mathrm{y}^* \leftarrow \textsc{Pop}(\widehat{\mathcal{Y}}_c)$ ▷ $\widehat{\mathcal{Y}}_c$ a copy of $\widehat{\mathcal{Y}}$
    $\underline{\kappa} \leftarrow \widehat{\kappa}_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}; \Lambda, H(\mathrm{y}^* - \mathrm{y}))$
    if $\underline{\kappa} < \underline{f}$ **then**
      $\underline{f} \leftarrow \underline{\kappa}$
    else
      $\underline{\delta} \leftarrow$ False
  while $\overline{\delta}$ **do**
    $\mathrm{y}^* \leftarrow \textsc{Pop}(\widehat{\mathcal{Y}}_c)$ ▷ $\widehat{\mathcal{Y}}_c$ a copy of $\widehat{\mathcal{Y}}$
    $\overline{\kappa} \leftarrow \widehat{\kappa}_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}; \Lambda, H(\mathrm{y} - \mathrm{y}^*))$
    if $\overline{\kappa} > \overline{f}$ **then**
      $\overline{f} \leftarrow \overline{\kappa}$
    else
      $\overline{\delta} \leftarrow$ False
  return $\underline{f}, \overline{f}$

D.2.3.3 Gradient Descent Interval Optimization

Second, we need a functional estimator for $w(\mathrm{y}, \mathbf{x})$. We use a neural network, $w(\mathrm{y}, \mathbf{x}; \boldsymbol{\omega})$, parameterized by $\boldsymbol{\omega}$ with sigmoid non-linearity on the output layer to satisfy the $w : \mathcal{Y} \times \mathcal{X} \to [0, 1]$ constraint.

For each (Λ, t) pair, we then need to solve the following optimization problems:

$$\underline{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \frac{1}{n} \sum_{i=1}^{n} f(w(\mathrm{y}, \cdot; \boldsymbol{\omega}); \mathbf{x}_i, \mathrm{t}, \Lambda, \boldsymbol{\theta}), \quad \mathbf{x}_i \in \mathcal{D},$$

and

$$\overline{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \frac{1}{n} \sum_{i=1}^{n} -f(w(\mathrm{y}, \cdot; \boldsymbol{\omega}); \mathbf{x}_i, \mathrm{t}, \Lambda, \boldsymbol{\theta}), \quad \mathbf{x}_i \in \mathcal{D},$$

where

$$\begin{aligned} &\mu(w(\mathrm{y}, \cdot; \boldsymbol{\omega}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}) \\ &\quad \coloneqq f(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) + \frac{S\left(w(\mathrm{y}, \mathbf{x}; \boldsymbol{\omega})(\mathrm{y} - f(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}))\right)}{(\Lambda^2 - 1)^{-1} + S(w(\mathrm{y}, \mathbf{x}; \boldsymbol{\omega}))}. \end{aligned}$$

Each of these problems can then be optimized using stochastic gradient descent [RM51; Bot98; Rud16] and error back-propogation [RHW86]. Since the optimization over $\boldsymbol{\omega}$ is non-convex, guarantees on this strategy finding the optimal solution have yet to be established. As an alternative, the line-search algorithm presented in [Jes+21b] can also be used with small modifications. Under the assumptions of Theorem 1 in [Jes+21b], with the additional assumption that T is a bounded random variable, we inherit their guarantees on the bound of the conditional average potential outcome.

The upper and lower bounds for the CAPO function under treatment $\mathrm{T} = \mathrm{t}$ and sensitivity parameter Λ can be estimated for any observed covariate value, $\mathbf{X} = \mathbf{x}$, as

$$\underline{f}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) = f(w(\mathrm{y}, \cdot; \underline{\boldsymbol{\omega}}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}),$$

and

$$\overline{f}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) = f(w(\mathrm{y}, \cdot; \overline{\boldsymbol{\omega}}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}).$$

The upper and lower bounds for the APO (dose-response) function under treatment $\mathrm{T} = \mathrm{t}$ and sensitivity parameter Λ can be estimated over any set of observed covariates $\mathcal{D}_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^{n}$, as

$$\underline{f}(\mathrm{t}; \Lambda, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \underline{f}(\mathbf{x}_i, \mathrm{t}; \Lambda, \boldsymbol{\theta}), \quad \mathbf{x}_i \in \mathcal{D}_{\mathbf{x}},$$

$$\overline{f}(\mathrm{t}; \Lambda, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \overline{f}(\mathbf{x}_i, \mathrm{t}; \Lambda, \boldsymbol{\theta}), \quad \mathbf{x}_i \in \mathcal{D}_{\mathbf{x}}.$$

Bibliography

[Agl+20] Virginia Aglietti et al. "Causal bayesian optimization". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3155–3164.

[Agl+21] Virginia Aglietti et al. "Dynamic causal Bayesian optimization". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10549–10560.

[Agl+23] Virginia Aglietti et al. "Constrained causal bayesian optimization". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 304–321.

[Alb89] Bruce A Albrecht. "Aerosols, cloud microphysics, and fractional cloudiness". In: *Science* 245.4923 (1989), pp. 1227–1230.

[AV17] 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 (2017).

[AV18] Ahmed M Alaa and Mihaela Van Der Schaar. "Bayesian nonparametric causal inference: Information rates and learning algorithms". In: *IEEE Journal of Selected Topics in Signal Processing* 12.5 (2018), pp. 1031–1046.

[AW19] Susan Athey and Stefan Wager. "Estimating treatment effects with causal forests: An application". In: *Observational studies* 5.2 (2019), pp. 37–51.

[Bis95] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[Bos+15] Michael G Bosilovich et al. "MERRA-2: Initial evaluation of the climate". In: (2015).

[Bot98] Leon Bottou. "Online learning and stochastic approximations". In: *Online learning in neural networks* 17.9 (1998), p. 142.

[Bou+13] Olivier Boucher et al. "Clouds and aerosols". In: *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2013, pp. 571–657.

[BP06] Bryan A Baum and Steven Platnick. "Introduction to MODIS cloud products". In: *Earth science satellite remote sensing*. Springer, 2006, pp. 74–91.

[Bra+23] Nicola Branchini et al. “Causal entropy optimization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 8586–8605.

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[Bro+20] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[BTG02] Francois-Marie Bréon, Didier Tanré, and Sylvia Generoso. “Aerosol effect on cloud droplet size monitored from satellite”. In: *Science* 295.5556 (2002), pp. 834–838.

[CGJ96] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. “Active learning with statistical models”. In: *Journal of artificial intelligence research* 4 (1996), pp. 129–145.

[CH20a] Carlos Cinelli and Chad Hazlett. “An omitted variable bias framework for sensitivity analysis of instrumental variables”. In: *Work. Pap* (2020).

[CH20b] Carlos Cinelli and Chad Hazlett. “Making sense of sensitivity: Extending omitted variable bias”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1 (2020), pp. 39–67.

[Che+18] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[Che+21] Victor Chernozhukov et al. “Omitted Variable Bias in Machine Learned Causal Models”. In: *arXiv preprint arXiv:2112.13398* (2021).

[CHH16] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. “Assessing sensitivity to unmeasured confounding using a simulated potential confounder”. In: *Journal of Research on Educational Effectiveness* 9.3 (2016), pp. 395–420.

[Chr+17] Matthew W Christensen et al. “Unveiling aerosol–cloud interactions–Part 1: Cloud contamination in satellite products enhances the aerosol indirect forcing estimate”. In: *Atmospheric Chemistry and Physics* 17.21 (2017), pp. 13151–13164.

[Cin+19] Carlos Cinelli et al. “Sensitivity analysis of linear structural causal models”. In: *International conference on machine learning*. PMLR. 2019, pp. 1252–1261.

[CK08] Marco Caliendo and Sabine Kopeinig. “Some practical guidance for the implementation of propensity score matching”. In: *Journal of economic surveys* 22.1 (2008), pp. 31–72.

[CS11] Matthew W Christensen and Graeme L Stephens. “Microphysical and macrophysical responses of marine stratocumulus polluted by underlying ships: Evidence of cloud deepening”. In: *Journal of Geophysical Research: Atmospheres* 116.D3 (2011).

[CV21a] Alicia Curth and Mihaela Van Der Schaar. “Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1810–1818.

[CV21b] Alicia Curth and Mihaela Van Der Schaar. "On inductive biases for heterogeneous treatment effect estimation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15883–15894.

[DAm+21] Alexander D'Amour et al. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2 (2021), pp. 644–654.

[Daw00] Philip Dawid. "Causal inference without counterfactuals". In: *Journal of the American statistical Association* 95.450 (2000), pp. 407–424.

[Daw21] Philip Dawid. "Decision-theoretic foundations for statistical causality". In: *Journal of Causal Inference* 9.1 (2021), pp. 39–77.

[Daw79] Philip Dawid. "Conditional independence in statistical theory". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1 (1979), pp. 1–15.

[DD09] Armen Der Kiureghian and Ove Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural safety* 31.2 (2009), pp. 105–112.

[Del+24] Tristan Deleu et al. "Joint bayesian inference of graphical structure and parameters with a single generative flow network". In: *Advances in Neural Information Processing Systems* 36 (2024).

[Den12] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[DF21] Alexander D'Amour and Alexander Franks. "Deconfounding Scores: Feature Representations for Causal Effect Estimation with Weak Overlap". In: *arXiv preprint arXiv:2104.05762* (2021).

[DG23] Jacob Dorn and Kevin Guo. "Sharp sensitivity analysis for inverse propensity weighting via quantile balancing". In: *Journal of the American Statistical Association* 118.544 (2023), pp. 2645–2657.

[DGH22] Edward De Brouwer, Javier Gonzalez, and Stephanie Hyland. "Predicting the impact of treatments over time with uncertainty aware neural differential equations." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 4705–4722.

[DGK24] Jacob Dorn, Kevin Guo, and Nathan Kallus. "Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding". In: *Journal of the American Statistical Association* (2024), pp. 1–12.

[Dia+22] Michael S Diamond et al. "Opinion: To assess marine cloud brightening's technical feasibility, we need to know what to study—and when to stop". In: *Proceedings of the National Academy of Sciences* 119.4 (2022).

[DL20] Alyson Douglas and Tristan L'Ecuyer. "Quantifying cloud adjustments and the radiative forcing due to aerosol–cloud interactions in satellite observations of warm marine clouds". In: *Atmospheric Chemistry and Physics* 20.10 (2020), pp. 6225–6241.

[DL21] Alyson Douglas and Tristan L'Ecuyer. "Global evidence of aerosol-induced invigoration in marine cumulus clouds". In: *Atmospheric Chemistry and Physics* 21.19 (2021), pp. 15103–15114.

[Dor+16] Vincent Dorie et al. “A flexible, interpretable framework for assessing sensitivity to unmeasured confounding”. In: *Statistics in medicine* 35.20 (2016), pp. 3453–3470.

[Dor+19] Vincent Dorie et al. “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition”. In: *Statistical Science* 34.1 (2019), pp. 43–68.

[Dor16] Vincent Dorie. “NPCI: Non-parametrics for causal inference”. In: *URL: https://github. com/vdorie/npci* (2016).

[Dos+20] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[DPM11] Kun Deng, Joelle Pineau, and Susan Murphy. “Active learning for personalizing treatment”. In: *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE. 2011, pp. 32–39.

[Dun75] Otis Dudley Duncan. *Introduction to structural equation models*. Academic Press, 1975.

[Dur+23] Joshua Durso-Finley et al. “Improving Image-Based Precision Medicine with Uncertainty-Aware Causal Models”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 472–481.

[DY23] Jacob Dorn and Luther Yap. “Sensitivity Analysis for Linear Estimands”. In: *arXiv preprint arXiv:2309.06305* (2023).

[FDF19] AlexanderM Fraznks, Alexander D’Amour, and Avi Feller. “Flexible sensitivity analysis for observational studies without observable implications”. In: *Journal of the American Statistical Association* (2019).

[FGR21] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. “On Statistical Bias In Active Learning: How and When to Fix It”. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=JiYq3eqTKY.

[FHW23] Edwin Fong, Chris Holmes, and Stephen G Walker. “Martingale posterior distributions”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.5 (2023), pp. 1357–1391.

[Fis36] Ronald Aylmer Fisher. “Design of experiments”. In: *British Medical Journal* 1.3923 (1936), p. 554.

[FMF24] Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. “Sharp bounds for generalized causal sensitivity analysis”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[Fra+23a] Kenneth A Frank et al. “Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science”. In: *Social Science Research* 110 (2023), p. 102815.

[Fra+23b] Dennis Frauen et al. “A neural framework for generalized causal sensitivity analysis”. In: *arXiv preprint arXiv:2311.16026* (2023).

[Gar+18] Marta Garnelo et al. “Neural processes”. In: *arXiv preprint arXiv:1807.01622* (2018).

[Gel+17] Ronald Gelaro et al. “The modern-era retrospective analysis for research and applications, version 2 (MERRA-2)”. In: *Journal of climate* 30.14 (2017), pp. 5419–5454.

[GG16] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[Gry+19] Edward Gryspeerdt et al. “Constraining the aerosol influence on cloud liquid water path”. In: *Atmospheric Chemistry and Physics* 19.8 (2019), pp. 5331–5347.

[HA15] José Miguel Hernández-Lobato and Ryan Adams. “Probabilistic backpropagation for scalable learning of bayesian neural networks”. In: *International conference on machine learning*. PMLR. 2015, pp. 1861–1869.

[Haa43] Trygve Haavelmo. “The statistical implications of a system of simultaneous equations”. In: *Econometrica, Journal of the Econometric Society* (1943), pp. 1–12.

[He+16] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[Her+16] Jose Hernandez-Lobato et al. “Black-box alpha divergence minimization”. In: *International conference on machine learning*. PMLR. 2016, pp. 1511–1520.

[Hes+23] Konstantin Hess et al. “Bayesian Neural Controlled Differential Equations for Treatment Effect Estimation”. In: *The Twelfth International Conference on Learning Representations*. 2023.

[HI04] Keisuke Hirano and Guido W Imbens. “The propensity score with continuous treatments”. In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164 (2004), pp. 73–84.

[Hil11] Jennifer L Hill. “Bayesian nonparametric modeling for causal inference”. In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240.

[HMG15] James Hensman, Alexander Matthews, and Zoubin Ghahramani. “Scalable variational Gaussian process classification”. In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 351–360.

[Hof+13] Matthew D Hoffman et al. “Stochastic variational inference”. In: *Journal of Machine Learning Research* (2013).

[Hol+23] David Holzmüller et al. “A framework and benchmark for deep batch active learning for regression”. In: *Journal of Machine Learning Research* 24.164 (2023), pp. 1–81.

[Hol86] Paul W Holland. “Statistics and causal inference”. In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[Hou+11] Neil Houlsby et al. “Bayesian Active Learning for Classification and Preference Learning”. In: *stat* 1050 (2011), p. 24.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[Hu+21] Yaowei Hu et al. “A generative adversarial framework for bounding confounded causal effects”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 13. 2021, pp. 12104–12112.

[Imb03] Guido W. Imbens. “Sensitivity to Exogeneity Assumptions in Program Evaluation”. In: *American Economic Review* 93.2 (2003), pp. 126–132. DOI: 10.1257/000282803321946921. URL: https://www.aeaweb.org/articles?id=10.1257/000282803321946921.

[Izm+18] P Izmailov et al. “Averaging weights leads to wider optima and better generalization”. In: *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. 2018, pp. 876–885.

[Jae+21] Andrew Jaegle et al. “Perceiver: General perception with iterative attention”. In: *International conference on machine learning*. PMLR. 2021, pp. 4651–4664.

[Jen06] Johan Ludwig William Valdemar Jensen. “Sur les fonctions convexes et les inégalités entre les valeurs moyennes”. In: *Acta mathematica* 30.1 (1906), pp. 175–193.

[Jes+20] Andrew Jesson et al. “Identifying causal-effect inference failure with uncertainty-aware models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11637–11649.

[Jes+21a] Andrew Jesson et al. “Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 30465–30478.

[Jes+21b] Andrew Jesson et al. “Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 4829–4838. URL: https://proceedings.mlr.press/v139/jesson21a.html.

[Jes+22] Andrew Jesson et al. “Scalable Sensitivity and Uncertainty Analyses for Causal-Effect Estimates of Continuous-Valued Interventions”. In: *Advances in Neural Information Processing Systems*. 2022.

[Jes+24] Andrew Jesson et al. “ReLU to the Rescue: Improve Your On-Policy Actor-Critic with Positive Advantages”. In: *ICML* (2024).

[KB14] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[KG17] Alex Kendall and Yarin Gal. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 5574–5584.

[Kha+18] Mohammad Khan et al. “Fast and scalable bayesian deep learning by weight-perturbation in adam”. In: *International conference on machine learning*. PMLR. 2018, pp. 2611–2620.

[Kir+21] Andreas Kirsch et al. “Stochastic batch acquisition for deep active learning”. In: *arXiv preprint arXiv:2106.12059* (2021).

[Kir23] Andreas Kirsch. “Black-Box Batch Active Learning for Regression”. In: *arXiv preprint arXiv:2302.08981* (2023).

[Kiv+21] Ian Kivlichan et al. “Measuring and Improving Model-Moderator Collaboration using Uncertainty Estimation”. In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 2021, pp. 36–53.

[KKS20] Niki Kilbertus, Matt J Kusner, and Ricardo Silva. “A Class of Algorithms for General Instrumental Variable Models”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 20108–20119. URL: https://proceedings.neurips.cc/paper/2020/file/e8b1cbd05f6e6a358a81dee52493dd06-Paper.pdf.

[KMZ19] Nathan Kallus, Xiaojie Mao, and Angela Zhou. “Interval estimation of individual-level causal effects under unobserved confounding”. In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 2281–2290.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[Kün+19] Sören R Künzel et al. “Metalearners for estimating heterogeneous treatment effects using machine learning”. In: *Proceedings of the national academy of sciences* 116.10 (2019), pp. 4156–4165.

[KVG19] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. “Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning”. In: *Advances in neural information processing systems* 32 (2019).

[KW13] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[LeC+98] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[LeC98] Yann LeCun. “The MNIST database of handwritten digits”. In: *http://yann. lecun. com/exdb/mnist/* (1998).

[Lee+23] Hyungi Lee et al. “Martingale posterior neural processes”. In: *The Eleventh International Conference on Learning Representations*. International Conference on Learning Representations. 2023.

[Lei+17] Christian Leibig et al. “Leveraging uncertainty information from deep neural networks for disease detection”. In: *Scientific reports* 7.1 (2017), pp. 1–14.

[LHT15] Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. “Stochastic expectation propagation”. In: *Advances in neural information processing systems* 28 (2015).

[Liu+20] Jeremiah Liu et al. “Simple and principled uncertainty estimation with deterministic deep learning via distance awareness”. In: (2020).

[Lou+17] Christos Louizos et al. “Causal effect inference with deep latent-variable models”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 6446–6456.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *Advances in Neural Information Processing Systems*. Ed. by

I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38Paper.pdf.

[Mac03] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[Mac95a] David JC MacKay. "Bayesian neural networks and density networks". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 354.1 (1995), pp. 73–80.

[Mac95b] David JC MacKay. "Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks". In: *Network: computation in neural systems* 6.3 (1995), p. 469.

[Mar+23] Myrl G Marmarelis et al. "Partial identification of dose responses with hidden confounders". In: *Uncertainty in Artificial Intelligence*. PMLR. 2023, pp. 1368–1379.

[Mar+24] Myrl G Marmarelis et al. "Ensembled Prediction Intervals for Causal Outcomes Under Hidden Confounding". In: *Causal Learning and Reasoning*. PMLR. 2024, pp. 18–40.

[Mas+21] V. Masson-Delmotte et al. "IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change". In: (2021).

[MFF24] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. "Partial counterfactual identification of continuous outcomes with a curvature sensitivity model". In: *Advances in Neural Information Processing Systems* 36 (2024).

[MH08] Laurens Van Der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[Mid+16] Joel A Middleton et al. "Bias amplification and bias unmasking". In: *Political Analysis* 24.3 (2016), pp. 307–323.

[Mül+21] Samuel Müller et al. "Transformers can do bayesian inference". In: *arXiv preprint arXiv:2112.10510* (2021).

[Myh+07] Gunnar Myhre et al. "Aerosol-cloud interaction inferred from MODIS satellite data and global aerosol models". In: *Atmospheric Chemistry and Physics* 7.12 (2007), pp. 3081–3101.

[Nea12] Radford M Neal. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media, 2012.

[Nea95] Radford M Neal. "BAYESIAN LEARNING FOR NEURAL NETWORKS". PhD thesis. University of Toronto, 1995.

[Ney23] Jerzy Neyman. "edited and translated by dorota m. dabrowska and terence p. speed (1990). on the application of probability theory to agricultural experiments. essay on principles. section 9". In: *Statistical Science* 5.4 (1923), pp. 465–472.

[Nie+21] Lizhen Nie et al. "VCNet and Functional Targeted Regularization For Learning Causal Effects of Continuous Treatments". In: *arXiv preprint arXiv:2103.07861* (2021).

[Nis72] Kenneth R Niswander. "The collaborative perinatal study of the National Institute of Neurological Diseases and Stroke". In: *The Woman and Their Pregnancies* (1972).

[NW21] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (2021), pp. 299–319.

[Opr+23] Miruna Oprescu et al. "B-Learner: Quasi-Oracle Bounds on Heterogeneous Causal Effects Under Hidden Confounding". In: (2023).

[Osb+21] Ian Osband et al. "Epistemic neural networks". In: *arXiv preprint arXiv:2107.08924* (2021).

[Ost19] Emily Oster. "Unobservable selection and coefficient stability: Theory and evidence". In: *Journal of Business & Economic Statistics* 37.2 (2019), pp. 187–204.

[Pad+22] Kirtan Padh et al. "Stochastic Causal Programming for Bounding Treatment Effects". In: *arXiv preprint arXiv:2202.10806* (2022).

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Pet+12] Maya L Petersen et al. "Diagnosing and responding to violations in the positivity assumption". In: *Statistical methods in medical research* 21.1 (2012), pp. 31–54.

[PZ13] D Painemal and P Zuidema. "The first aerosol indirect effect quantified through airborne remote sensing during VOCALS-REx". In: *Atmospheric Chemistry and Physics* 13.2 (2013), pp. 917–931.

[QWZ21] Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. "Budgeted Heterogeneous Treatment Effect Estimation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8693–8702.

[Rag+19] Maithra Raghu et al. "The algorithmic automation problem: Prediction, triage, and human effort". In: *arXiv preprint arXiv:1903.12220* (2019).

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[RM51] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[RN10] Carl Edward Rasmussen and Hannes Nickisch. "Gaussian processes for machine learning (GPML) toolbox". In: *The Journal of Machine Learning Research* 11 (2010), pp. 3011–3015.

[Ros14] Paul R Rosenbaum. "Sensitivity analysis in observational studies". In: *Wiley StatsRef: Statistics Reference Online* (2014).

[RR13] Thomas S Richardson and James M Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30 (2013), p. 2013.

[RR83] P. R. Rosenbaum and D. B. Rubin. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 45.2 (1983), pp. 212–218. ISSN: 00359246. URL: http://www.jstor.org/stable/2345524.

[RRS00] James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models". In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 1–94.

[Rub74] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[Rub80] Donald B Rubin. "Randomization analysis of experimental data: The Fisher randomization test comment". In: *Journal of the American Statistical Association* 75.371 (1980), pp. 591–593.

[Rud16] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

[SBV19] Claudia Shi, David Blei, and Victor Veitch. "Adapting neural networks for the estimation of treatment effects". In: *Advances in Neural Information Processing Systems*. 2019, pp. 2503–2513.

[Sch+16] Nick AJ Schutgens et al. "Will a perfect model agree with perfect observations? The impact of spatial sampling". In: *Atmospheric Chemistry and Physics* 16.10 (2016), pp. 6335–6353.

[Sch+20] Patrick Schwab et al. "Learning counterfactual representations for estimating individual dose-response curves". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5612–5619.

[Sek08] Jasjeet S Sekhon. "The Neyman-Rubin model of causal inference and estimation via matching methods". In: *The Oxford handbook of political methodology* 2 (2008), pp. 1–32.

[SF09] Bjorn Stevens and Graham Feingold. "Untangling aerosol effects on clouds and precipitation in a buffered system". In: *Nature* 461.7264 (2009), pp. 607–613.

[SGC22] Arvid Sjölander, Erin E Gabriel, and Iuliana Ciocănea-Teodorescu. "Sensitivity analysis for causal effects with generalized linear models". In: *Journal of Causal Inference* 10.1 (2022), pp. 441–479.

[Sha48] Claude E Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[Shi+18] Yishai Shimoni et al. "Benchmarking framework for performance-evaluation of causal inference analysis". In: *arXiv preprint arXiv:1802.05046* (2018).

[SJS17] 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-Volume 70*. JMLR. org. 2017, pp. 3076–3085.

[Sno+15] Jasper Snoek et al. “Scalable bayesian optimization using deep neural networks”. In: *International conference on machine learning*. PMLR. 2015, pp. 2171–2180.

[Sri+09] Niranjan Srinivas et al. “Gaussian process optimization in the bandit setting: No regret and experimental design”. In: *arXiv preprint arXiv:0912.3995* (2009).

[SS77] Herbert A Simon and Herbert A Simon. “Causal ordering and identifiability”. In: *Models of Discovery: And Other Topics in the Methods of Science* (1977), pp. 53–80.

[Sun+18] Shengyang Sun et al. “FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS”. In: *International Conference on Learning Representations*. 2018.

[Sun+19] Iiris Sundin et al. “Active learning for decision-making from imbalanced observational data”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6046–6055.

[Tan06] Zhiqiang Tan. “A Distributional Approach for Causal Inference Using Propensity Scores”. In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1619–1637. DOI: 10.1198/016214506000000023. eprint: https://doi.org/10.1198/016214506000000023. URL: https://doi.org/10.1198/016214506000000023.

[Tig+22] Panagiotis Tigas et al. “Interventions, where and how? experimental design for causal models at scale”. In: *Advances in neural information processing systems* 35 (2022), pp. 24130–24143.

[Tig+23] Panagiotis Tigas et al. “Differentiable multi-target causal bayesian experimental design”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 34263–34279.

[Tit+85] D Michael Titterington et al. *Statistical analysis of finite mixture distributions*. Vol. 198. John Wiley & Sons Incorporated, 1985.

[Tol+17] Velle Toll et al. “Volcano and ship tracks indicate excessive aerosol-induced cloud water increases in a climate model”. In: *Geophysical research letters* 44.24 (2017), pp. 12–492.

[Tol+19] Velle Toll et al. “Weak average liquid-cloud-water response to anthropogenic aerosols”. In: *Nature* 572.7767 (2019), pp. 51–55.

[Tom+15] Jonathan Tompson et al. “Efficient object localization using convolutional networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 648–656.

[Tot+22] Christian Toth et al. “Active bayesian causal inference”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16261–16275.

[Two77] Sean Twomey. “The influence of pollution on the shortwave albedo of clouds”. In: *Journal of the atmospheric sciences* 34.7 (1977), pp. 1149–1152.

[Van+21] Joost Van Amersfoort et al. “Improving Deterministic Uncertainty Estimation in Deep Learning for Classification and Regression”. In: *arXiv preprint arXiv:2102.11409* (2021).

[Vas+17] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[VZ20] Victor Veitch and Anisha Zaveri. *Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding*. 2020. arXiv: 2003.01747 [stat.ME].

[Wen+23] Hechuan Wen et al. “To Predict or to Reject: Causal Effect Estimation with Uncertainty on Networked Data”. In: *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2023, pp. 1415–1420.

[Wer74] P. J. Werbos. “Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences”. PhD thesis. Harvard University, 1974.

[WHH06] Neil A Weiss, Paul T Holmes, and Michael Hardy. *A course in probability*. Pearson Addison Wesley Boston, MA, USA: 2006, pp. 385–386.

[Wil+16] Andrew Gordon Wilson et al. “Deep kernel learning”. In: *Artificial intelligence and statistics*. PMLR. 2016, pp. 370–378.

[WR06] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA, 2006.

[Wri34] Sewall Wright. “The method of path coefficients”. In: *The annals of mathematical statistics* 5.3 (1934), pp. 161–215.

[Xie+21] Sang Michael Xie et al. “An Explanation of In-context Learning as Implicit Bayesian Inference”. In: *International Conference on Learning Representations*. 2021.

[Yad+18] Steve Yadlowsky et al. “Bounds on the conditional and average treatment effect with unobserved confounding factors”. In: *arXiv preprint arXiv:1808.09521* (2018).

[Yin+24] Mingzhang Yin et al. “Conformal sensitivity analysis for individual treatment effects”. In: *Journal of the American Statistical Association* 119.545 (2024), pp. 122–135.

[YJV18] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. “GANITE: Estimation of individualized treatment effects using generative adversarial nets”. In: *International conference on learning representations*. 2018.

[Zak04] Elias Zakon. *Mathematical analysis*. The Trillia Group, 2004.

[ZB20] Junzhe Zhang and Elias Bareinboim. “Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach.” In: *JMLR workshop and conference proceedings*. Vol. 119. 2020.

[Zha+23] Jiaqi Zhang et al. “Active learning for optimal intervention design in causal models”. In: *Nature Machine Intelligence* 5.10 (2023), pp. 1066–1075.

[ZSB19] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. “Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.4 (2019), pp. 735–761.