Generalization Bounds for Causal Regression: Insights, Guarantees and Sensitivity Analysis

Daniel Csillag [1] **Claudio José Struchiner** [1] **Guilherme Tegoni Goedert** [1]

Abstract

Many algorithms have been recently proposed for causal machine learning. Yet, there is little to no theory on their quality, especially considering finite samples. In this work, we propose a theory based on generalization bounds that provides such guarantees. By introducing a novel change-of-measure inequality, we are able to tightly bound the model loss in terms of the deviation of the treatment propensities over the population, which we show can be empirically limited. Our theory is fully rigorous and holds even in the face of hidden confounding and violations of positivity. We demonstrate our bounds on semi-synthetic and real data, showcasing their remarkable tightness and practical utility.

1. Introduction

Causal machine learning plays an increasingly important role in many application domains including economics, medicine, education research, and more. At the core of causal ML is reasoning about potential outcomes. For instance, using covariates X to make predictions about the potential outcomes Y^1 and Y^0, which correspond to *what would happen* if the treatment were to be administered ($T = 1$) and if the treatment were not to be administered ($T = 0$). This reasoning differs crucially from simply predicting the actual outcome Y given those covariates, which is subject to biases in the training data.

The fundamental problem of causal ML is that the potential outcomes cannot be both observed at the same time. For instance, whenever $T = 1$ we can only observe $Y = Y^1$, and Y^0 could vary unpredictably. To amend this, some strong assumptions are introduced, with the most typical being *ignorability* (that $Y^1, Y^0 \perp\!\!\!\perp T|X$) and *positivity* (that for all X and a, $0 < \mathbb{P}[T = a|X] < 1$). When these don't hold we are in the field of *sensitivity analysis*. In this context, we may assume that there are certain unobserved covariates U which must be included to satisfy ignorability.

We consider two key tasks of causal ML: (i) outcome regression, in which we seek to predict the potential outcomes Y^a given the covariates, and (ii) individual treatment effect estimation, in which we seek to predict the treatment effects $Y^1 - Y^0$ given the covariates.

Many methods have been proposed for these tasks. There are classical approaches based on linear models (Angrist & Imbens, 1995; Hirano & Imbens, 2001) as well as more recent approaches based on decision trees (Athey & Imbens, 2015; Wager & Athey, 2015; Athey et al., 2016), neural networks (Shalit et al., 2017; Johansson et al., 2016; Yoon et al., 2018; Zhang et al., 2020) and even some model-agnostic approaches (Künzel et al., 2017; Nie & Wager, 2017; Oprescu et al., 2023).

However, all these methods still lack a rigorous supporting theory, with there being fundamental questions that have not been fully answered. For example, how well can we expect these procedures to extract causal relations from the data? How many samples are enough? How do these procedures fare when causal assumptions are violated (e.g., because there are unobserved confounders)?

In this paper, we seek to provide satisfying answers to these questions through the lens of generalization bounds. By leveraging a change-of-measure inequality based on an f-divergence (namely, the Pearson χ^2 divergence) we are able to tightly bound the (unobservable) complete causal loss in terms of the (observable) conditional loss plus some additional highly interpretable terms:

Theorem 1.1 (Informal)**.** *For any (decomposable) loss function and any $\lambda > 0$,*

$$\underbrace{\mathbb{E}[\text{Loss}]}_{\textit{unobservable...}} \leq \underbrace{\mathbb{E}[\text{Loss}|T = a]}_{\textit{observable!}} + \lambda \cdot \Delta + \sigma^2/4\lambda$$

where Δ is a term that quantifies how far we deviate from a randomized control trial.

However, the result above suffers from the fact that Δ fundamentally depends on unobservable quantities. But it can –

[1]School of Applied Mathematics, Fundação Getúlio Vargas, Rio de Janeiro, Brazil. Correspondence to: Daniel Csillag <daniel.csillag@fgv.br>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

quite surprisingly – be empirically upper bounded, leading to a new form of sensitivity analysis:

Theorem 1.2 (Informal)**.** *Let ν be an (arbitrary) propensity scoring model. Then the following holds:*

$$\Delta \leq C \cdot \mathrm{BrierScore}(\nu, T) + D$$

where D quantifies how our reweighed samples deviate from being balanced, C is an universal constant, and $\mathrm{BrierScore}(\nu, T) = \mathbb{E}[(\nu(X) - T)^2]$.

We demonstrate empirically that these bounds are tight and, besides aiding in a theoretical backbone for causal regression, are also useful in practice for model selection.

Our main contributions are:

- Novel generalization bounds applicable to many causal regression algorithms, which shed light on the applicability and efficacy of such procedures. Our bounds are general, assumption-light, framework-agnostic (i.e., can be used in conjunction with Rademacher bounds, VC bounds, PAC-Bayes, etc.) and hold even in the lack of ignorability or positivity.
- Relaxed versions of our bounds which are entirely empirically boundable with high probability, allowing for practical bounding of the unobservable counterfactual losses.
- A change-of-measure inequality based on the Pearson χ^2 f-divergence, which is remarkably tight (Figure 2) and particularly applicable to causal inference problems.
- Our bounds show that we are able to learn to estimate treatment effects while optimizing losses other than the mean squared loss. For example, we can use the mean absolute error for robust estimation, or the quantile loss[1] for quantile regression, a feat previously thought impossible in general.

Related work There has been much work on asymptotic guarantees for causal regression algorithms, e.g., (Künzel et al., 2017; Nie & Wager, 2017; Oprescu et al., 2023; Athey et al., 2019). However, besides not being applicable to finite samples, they are often restricted to specific learning classes and/or make restrictive modelling assumptions. A notable exception to this is (Johansson et al., 2022), which also establishes some generalization bounds on causal regression. However, their bounds are orders of magnitude looser than ours (see Section 3.1) and are restricted to bounds based on the pseudo-VC-dimension and mean squared error. Also closely related are generalization bounds for domain adaptation (Cortes et al., 2010; Blitzer et al., 2007; Germain et al., 2017; 2016; Mansour et al., 2009). Under ignorability, the observed and complete distributions differ by a covariate shift, and causal inference becomes a domain adaptation problem. However, these bounds generally give little intuition when applied in a causal context.

[1] also known as the "pinball" loss.

2. Novel Bounds for Causal Regression

In order to bridge the gap between the complete data distribution and the observed data distribution, we develop a novel change-of-measure inequality based on Pearson's χ^2-divergence:

Definition 2.1 (Pearson's χ^2 divergence)**.** Let $\mathcal{H}$ be any arbitrary domain, and denote by P and Q the probability measures over the Borel σ-field on $\mathcal{H}$. The χ^2 divergence between Q and P, denoted $\chi^2(Q\|P)$, is given by

$$\chi^2(Q\|P) := \mathbb{E}_P\left[\left(\frac{\mathrm{d}Q}{\mathrm{d}P} - 1\right)^2\right].$$

Lemma 2.2. *Let P and Q be as in Definition 2.1. For any $\lambda > 0$,*

$$\mathbb{E}_Q[\phi] - E \leq \mathbb{E}_P[\phi] \leq \mathbb{E}_Q[\phi] + E$$

where

$$E := \lambda \cdot \chi^2(Q\|P) + \frac{1}{4\lambda}\mathrm{Var}_P(\phi).$$

Moreover, the bound is optimized for

$$\lambda^\star = \sqrt{\mathrm{Var}_P(\phi)/4\chi^2(Q\|P)}.$$

Note how, if $P = Q$, then the bound in Lemma 2.2 becomes an equality by taking $\lambda \to \infty$, showing that it is reasonably tight. This is in contrast to, e.g., the inequalities in (Ohnishi & Honorio, 2021), for which this is not the case.

2.1. Outcome regression

For our first setting, we wish to use our covariates X to predict the potential outcome Y^a – i.e., what would happen to that individual if he received treatment $T = a$. To do so, we consider a sample-reweighted empirical risk minimization problem: that is, we are solving for

$$\begin{aligned} h^\star &= \arg\min_{h \in L^2} \mathbb{E}[w(X)(h(X) - Y)^2 | T = a] \\ &\approx \arg\min_{h \in L^2} \frac{1}{n_{T=a}} \sum_{T_i = a} w(X_i)(h(X_i) - Y_i)^2. \end{aligned}$$

The purpose of sample reweighing is to bridge the gap between the observed distribution (conditional on $T = a$) and the complete distribution (unconditional on $T = a$). In particular, the idea is that $Y^a|X$ is fixed and that reweighing by $w(X)$ changes the distribution over X from $P_{X|T=a}$ to some $P_{\widetilde{X}|T=a} \approx P_X$, with $\mathrm{d}P_{\widetilde{X}|T=a}/\mathrm{d}P_{X|T=a} = w(X)$.

This procedure forms the basis for many notable algorithms and is a fundamental building block of causal meta-learners (Künzel et al., 2017), one of the main objects of study in this work. The main issue to tackle is quantifying the gap between the observed distribution $Y, X|T = a$ and the complete distribution Y^a, X. To this end, we use Lemma 2.2:

$$\mathbb{E}[(Y^a - h(X))^2] \leq \mathbb{E}[w(X)(Y - h(X))^2|T = a] + \lambda \underbrace{\chi^2(P_{Y,\widetilde{X}|T=a} \| P_{Y^a,X})}_{\Delta_{T=a}} + \frac{1}{4\lambda} \underbrace{\mathrm{Var}[(Y^a - h(X))^2]}_{\sigma^2_{T=a}};$$

We can now leverage the causal nature of our distribution shift in order to precisely quantify this bound. In particular, let us work out the χ^2 term. Under SUTVA and ignorability with regards to X and U and a simple application of Bayes' rule,[2] we have that

$$\frac{\mathrm{d}P_{Y,\widetilde{X},U|T=a}}{\mathrm{d}P_{Y^a,X,U}} = w(X)\frac{\mathrm{d}P_{Y,X,U|T=a}}{\mathrm{d}P_{Y^a,X,U}} = w(X)\frac{\mathrm{d}P_{X,U|T=a}}{\mathrm{d}P_{X,U}} = w(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]},$$

and so

$$\Delta_{T=a} = \chi^2(P_{Y,\widetilde{X}|T=a} \| P_{Y^a,X}) = \mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right],$$

from which we derive our first major result:

Theorem 2.3 (Upper bound on outcome regression loss in expectation)**.** *For any $\lambda > 0$, loss function $\ell(\cdot, \cdot)$ and nonnegative reweighing function $w(X)$ with $\mathbb{E}[w(X)|T = a] = 1$,*

$$\mathbb{E}[\ell(Y^a, h(X))] \leq \mathbb{E}[w(X_i)\ell(Y_i, h(X_i))|T = a] + E,$$

where $\sigma^2 := \mathrm{Var}[\ell(Y^a, h(X))]$ and

$$E = \lambda \underbrace{\mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right]}_{\Delta_{T=a}} + \frac{\sigma^2}{4\lambda}.$$

Let's consider some cases. First, the typical scenario in the meta-learners literature, where $w \equiv 1$: in that case, the $\Delta_{T=a}$ term becomes

$$\Delta_{T=a} = \mathbb{E}\left[\left(\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right],$$

i.e., becomes nil when $\mathbb{P}[T = a|X, U] = \mathbb{P}[T = a]$ almost everywhere – which is the case when the data comes from a randomized control trial. When this is not the case, this term (for $w \equiv 1$) is effectfully a measure of how far the study is from being a randomized control trial.

Interestingly, it implies that small local deviations from an RCT (e.g., because for some small part of the population there was some bias in the treatment assignment) should lead to small deviations between the complete and the observed distributions.

Now, let's consider the case where we choose weights to bridge over the distributions – in particular, consider the optimal weights under the standard ignorability and postivity assumptions sans U (i.e., when there are no unobserved confounders) $w^\star(X) = \mathrm{d}P_{Y^a,X}/\mathrm{d}P_{Y,X|T=a} = \mathbb{P}[T = a]/\mathbb{P}[T = a|X]$. In that case, the divergence term becomes

$$\Delta_{T=a} = \mathbb{E}\left[\left(\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a|X]} - 1\right)^2\right].$$

Under these assumptions, we would have that $\mathbb{P}[T = a|X, U] = \mathbb{P}[T = a|X]$ and so $\Delta_{T=a} = 0$. When there are unobserved confounders, $\Delta_{T=a}$ quantifies by how much more certain we become about the treatment mechanism with their information – i.e., how much the confounders can improve our understanding of the treatment assignment mechanism.

So far, our bounds crucially involve the true propensity scores $\mathbb{P}[T = a|X, U]$, which are unknown and unobservable in practice. We can relax the bound from Theorem 2.3 in order to solve this through the use of a relaxed triangular inequality:

$$\Delta_{T=a} = \mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right] \leq 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - w(X)\frac{\mathbb{1}[T = a]}{\mathbb{P}[T = a]}\right)^2\right] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T = a]}{\mathbb{P}[T = a]} - 1\right)^2\right].$$

Which can then be rearranged into

$$\Delta_{T=a} \leq \frac{2}{\mathbb{P}[T = a]^2}\mathbb{E}\Big[w(X)^2(\mathbb{P}[T = a|X, U] - \mathbb{1}[T = a])^2\Big] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T = a]}{\mathbb{P}[T = a]} - 1\right)^2\right].$$

It can be shown that $\mathbb{P}[T = a|X, U]$ optimizes the term $\mathbb{E}\left[w(X)^2\left(\cdot - \mathbb{1}[T = a]\right)^2\right]$. So, by replacing it with any

[2] SUTVA asserts that data is i.i.d. and that $P_{Y^a|T=a} = P_{Y|T=a}$. Ignorability with respect to X and U states that $Y^1, Y^0 \perp\!\!\!\perp T|X, U$.

$\nu(X)$ the bound remains valid, and we obtain that

$$\Delta_{T=a} \leq \frac{2}{\mathbb{P}[T=a]^2} \underbrace{\mathbb{E}\left[w(X)^2 \left(\nu(X) - \mathbb{1}[T=a]\right)^2\right]}_{\text{BrierScore}_{w^2}(\nu,T)} + 2\,\underbrace{\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2\right]}_{D}.$$

This is a looser bound than the previous one. For example, while the previous one gave us that, under an RCT, $\Delta_{T=a} = 0$, this time we'd conclude merely that $\Delta_{T=a} \leq C \cdot \text{BrierScore}_{w^2}(\nu, T) + 2D$, which is generally greater than zero. Nevertheless, we show in Section 3 that this relaxed bound can still be of great practical use and is sufficiently tight in practice.

Theorem 2.4 (Empirical upper bound on outcome regression loss in expectation)**.** *Let $\Delta_{T=a}$ be as in Theorem 2.3. It holds that, for any ν,*

$$\Delta_{T=a} \leq 2\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2 \left(\nu(X) - \mathbb{1}[T=a]\right)^2\right] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2\right].$$

Finally, we produce PAC-style (Valiant, 1984) finite-sample results for the bound in Theorem 2.4, concretely showing how its estimation would go in practice. For brevity, our results here are based on the Rademacher complexity, but we emphasize that our approach works just as easily with other frameworks (e.g., VC, PAC-Bayes, algorithmic stability).

Corollary 2.5 (PAC empirical upper bound on outcome regression loss)**.** *Suppose that ℓ is a loss function bounded in $[0, M]$, and $w(X)$ is a nonnegative reweighing function bounded in $[0, w_{\max}]$ with $\mathbb{E}[w(X)|T = a] = 1$. Then, for any $\lambda > 0$, with probability at least $1 - \delta$ over the draw of the training data $(X_i, T_i, Y_i)_{i=1}^n$, for all $h \in \mathcal{H}$ and $\nu \in \mathcal{H}_\nu$,*

$$\mathbb{E}[\ell(Y^a, h(X))] \leq \frac{1}{n_{T=a}} \sum_{T_i=a} w(X_i)\ell(Y_i, h(X_i)) + \lambda\widehat{\Delta}_{T=a} + \frac{M^2}{16\lambda} + 2\mathfrak{R}(w \cdot \ell \circ \mathcal{H}) + 2\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu) + \left(M w_{\max} + C(w_{\max})\sqrt{n_{T=a}/n}\right)\sqrt{\frac{\log 2/\delta}{2n_{T=a}}}$$

where

$$\widehat{\Delta}_{T=a} := \frac{2\lambda}{n}\sum_{i=1}^n \left(\frac{w(X_i)}{\mathbb{P}[T=a]}\right)^2 \left(\nu(X_i) - \mathbb{1}[T_i=a]\right)^2 + \frac{2\lambda}{n}\sum_{i=1}^n \left(w(X_i)\frac{\mathbb{1}[T_i=a]}{\mathbb{P}[T=a]} - 1\right)^2,$$

$\mathfrak{R}(w \cdot \ell \circ \mathcal{H})$ and $\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu)$ are the Rademacher complexities of $\mathcal{H}$ and $\mathcal{H}_\nu$ composed with their respective loss functions/means, and $C(w_{\max})$ is a constant nonnegative quantity defined in the proof.

The right-hand-side of this bound has many terms: first is the empirical observable loss, reweighed by our $w(X)$. Next is the empirical bound for $\Delta_{T=a}$, $\widehat{\Delta}_{T=a}$, along with $M^2/16$ corresponding to the variance term $\sigma^2/4$, both reweighed by the λ as in Lemma 2.2. Next are the Rademacher complexities $\mathfrak{R}(\mathcal{H})$ and $\mathfrak{R}(\mathcal{H}_\nu)$ correcting for the complexity of the algorithms we are fitting. Finally, we have a term bounding the tail of the distribution of the reweighted empirical loss, notably as a factor of $M w_{\max}$, which comes divided by the number of observable samples.

This bound neatly exhibits the cost of using a reweighing w (e.g., the optimal $w^\star$) to bridge over the distributions: if there is some point that is extremely unlikely to be observed, then that point will have a very large weight, increasing $w_{\max}$ substantially. Thankfully, this is not irremediable: by "simply" using more data, the effect of this term on the bound can be reduced until it no longer dominates the rest.

2.2. Causal Meta-learners

We now shift our attention to using our covariates X to predict the individual treatment effects $Y^1 - Y^0$ – i.e., what would be the benefit of applying the treatment versus doing nothing to this particular individual.

We are particularly interested in causal meta-learners: procedures that leverage previously-established (or purpose-built) regression algorithms as oracles in an estimating equation. We focus here on the most common meta-learners available: the T-learner, the S-learner and the X-learner (Künzel et al., 2017). Nevertheless, our analysis is also directly applicable to many other common causal regression methods such as the Causal BART (Hill, 2011; Hahn et al., 2017) and TARNet/CFRNet (Shalit et al., 2017), since these can be seen as special cases of T- or S-learners.

Throughout this section, we focus on losses that satisfy a sort of relaxed triangular inequality:

Assumption 2.6. The loss function $\ell(\cdot, \cdot)$ can be restated as

$$\ell(Y, \widehat{Y}) = \psi(Y - \widehat{Y}),$$

with ψ satisfying a relaxed subadditive condition: there is some C such that, for any x, y, $\psi(x \pm y) \leq C(\psi(x) + \psi(y))$.

Most loss functions of interest satisfy Assumptions 2.6:

- Mean Squared Error: take $\psi(x) = x^2$ and $C = 2$.
- Mean Absolute Error: take $\psi(x) = |x|$ and $C = 1$.

- α-Quantile Loss: take $\psi(x) = x\alpha\mathbb{1}[x \geq 0] - x(1-\alpha)\mathbb{1}[x < 0]$ and $C = \max\{\alpha, 1-\alpha\}/\min\{\alpha, 1-\alpha\}$.
- 0-1 Loss[3]: take $\psi(x) = |x|$ and $C = 1$.

Such an assumption allows us to employ a sort of relaxed triangle inequality in order to decompose the loss of a meta-learner into the losses of its individual components. This technique has been widely applied in previous works (e.g. (Künzel et al., 2017; Johansson et al., 2022)), though previously restricted to only the mean squared error. Through the introduction of Assumption 2.6, we are able to make our results substantially more general – the implications of which we discuss in Section 2.3.

2.2.1. T-LEARNERS AND S-LEARNERS

In these types of models, we first obtain two functions h^1 and h^0 that have been trained to approximate $Y|X, T = 1$ and $Y|X, T = 0$ respectively. In the case of T-learners, this is done by independently training h^1 and h^0 on the samples where $T = 1$ and $T = 0$, while for S-learners it is done by training a single model $h : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ to use X and T to predict Y, and taking $h^1 = h(\cdot, 1)$ and $h^0 = h(\cdot, 0)$. With h^1 and h^0 in hand, we can predict the individual treatment effect as $h^1(X) - h^0(X)$.

To bound the losses of our predictions $h^1(X) - h^0(X)$, we can separate it in terms of the losses of the individual functions h^1 and h^0 through the use of Assumption 2.6:

$$\begin{aligned}&\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))]\\&\quad\leq C(\mathbb{E}[\ell(Y^1, h^1(X))] + \mathbb{E}[\ell(Y^0, h^0(X))])\end{aligned}$$

From there on, we can use the bounds developed in Section 2.1 to obtain bounds in terms of the observable losses:

Proposition 2.7 (Upper bound on T-/S-learner loss in expectation)**.** *Let ℓ be a loss function satisfying Assumption 2.6 with constant C, and let $w^1(X), w^0(X)$ be nonnegative reweighing functions with $\mathbb{E}[w^a(X)|T = a] = 1$ for $a = 0, 1$. For any $\lambda_1, \lambda_0 > 0$,*

$$\begin{aligned}&\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))]\\&\quad\leq C\Big(\mathbb{E}[w^1(X)\ell(Y, h^1(X))|T = 1]\\&\qquad+ \mathbb{E}[w^0(X)\ell(Y, h^0(X))|T = 0]\\&\qquad+ \underbrace{\lambda_1\Delta_{T=1} + \lambda_0\Delta_{T=0}}_{\Delta_{\mathrm{T-learner}}} + \underbrace{\sigma^2_{T=1}/4\lambda_1 + \sigma^2_{T=0}/4\lambda_0}_{\sigma^2/4}\Big),\end{aligned}$$

where $\sigma^2_{T=a} := \mathrm{Var}[\ell(Y^a, h^a(X))]$ and

$$\Delta_{T=a} = \mathbb{E}\left[\left(w^a(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right]$$

$$\begin{aligned}&\leq \frac{2}{\mathbb{P}[T = a]^2}\mathbb{E}\left[w^a(X)^2\left(\nu(X) - \mathbb{1}[T = a]\right)^2\right]\\&\quad+ 2\mathbb{E}\left[\left(w^a(X)\frac{\mathbb{1}[T = a]}{\mathbb{P}[T = a]} - 1\right)^2\right].\end{aligned}$$

We can also produce a finite-sample bound akin to Corollary 2.5. Again, we only present here a bound based on the Rademacher complexity, but the same ideas can be applied to other paradigms.

Corollary 2.8 (PAC empirical upper bound on the loss of a T-/S-learner)**.** *Let ℓ be a loss function bounded in $[0, M]$ satisfying Assumption 2.6 with constant C and let $w^1(X)$, $w^0(X)$ be nonnegative reweighing functions bounded in $[0, w_{max}]$ and with $\mathbb{E}[w^a(X)|T = a] = 1$ for $a = 0, 1$. Then, for any $\lambda_1, \lambda_0 > 0$, with probability at least $1 - \delta$ over the draw of the training data $(X_i, T_i, Y_i)_{i=1}^n$, for all $h^1, h^0 \in \mathcal{H}$ and $\nu \in \mathcal{H}_\nu$,*

$$\begin{aligned}&\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))]\\&\leq C\Bigg(\frac{1}{n_{T=1}}\sum_{T_i=1} w^1(X_i)\ell(Y_i, h^1(X_i))\\&\quad+ \frac{1}{n_{T=0}}\sum_{T_i=0} w^0(X_i)\ell(Y_i, h^0(X_i))\\&\quad+ \lambda_1\widehat{\Delta}_{T=1} + \lambda_0\widehat{\Delta}_{T=0} + \frac{M^2}{16\lambda_1} + \frac{M^2}{16\lambda_0}\\&\quad+ 2\mathfrak{R}(w^1 \cdot \ell \circ \mathcal{H}) + 2\mathfrak{R}(w^0 \cdot \ell \circ \mathcal{H}) + 2\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu)\\&\quad+ \left(cMw_{\max} + C(w_{\max})\sqrt{n_{T=\min}/n}\right)\sqrt{\frac{\log 3/\delta}{2n_{T=\min}}}\Bigg)\end{aligned}$$

where

$$\begin{aligned}\widehat{\Delta}_{T=a} :=& \frac{2\lambda}{n}\sum_{i=1}^n\left(\frac{w^a(X_i)}{\mathbb{P}[T = a]}\right)^2(\nu(X_i) - \mathbb{1}[T_i = a])^2\\&+ \frac{2\lambda}{n}\sum_{i=1}^n\left(w^a(X_i)\frac{\mathbb{1}[T_i = a]}{\mathbb{P}[T = a]} - 1\right)^2,\end{aligned}$$

$n_{T=\min} = \min\{n_{T=1}, n_{T=0}\}$, $\mathfrak{R}(w^a \cdot \ell \circ \mathcal{H})$ and $\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu)$ are the Rademacher complexities of $\mathcal{H}$ and $\mathcal{H}_\nu$ composed with their respective loss functions/means, and c and $C(w_{\max})$ are a constant nonnegative quantities defined in the proof, with $1 \leq c \leq 2$.

It should be noted that these bounds do not account for interactions between h^1 and h^0. Quantifying such interactions requires some consideration for the structure of the models themselves, which would lead to losing the mostly-model-agnostic flavor of our results.

2.2.2. X-LEARNERS

An X-learner follows a slightly more involved procedure than the T- and S-learners. First, one estimates h^1 and

[3] In the case of binary classification, taking both the label and predictions $Y, \widehat{Y}$ to assume numeric values in $\{0, 1\}$.

h^0 as in a T-learner. With that, new models τ^1 and τ^0 are trained to regress on $Y^1 - h^0(X)$ and $h^1(X) - Y^0$, respectively, as some sort of pseudo-treatment-effect labels. The individual treatment effect is then estimated as $e(X)\tau^1(X) + (1 - e(X))\tau^0(X)$, for some $e(x)$ (often an estimate of the propensity score $\mathbb{P}[T = 1|X]$).

Similar to what we did in Section 2.2.1, we will bound the loss of the X-learner estimator by separating it in terms of the losses of its individual components h^1, h^0, τ^1 and τ^0.

For ease of notation, let $\bar{e}(x) = 1 - e(x)$ and $\ell_e(Y, \widehat{Y}) := \ell(e(X)Y, e(X)\widehat{Y})$. Then, by Assumption 2.6:

$$\begin{aligned}
&\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + \bar{e}(X)\tau^0(X))] \\
&\leq C^2(\mathbb{E}[\ell_{\bar{e}}(Y^1, h^1(X))] + \mathbb{E}[\ell_e(Y^0, h^0(X))] \\
&\qquad + \mathbb{E}[\ell_e(Y^1 - h^0(X), \tau^1(X))] \\
&\qquad + \mathbb{E}[\ell_{\bar{e}}(h^1(X) - Y^0, \tau^0(X))]).
\end{aligned}$$

And again, all that remains is to use the bounds from Section 2.1 to obtain observable bounds for the complete loss of the X-learner.

Proposition 2.9 (Upper bound on X-learner loss in expectation)**.** *Let ℓ be a loss function satisfying Assumption 2.6 with constant C and let w^1, w^0 be nonnegative reweighting functions with $\mathbb{E}[w^a(X)|T = a] = 1$ for $a = 0, 1$. For any $\lambda_1, \lambda_0, \lambda_{0,1}, \lambda_{1,0} > 0$,*

$$\begin{aligned}
&\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + \bar{e}(X)\tau^0(X))] \\
&\leq C^2\Big(\mathbb{E}[w^1(X)\ell_{\bar{e}}(Y, h^1(X))|T = 1] \\
&+ \mathbb{E}[w^0(X)\ell_e(Y, h^0(X))|T = 0] \\
&+ \mathbb{E}[w^1(X)\ell_e(Y^1 - h^0(X), \tau^1(X))|T = 1] \\
&+ \mathbb{E}[w^0(X)\ell_{\bar{e}}(h^1(X) - Y^0, \tau^0(X))|T = 0] \\
&+ \underbrace{(\lambda_1 + \lambda_{1,0})\Delta_{T=1} + (\lambda_0 + \lambda_{0,1})\Delta_{T=0}}_{\Delta_{\text{X-learner}}} \\
&+ \underbrace{\sigma^2_{T=1}/4\lambda_1 + \sigma^2_{T=0}/4\lambda_0 + \sigma^2_{1,0}/4\lambda_{1,0} + \sigma^2_{0,1}/4\lambda_{0,1}}_{\sigma^2/4}\Big),
\end{aligned}$$

where $\sigma^2_{T=a} := \mathrm{Var}[\ell(Y^a, h^a(X))]$, $\sigma^2_{a,b} := \mathrm{Var}[\ell(Y^a - h^b(X), \tau^a(X))]$ and

$$\begin{aligned}
\Delta_{T=a} &= \mathbb{E}\left[\left(w^a(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right] \\
&\leq \frac{2}{\mathbb{P}[T = a]^2}\mathbb{E}\left[w^a(X)^2\left(\nu(X) - \mathbb{1}[T = a]\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\left(w^a(X)\frac{\mathbb{1}[T = a]}{\mathbb{P}[T = a]} - 1\right)^2\right].
\end{aligned}$$

We can also obtain finite-sample bounds akin to our Corollary 2.8, which can be found in Appendix A as Corollary A.13.

Just as in our analysis of T- and S- learners, we are blind to potential improvements to the bounds due to interactions between the individual regressions, in the spirit of keeping our analysis reasonably model-agnostic.

With both bounds for T/S-learners and for X-learners in hand, a natural question arises: can our theory suggest when one should be preferred over the other? Since our bounds for these classes of models share many terms, it is not clear a priori whether one method should be preferred over the other. In the end, it appears that any possible advantage originates from the potentially improved inductive biases of the individual regressions. This is consistent with previous findings (Künzel et al., 2017; Dorie et al., 2017).

2.3. Beyond the Mean Squared Loss

One remarkable aspect of the bounds developed in Section 2.2 is how they are general in relation to the loss. That is, they can limit other losses such as the mean absolute error (for robust regression) and the quantile loss (for quantile regression) just as well as the more typical mean squared loss. Moreover, minimizing the same loss on each component of the meta-learner implies minimizing the total loss of the meta-learner as a whole.

This is a rather surprising result. For instance, it was previously understood that the conditional quantile treatment effect $Q_\alpha(Y^1 - Y^0|X)$ was unidentifiable in general. Yet we show that by training a T-learner optimizing the quantile loss, for example, we can estimate it by first approximating $h^1(X) \approx Q_\alpha(Y^1|X)$ and $h^0 \approx Q_\alpha(Y^0|X)$, leading to

$$h^1(X) - h^0(X) \approx Q_\alpha(Y^1 - Y^0|X).$$

This is counterintuitive, since $Q_\alpha(Y^1|X) - Q_\alpha(Y^0|X) \neq Q_\alpha(Y^1 - Y^0|X)$ in general! The key to our argument is that we are *not* asserting that $h^1(X) - h^0(X)$ will always converge to $Q_\alpha(Y^1 - Y^0|X)$. We have bounded the complete quantile loss of the T-learner, but we have never proven that it is even close to the optimal one (which is what would then imply convergence to the conditional quantile).

Nevertheless, we do show that by improving how well we estimate the individual conditional quantiles (in terms of the quantile loss) we also improve how well we estimate the conditional quantile of the treatment effect. At the limit, if h^1 and h^0 achieve zero quantile loss (meaning they predict the potential outcomes exactly) and we have a negligible divergence term (e.g., $\Delta = 0$), then the quantile loss for the treatment effect will also be zero (meaning we will predict the individual treatment effects exactly).

3. Experiments and Applications

In order to assess the tightness and efficacy of our bounds, we empirically evaluate them on three semi-synthetic

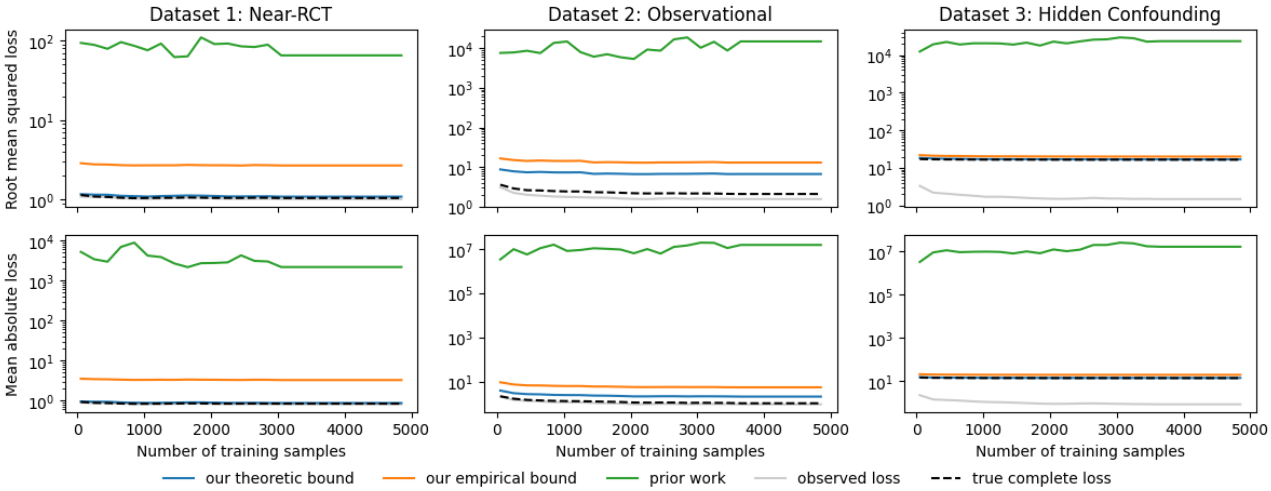


Figure 1. **Tightness of our bounds.** Comparison between our bounds and those of (Johansson et al., 2022), both for the complete loss of the estimation of the potential outcome Y^1. Additional images for other tasks (e.g., estimation of treatment effects) are available in Appendix C. Our "theoretic" and "empirical" bounds correspond in Theorems 2.3 and 2.4, and "prior work" refers to Corollary 1 of (Johansson et al., 2022). Our theoretic bound is quite tight, being very close to the complete loss (which is unobservable in practice). Our empirical bound, while somewhat looser than the theoretic bound, is still *substantially* tighter than the available prior work.

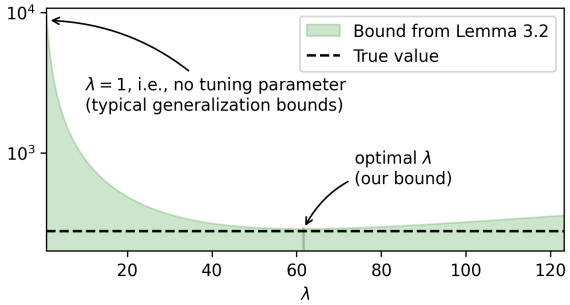


Figure 2. **Importance of the tuning parameter λ in Lemma 2.2.** An illustration of the bound in Lemma 2.2 (shaded in green) over different values of its tuning parameter λ. Change-of-measure inequalities (e.g., (Ohnishi & Honorio, 2021)) typically do not have a tuning parameter, which corresponds to taking $\lambda = 1$ in our lemma. As can be seen in the figure, being able to optimally select λ substantially tightens our bounds.

datasets of increasing difficulty: first simulating a randomized control trial and later in simulations of observational data with fully observed and hidden confounders. Finally, we tackle a real-world application of model selection in a Parkinson's telemonitoring dataset.

Experiments were run on an AMD Ryzen 9 5950X CPU (2.2GHz/5.0GHz, 32 threads) with 64GB of RAM. Nevertheless, the relevant code is lightweight and should easily run on weaker hardware. More details can be found in Appendix B, and the code is available at `https://github.com/dccsillag/experiments-causal-generalization-bounds`.

3.1. Experiments on Semi-Synthetic Data

To evaluate our procedures we use semi-synthetic data (i.e., synthetic data that attempts to reproduce real data) as this allows us to have access to the potential outcomes Y^1, Y^0.

We use the following datasets, ordered by increasing difficulty. **Learned IHDP**: Results of a randomized control trial simulated with generative models trained on the IHDP (Hill, 2011) dataset. **ACIC16**: Simulated observational data from (Dorie et al., 2017) with fully observed confounding, satisfying ignorability and positivity assumptions; **Confounded ACIC16**: Observational data with significant unobserved confounding and no positivity. This is also the ACIC16 dataset, but modified so that there are hidden confounders. This does not satisfy ignorability nor positivity, making it an extremely challenging dataset to work with.

Figure 1 demonstrates the tightness of our bounds, especially in comparison to the prior work of (Johansson et al., 2022). In it, we visualize the observed loss, the (unobservable) complete loss, our bounds and the closest corresponding bound from (Johansson et al., 2022), varying the number of training samples, all for the task of estimating the potential outcome Y^1.

Our theoretical bound is remarkably tight both in the Near-RCT data and in the observational datasets with hidden confounding, matching the complete loss almost exactly. Furthermore, in the hidden confounding case even our empirical bound – which is generally looser than the theoretic one – is exceptionally tight, possibly due to the inapplicability of the positivity assumption. In the observable dataset

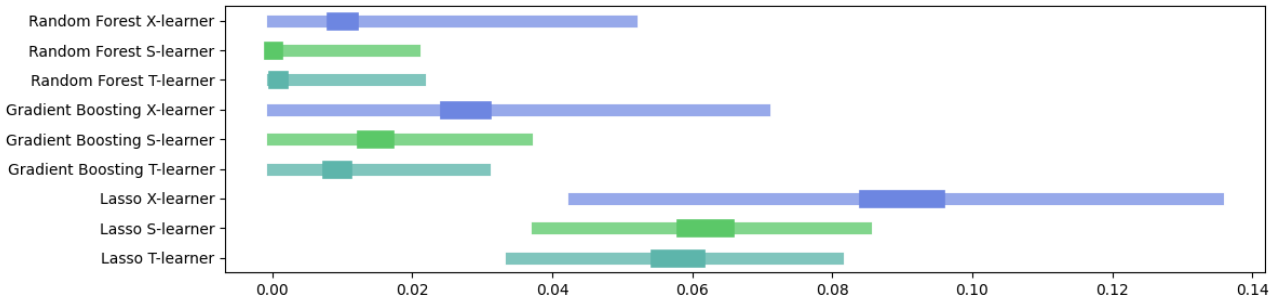


Figure 3. **Application: model selection on real data.** The plot compares multiple models for treatment effect estimation: bars correspond to our bounds on the complete loss of the models, while the knobs in the middle correspond to standard bootstrapped confidence intervals for the loss on the observed distributions. Note how some models (e.g., R.F. T-learner) appear strictly better than others (e.g., G.B. X-learner) only if ours bounds are not considered. The R.F. and G.B. T/S-learners remain strictly better than the Lasso-based models.

with fully observed confounding there is a wider gap between our bounds and the complete loss, though they are all about the same order of magnitude. On all accounts, the previous bound of (Johansson et al., 2022) is multiple orders of magnitude looser.

A significant contributor to the tightness of our bounds is the tuning parameter λ in Lemma 2.2. Figure 2 showcases its importance, displaying the bound from Lemma 2.2 on our dataset with hidden confounding for various values of λ. Most existing bounds correspond to taking $\lambda = 1$ (e.g., those of (Ohnishi & Honorio, 2021) and those of (Johansson et al., 2022), though it is arguable whether such a tuning parameter could even be introduced into their IPM-based bounds). By allowing optimization over the λ parameter we introduced, our bounds can become over $30\times$ tighter.

3.2. Application on real data

In this section, we work on top of the Parkinson's Telemonitoring dataset of (Tsanas et al., 2009), containing features from voice measurements paired with standardized scores describing the progression of the disease, along with the subject's age, gender, and time since recruitment to the study. The goal is to assess the effect of sex on the progression of the disease. There are likely unobserved confounders (e.g., not enough data about the subject to construct a baseline), and at the same time there is enough data to nearly violate positivity (voice data can be a good predictor of sex).

In order to evaluate a causal link between sex and disease progression, we wish to train a model that estimates the individual treatment effect of the standardized UPDRS score on the subject being male vs. female. To this end, we have trained a myriad of models consisting of T-learners, S-learners and X-learners based on a Lasso, Gradient Boosting and Random Forests.

To assess the quality of our models, we utilize our bounds from Propositions 2.7 and 2.9. We limit the Δ term via our empirial bounds, and posit a conservative upper bound on the variance of the loss; we note that, on the observed distributions, the loss typically has a variance of around 10^{-6} to 10^{-8}, leading us to conservatively estimate the variance of the loss as below 10^{-5} to give some margin to increased deviation in the complete distribution.

The results of this procedure can be seen in Figure 3. Prior to the "correction" given by our bounds, we would have confidently concluded that the Random Forest-based S- and T-learners outperformed all others. After considering the gap between the observed and complete distributions, however, this no longer appears to be the case, with these models now performing roughly on par with the gradient boosting-based meta-learners. However, even after this correction, it is clear that there is an advantage to using more expressive models over our linear meta-learners, these being confidently outperformed by the Random Forest backed T- and S-learners. Finally, the use of an X-learner does not seem worth it. They are worse in terms of the observable loss and have wider error bars due to the bound – one should probably opt here instead for either the Random Forest T- or S-learner.

4. Conclusion

In this work, we've introduced generalization bounds for outcome regression and treatment effect estimation, establishing rigorously that modern causal ML procedures *can* properly estimate causal quantities. Our results hold regardless of hidden confounding or lack of positivity. Moreover, our bounds show that existing procedures for CATE estimation can be adapted for other tasks, such as estimation of quantiles of individual treatment effects – which was previously thought to be infeasible – simply by changing the loss function in the individual optimizations. We also empirically showcase the tightness and practical utility of our bounds on semi-synthetic and real data, where it is shown

to vastly outperform the bounds from the closest matching prior work. We expect that our results can be of great use not only as motivation for using existing algorithms, but also as a backbone for new algorithms on diverse causal inference tasks.

Acknowledgements

We would like to thank the reviewers for their helpful and constructive comments. This work was partially funded by CNPq and FAPERJ.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and Causality. There are many potential societal consequences of our work through applications to various domains such as medicine and economics, but none which we feel must be specifically highlighted here.

References

Angrist, J. D. and Imbens, G. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90:431–442, 1995. URL https://api.semanticscholar.org/CorpusID:8384694.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113:7353 – 7360, 2015. URL https://api.semanticscholar.org/CorpusID:16171120.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 2016. URL https://api.semanticscholar.org/CorpusID:51735142.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. C., and Vaughan, J. W. Learning bounds for domain adaptation. In *Neural Information Processing Systems*, 2007. URL https://api.semanticscholar.org/CorpusID:2497886.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Neural Information Processing Systems*, 2010. URL https://api.semanticscholar.org/CorpusID:2555196.

Curth, A. and van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021. URL https://api.semanticscholar.org/CorpusID:231709566.

Dorie, V., Hill, J. L., Shalit, U., Scott, M. A., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 2017. URL https://api.semanticscholar.org/CorpusID:51992418.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018. URL https://api.semanticscholar.org/CorpusID:84834062.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A new pac-bayesian perspective on domain adaptation. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 859–868, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/germain16.html.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. Pac-bayes and domain adaptation. *Neurocomputing*, 379:379–397, 2017. URL https://api.semanticscholar.org/CorpusID:53493590.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Econometrics: Multiple Equation Models eJournal*, 2017. URL https://api.semanticscholar.org/CorpusID:34019969.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217 – 240, 2011. URL https://api.semanticscholar.org/CorpusID:122155840.

Hirano, K. and Imbens, G. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2:259–278, 2001. URL https://api.semanticscholar.org/CorpusID:3346892.

Johansson, F. D., Shalit, U., and Sontag, D. A. Learning representations for counterfactual inference. *ArXiv*, abs/1605.03661, 2016. URL https://api.semanticscholar.org/CorpusID:8558103.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022. URL http://jmlr.org/papers/v23/19-511.html.

Jolicoeur-Martineau, A., Fatras, K., and Kachman, T. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees, 2023.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116:4156 – 4165, 2017. URL https://api.semanticscholar.org/CorpusID:73455742.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *ArXiv*, abs/0902.3430, 2009. URL https://api.semanticscholar.org/CorpusID:6178817.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. Foundations of machine learning. In *Adaptive computation and machine learning*, 2012. URL https://api.semanticscholar.org/CorpusID:263010642.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2017. URL https://api.semanticscholar.org/CorpusID:85529052.

Ohnishi, Y. and Honorio, J. Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1711–1719. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/ohnishi21a.html.

Oprescu, M., Dorn, J., Ghoummaid, M., Jesson, A., Kallus, N., and Shalit, U. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:258291549.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ruderman, A., Reid, M. D., García-García, D., and Petterson, J. Tighter variational representations of f-divergences via restriction to probability measures. *ArXiv*, abs/1206.4664, 2012. URL https://api.semanticscholar.org/CorpusID:288983.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57:884–893, 2009. URL https://api.semanticscholar.org/CorpusID:7382779.

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27:1134–1142, 1984. URL https://api.semanticscholar.org/CorpusID:59712.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113:1228 – 1242, 2015. URL https://api.semanticscholar.org/CorpusID:15676251.

Yoon, J., Jordon, J., and van der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:65516833.

Zhang, Y., Bellot, A., and van der Schaar, M. Learning overlapping representations for the estimation of individualized treatment effects. *ArXiv*, abs/2001.04754, 2020. URL https://api.semanticscholar.org/CorpusID:210473399.

A. Theoretical Results

A.1. Change of Measure

Lemma A.1 (Lemma 2.2 in the main body)**.** *Let P and Q be as in Definition 2.1. For any $\lambda > 0$,*

$$\mathbb{E}_Q[\phi] - E \leq \mathbb{E}_P[\phi] \leq \mathbb{E}_Q[\phi] + E$$

where

$$E := \lambda \cdot \chi^2(Q\|P) + \frac{1}{4\lambda}\operatorname{Var}_P(\phi).$$

Moreover, the bound is optimized for

$$\lambda^\star = \sqrt{\operatorname{Var}_P(\phi)/4\chi^2(Q\|P)}.$$

Proof. We build upon the variational representation framework of (Ruderman et al., 2012) and (Ohnishi & Honorio, 2021). By Lemma 2 of (Ohnishi & Honorio, 2021), we have that, for all R and S, for any ψ,

$$\mathbb{E}_Q[\psi] \leq \mathbb{E}_P[\psi] + \chi^2(Q\|P) + \frac{1}{4}\operatorname{Var}_P[\psi].$$

Consider two possible choices of ψ: $\psi_0 = \lambda\phi$ and $\psi_1 = -\lambda\phi$. We then have that:

$$\mathbb{E}_Q[\lambda\phi] = \lambda\mathbb{E}_Q[\phi] \leq \mathbb{E}_P[\lambda\phi] + \chi^2(Q\|P) + \frac{1}{4}\operatorname{Var}_P[\lambda\phi] = \lambda\mathbb{E}_P[\phi] + \chi^2(Q\|P) + \frac{1}{4}\lambda^2\operatorname{Var}_P[\phi]$$
$$\mathbb{E}_Q[-\lambda\phi] = -\lambda\mathbb{E}_Q[\phi] \leq \mathbb{E}_P[-\lambda\phi] + \chi^2(Q\|P) + \frac{1}{4}\operatorname{Var}_P[-\lambda\phi] = -\lambda\mathbb{E}_P[\phi] + \chi^2(Q\|P) + \frac{1}{4}\lambda^2\operatorname{Var}_P[\phi].$$

Rearranging, we obtain that, for any $\lambda > 0$,

$$\mathbb{E}_Q[\phi] - \lambda \cdot \chi^2(Q\|P) - \frac{1}{4\lambda}\operatorname{Var}_P[\phi] \leq \mathbb{E}_P[\phi] \leq \mathbb{E}_Q[\phi] + \lambda \cdot \chi^2(Q\|P) + \frac{1}{4\lambda}\operatorname{Var}_P[\phi].$$

Finally, this bound is tightest by optimizing

$$\lambda^\star = \arg\min_\lambda \left(\mathbb{E}_Q[\phi] + \lambda \cdot \chi^2(Q\|P) + \frac{1}{4\lambda}\operatorname{Var}_P(\phi)\right) = \arg\min_\lambda \left(\lambda \cdot \chi^2(Q\|P) + \frac{1}{4\lambda}\operatorname{Var}_P(\phi)\right).$$

And taking the derivative in respect to λ and equating it to zero gives the desired result. □

A.2. Bounds in Expectation

Theorem A.2 (Theorem 2.3 in the main body)**.** *For any $\lambda > 0$, loss function $\ell(\cdot,\cdot)$ and nonnegative reweighing function $w(X)$ with $\mathbb{E}[w(X)|T = a] = 1$,*

$$\mathbb{E}[\ell(Y^a, h(X))] \leq \mathbb{E}[w(X_i)\ell(Y_i, h(X_i))|T = a] + E,$$

where $\sigma^2 := \operatorname{Var}[\ell(Y^a, h(X))]$ and

$$E = \lambda \underbrace{\mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T = a|X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right]}_{\Delta_{T=a}} + \frac{\sigma^2}{4\lambda}.$$

Proof. First note that since $\mathbb{E}[w(X)|T = a] = 1$, the reweighting $w(X)$ induces a distribution $P_{\widetilde{X}}$ over the X with $\mathrm{d}P_{\widetilde{X}}/\mathrm{d}P_{X|T=1} = w(X)$.

By Lemma 2.2:

$$\mathbb{E}[(Y^a - h(X))^2] \leq \mathbb{E}[w(X)(Y^a - h(X))^2|T = a] + \lambda\chi^2(P_{Y^a,\widetilde{X}|T=a}\|P_{Y^a,X}) + \frac{1}{4\lambda}\operatorname{Var}[(Y^a - h(X))^2]; \quad (1)$$

Under SUTVA and ignorability with regards to X and U and using the fact that $\mathrm{d}P_{\widetilde{X}}/\mathrm{d}P_{X|T=1} = w(X)$ and a simple application of Bayes' rule, we have that

$$\begin{aligned}
\frac{\mathrm{d}P_{Y,\widetilde{X},U|T=a}}{\mathrm{d}P_{Y,X,U}} &= \frac{\mathrm{d}P_{Y,U|\widetilde{X},T=a}\cdot \mathrm{d}P_{\widetilde{X}|T=a}}{\mathrm{d}P_{Y^a,X,U}} = w(X)\frac{\mathrm{d}P_{Y,U|X,T=a}\cdot \mathrm{d}P_{X|T=a}}{\mathrm{d}P_{Y^a,X,U}} = w(X)\frac{\mathrm{d}P_{Y,X,U|T=a}}{\mathrm{d}P_{Y^a,X,U}} \\
&= w(X)\frac{\mathrm{d}P_{Y^a,X,U|T=a}}{\mathrm{d}P_{Y^a,X,U}} = w(X)\frac{\mathrm{d}P_{Y^a|X,U,T=a}\cdot \mathrm{d}P_{X,U|T=a}}{\mathrm{d}P_{Y^a|X,U}\cdot \mathrm{d}P_{X,U}} = w(X)\frac{\mathrm{d}P_{Y^a|X,U}\cdot \mathrm{d}P_{X,U|T=a}}{\mathrm{d}P_{Y^a|X,U}\cdot \mathrm{d}P_{X,U}} \\
&= w(X)\frac{\mathrm{d}P_{X,U|T=a}}{\mathrm{d}P_{X,U}} = w(X)\frac{\mathrm{d}P_{T=a|X,U}\cdot \mathrm{d}P_{X,U}/\mathrm{d}P_{T=a}}{\mathrm{d}P_{X,U}} = w(X)\frac{\mathrm{d}P_{T=a|X,U}}{\mathrm{d}P_{T=a}} \\
&= w(X)\frac{\mathbb{P}[T=a|X,U]}{\mathbb{P}[T=a]},
\end{aligned}$$

and thus

$$\chi^2(P_{Y,\widetilde{X}|T=a}\|P_{Y^a,X}) = \mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T=a|X,U]}{\mathbb{P}[T=a]}-1\right)^2\right].$$

Plugging this into Equation 1, we conclude. □

Theorem A.3 (Theorem 2.4 in the main body)**.** *Let $\Delta_{T=a}$ be as in Theorem 2.3. It holds that, for any ν,*

$$\Delta_{T=a} \leq 2\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2\right] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right].$$

Proof. We can relax the bound from Theorem 2.3 in order to solve this through the use of a relaxed triangular inequality:

$$\begin{aligned}
\Delta_{T=a} &= \mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T=a|X,U]}{\mathbb{P}[T=a]}-1\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T=a|X,U]}{\mathbb{P}[T=a]}-w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}\right)^2\right] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right]
\end{aligned}$$

Which can then be rearranged into

$$\frac{2}{\mathbb{P}[T=a]^2}\mathbb{E}\left[w(X)^2\left(\mathbb{P}[T=a|X,U]-\mathbb{1}[T=a]\right)^2\right] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right].$$

Note that $\mathbb{E}[w(X)^2\left(\nu(X)-\mathbb{1}[T=a]\right)^2]$ is the Brier Score (or Mean Squared Error) of $\nu(X)$ predicting $\mathbb{1}[T=a]$, reweighted by the covariates X according to $w(X)$. Therefore, it is minimized by $\nu^\star(X,U) = \mathbb{E}[\mathbb{1}[T=a]|X,U] = \mathbb{P}[T=a|X,U]$. Therefore, it holds that

$$\forall nu. \quad \mathbb{E}\left[w(X)^2\left(\mathbb{P}[T=a|X,U]-\mathbb{1}[T=a]\right)^2\right] \leq \mathbb{E}\left[w(X)^2\left(\nu(X)-\mathbb{1}[T=a]\right)^2\right],$$

And we obtain that

$$\Delta_{T=a} \leq \frac{2}{\mathbb{P}[T=a]^2}\mathbb{E}\left[w(X)^2\left(\nu(X)-\mathbb{1}[T=a]\right)^2\right] + 2\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right].$$

□

Before we continue, let us prove a fundamental lemma about the pinball function:

Lemma A.4. *Let* $\mathrm{pinball}_\alpha(x) = x\alpha\mathbb{1}[x\geq 0] - x(1-\alpha)\mathbb{1}[x<0]$. *Then, for all $x\in\mathbb{R}$ and $\alpha\in(0,1)$, it holds that*

$$x\alpha \leq \mathrm{pinball}_\alpha(x) \qquad \textit{and} \qquad -x(1-\alpha) \leq \mathrm{pinball}_\alpha(x).$$

Proof. To prove that $x\alpha \leq \text{pinball}_\alpha(x)$, we consider two cases:

1. If $x \geq 0$, then $\text{pinball}_\alpha(x) = x\alpha$ and equality holds trivially.
2. If $x < 0$, then $x\alpha < 0$. But certainly $\text{pinball}_\alpha(x) = -x(1-\alpha) \geq 0$, so the inequality holds.

To prove that $-x(1-\alpha) \leq \text{pinball}_\alpha(x)$ we similarly consider two cases:

1. If $x > 0$, then $-x(1-\alpha) < 0$. But certainly $\text{pinball}_\alpha(x) = x\alpha \geq 0$, so the inequality holds.
2. If $x \leq 0$, then $\text{pinball}_\alpha(x) = -x(1-\alpha)$, and the equality holds trivially.

□

Proposition A.5 (Applicability of Assumption 2.6)**.** *Let $\ell(\cdot,\cdot)$ be the mean squared error, mean absolute error, quantile loss or 0-1 loss. Then it holds that there exists some C and ψ that satisfy Assumption 2.6 for ℓ.*

Proof. For the mean squared error, write $\ell(y,\widehat{y}) = (y-\widehat{y})^2$, thus having $\psi(x) = x^2$. Then, by a relaxed triangular inequality, $(y \pm \widehat{y})^2 \leq 2 \cdot (y^2 + (\pm\widehat{y})^2) = 2 \cdot (y^2 + \widehat{y}^2)$, meaning we can take $C = 2$.

For the mean absolute error, write $\ell(y,\widehat{y}) = |y-\widehat{y}|$, with $\psi(x) = |x|$. Then, by the standard triangular inequality, $|y \pm \widehat{y}| \leq |y| + |\pm\widehat{y}| = |y| + |\widehat{y}|$, meaning we can take $C = 1$.

For the α-quantile loss, write $\ell(y,\widehat{y}) = \text{pinball}_\alpha(y-\widehat{y})$, with $\psi(x) = \text{pinball}_\alpha(x) = x\alpha\mathbb{1}[x \geq 0] - x(1-\alpha)\mathbb{1}[x < 0]$. To prove the triangular-type inequality, we consider two cases and use Lemma A.4:

1. When $y - \widehat{y} \geq 0$, it holds that

$$\text{pinball}_\alpha(y-\widehat{y}) = (y-\widehat{y})\alpha = y\alpha + \frac{\alpha}{1-\alpha}[-\widehat{y}(1-\alpha)] \leq \text{pinball}_\alpha(y) + \frac{\alpha}{1-\alpha}\text{pinball}_\alpha(\widehat{y}).$$
$$\text{pinball}_\alpha(y+\widehat{y}) = (y+\widehat{y})\alpha = y\alpha + \widehat{y}\alpha \leq \text{pinball}_\alpha(y) + \text{pinball}_\alpha(\widehat{y}).$$

2. When $y - \widehat{y} < 0$, it holds that

$$\text{pinball}_\alpha(y-\widehat{y}) = -(y-\widehat{y})(1-\alpha) = -y(1-\alpha) + \frac{1-\alpha}{\alpha}[\widehat{y}\alpha] \leq \text{pinball}_\alpha(y) + \frac{1-\alpha}{\alpha}\text{pinball}_\alpha(\widehat{y}).$$
$$\text{pinball}_\alpha(y+\widehat{y}) = -(y+\widehat{y})(1-\alpha) = -y(1-\alpha) - \widehat{y}(1-\alpha) \leq \text{pinball}_\alpha(y) + \text{pinball}_\alpha(\widehat{y}).$$

Joining both, we conclude that

$$\begin{aligned}
\text{pinball}_\alpha(y \pm \widehat{y}) &\leq \text{pinball}_\alpha(y) + \max\left\{\frac{\alpha}{1-\alpha}, \frac{1-\alpha}{\alpha}, 1\right\}\text{pinball}_\alpha(\widehat{y}) \\
&\leq \max\left\{\frac{\alpha}{1-\alpha}, \frac{1-\alpha}{\alpha}, 1\right\}(\text{pinball}_\alpha(y) + \text{pinball}_\alpha(\widehat{y})) \\
&= \max\left\{\frac{\alpha}{1-\alpha}, \frac{1-\alpha}{\alpha}\right\}(\text{pinball}_\alpha(y) + \text{pinball}_\alpha(\widehat{y})) \\
&= \frac{\max\{\alpha, 1-\alpha\}}{\min\{\alpha, 1-\alpha\}}(\text{pinball}_\alpha(y) + \text{pinball}_\alpha(\widehat{y})).
\end{aligned}$$

Finally, for the 0-1 loss (in the binary case), consider $y, \widehat{y} \in \{0,1\}$ and write $\ell(y,\widehat{y}) = |y-\widehat{y}|$. Then the same logic as in the mean absolute error holds. □

Proposition A.6 (Proposition 2.7 of the main body)**.** *Let ℓ be a loss function satisfying Assumption 2.6 with constant C, and let $w^1(X), w^0(X)$ be nonnegative reweighing functions. For any $\lambda_1, \lambda_0 > 0$,*

$$\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))] \leq C\Big(\mathbb{E}[w^1(X)\ell(Y, h^1(X))|T=1] + \mathbb{E}[w^0(X)\ell(Y, h^0(X))|T=0] + \underbrace{\lambda_1\Delta_{T=1} + \lambda_0\Delta_{T=0}}_{\Delta_{\mathrm{T-learner}}} + \underbrace{\sigma^2_{T=1}/4\lambda_1 + \sigma^2_{T=0}/4\lambda_0}_{\sigma^2/4}\Big),$$

Where $\sigma^2_{T=a} := \mathrm{Var}[\ell(Y^a, h^a(X))]$ and

$$\Delta_{T=a} = \mathbb{E}\left[\left(w^a(X)\frac{\mathbb{P}[T=a|X,U]}{\mathbb{P}[T=a]} - 1\right)^2\right].$$
$$\leq \frac{2}{\mathbb{P}[T=a]^2}\mathbb{E}\left[w^a(X)^2\left(\nu(X) - \mathbb{1}[T=a]\right)^2\right] \quad + 2\mathbb{E}\left[\left(w^a(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2\right].$$

Proof. Using Assumption 2.6,

$$\begin{aligned}
&\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))] \\
&\quad = \mathbb{E}[\psi\left((Y^1 - Y^0) - (h^1(X) - h^0(X))\right)] \\
&\quad = \mathbb{E}[\psi\left((Y^1 - h^1(X)) - (Y^0 - h^0(X))\right)] \\
&\quad \leq \mathbb{E}[\psi(Y^1 - h^1(X)) + \psi(Y^0 - h^0(X))] = \mathbb{E}[\psi(Y^1 - h^1(X))] + \mathbb{E}[\psi(Y^0 - h^0(X))].
\end{aligned}$$

Then, the rest follows by applying Theorem 2.3 and 2.4. □

Proposition A.7 (Proposition 2.9 of the main body)**.** *Let ℓ be a loss function satisfying Assumption 2.6 with constant C and let w^1, w^0 be nonnegative reweighting functions. For any $\lambda_1, \lambda_0, \lambda_{0,1}, \lambda_{1,0} > 0$,*

$$\begin{aligned}
\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + \bar{e}(X)\tau^0(X))] \leq C^2\Big(&\mathbb{E}[w^1(X)\ell_{\bar{e}}(Y, h^1(X))|T=1] + \mathbb{E}[w^0(X)\ell_e(Y, h^0(X))|T=0] \\
&+ \mathbb{E}[w^1(X)\ell_e(Y^1 - h^0(X), \tau^1(X))|T=1] + \mathbb{E}[w^0(X)\ell_{\bar{e}}(h^1(X) - Y^0, \tau^0(X))|T=0] \\
&+ \underbrace{(\lambda_1 + \lambda_{1,0})\Delta_{T=1} + (\lambda_0 + \lambda_{0,1})\Delta_{T=0}}_{\Delta_{\mathrm{X-learner}}} + \underbrace{\sigma^2_{T=1}/4\lambda_1 + \sigma^2_{T=0}/4\lambda_0 + \sigma^2_{1,0}/4\lambda_{1,0} + \sigma^2_{0,1}/4\lambda_{0,1}}_{\sigma^2/4}\Big),
\end{aligned}$$

where $\sigma^2_{T=a} := \mathrm{Var}[\ell(Y^a, h^a(X))]$, $\sigma^2_{a,b} := \mathrm{Var}[\ell(Y^a - h^b(X), \tau^a(X))]$ and

$$\Delta_{T=a} = \mathbb{E}\left[\left(w^a(X)\frac{\mathbb{P}[T=a|X,U]}{\mathbb{P}[T=a]} - 1\right)^2\right]$$
$$\leq \frac{2}{\mathbb{P}[T=a]^2}\mathbb{E}\left[w^a(X)^2\left(\nu(X) - \mathbb{1}[T=a]\right)^2\right] + 2\mathbb{E}\left[\left(w^a(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2\right].$$

Proof. Using Assumption 2.6,

$$\begin{aligned}
&\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + (1-e(X))\tau^0(X))] \\
&= \mathbb{E}[\psi\left((Y^1 - Y^0) - (e(X)\tau^1(X) + (1-e(X))\tau^0(X))\right)] \\
&= \mathbb{E}[\psi\big(Y^1 - Y^0 - e(X)(\tau^1(X) - (Y^1 - h^0(X)) + (Y^1 - h^0(X))) \\
&\qquad - (1-e(X))(\tau^0(X) - (h^1(X) - Y^0) + (h^1(X) - Y^0))\big)] \\
&= \mathbb{E}[\psi\big(Y^1 - Y^0 - e(X)Y^1 + e(X)h^0(X) - (1-e(X))h^1(X) + (1-e(X))Y^0 \\
&\qquad - e(X)(\tau^1(X) - (Y^1 - h^0(X))) - (1-e(X))(\tau^0(X) - (h^1(X) - Y^0))\big)] \\
&= \mathbb{E}[\psi\big((1-e(X))(Y^1 - h^1(X)) - e(X)(Y^0 - h^0(X)) \\
&\qquad - e(X)(\tau^1(X) - (Y^1 - h^0(X))) - (1-e(X))(\tau^0(X) - (h^1(X) - Y^0))\big)] \\
&\leq C(\mathbb{E}[\psi\left((1-e(X))(Y^1 - h^1(X)) - e(X)(Y^0 - h^0(X))\right)] \\
&\qquad + \mathbb{E}[\psi\left(e(X)(\tau^1(X) - (Y^1 - h^0(X))) + (1-e(X))(\tau^0(X) - (h^1(X) - Y^0))\right)]) \\
&\leq C^2(\mathbb{E}[\psi\left((1-e(X))(Y^1 - h^1(X))\right)] + \mathbb{E}[\psi\left(e(X)(Y^0 - h^0(X))\right)] \\
&\qquad + \mathbb{E}[\psi\left(e(X)(\tau^1(X) - (Y^1 - h^0(X)))\right)] + \mathbb{E}[\psi\left((1-e(X))(\tau^0(X) - (h^1(X) - Y^0))\right)]) \\
&= C^2(\mathbb{E}[\ell_{\bar{e}}(Y^1, h^1(X))] + \mathbb{E}[\ell_e(Y^0, h^0(X))] + \mathbb{E}[\ell_e(Y^1 - h^0(X), \tau^1(X))] + \mathbb{E}[\ell_{\bar{e}}(h^1(X) - Y^0, \tau^0(X))])
\end{aligned}$$

Then, the rest follows by applying Theorems 2.3 and 2.4. □

A.3. Finite-sample Bounds

Throughout this section, we rely on bounds on the random variables within the expectations that appear on the right-hand-side of the results in Section A.2.

Lemma A.8. *Suppose that ℓ is a loss function bounded in $[0, M]$, $w(X)$ is a nonnegative reweighing function bounded in $[0, w_{\max}]$, and that $\nu(X)$ is bounded in $[0, 1]$. It then holds that, for all x, y:*

$$0 \leq w(x)\ell(y, h(x)) \leq w_{\max} M$$

$$0 \leq \left(\frac{w(x)}{\mathbb{P}[T=a]}\right)^2 (\nu(x) - \mathbb{1}[T=a])^2 + \left(w(x)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2 \leq \left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2 + \max\left\{1, \left(\frac{w_{\max}}{\mathbb{P}[T=a]} - 1\right)^2\right\}.$$

Proof. For the first bound, simply note that both $w(x)$ and $\ell(y, h(x))$ are nonnegative, and that

$$w(x)\ell(y, h(x)) \leq w_{\max}\ell(y, h(x)) \leq w_{\max} M.$$

The second bound is slightly more involved.

$$0 \leq \frac{w(x)}{\mathbb{P}[T=a]} \leq \frac{w_{\max}}{\mathbb{P}[T=a]} \implies \left(\frac{w(x)}{\mathbb{P}[T=a]}\right)^2 \leq \left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2$$

and

$$-1 \leq \nu(x) - \mathbb{1}[T=a] \leq 1 \implies 0 \leq (\nu(x) - \mathbb{1}[T=a])^2 \leq 1.$$

Together, we get that

$$0 \leq \left(\frac{w(x)}{\mathbb{P}[T=a]}\right)^2 (\nu(x) - \mathbb{1}[T=a])^2 \leq \left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2.$$

Next,

$$-1 \leq w(x)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1 \leq \frac{w_{\max}}{\mathbb{P}[T=a]} - 1 \implies 0 \leq \left(w(x)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2 \leq \max\left\{1, \left(\frac{w_{\max}}{\mathbb{P}[T=a]} - 1\right)^2\right\}.$$

Combining everything, we get that

$$0 \leq \left(\frac{w(x)}{\mathbb{P}[T=a]}\right)^2 (\nu(x) - \mathbb{1}[T=a])^2 + \left(w(x)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]} - 1\right)^2 \leq \left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2 + \max\left\{1, \left(\frac{w_{\max}}{\mathbb{P}[T=a]} - 1\right)^2\right\}.$$

□

We base ourselves on a variation of the standard PAC Rademacher Complexity bound presented in (Mohri et al., 2012):

Lemma A.9 (Theorem 3.3 in (Mohri et al., 2012))**.** *Let $\mathcal{G}$ be family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample $(Z_i)_{i=1}^n$, the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(Z)] \leq \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2\mathfrak{R}(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2n}}.$$

Lemma A.10. *Let $\mathcal{G}$ be family of functions mapping from $\mathcal{Z}$ to $[0,M]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample $(Z_i)_{i=1}^n$, the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(Z)] \leq \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2\mathfrak{R}(\mathcal{G}) + M\sqrt{\frac{\log 1/\delta}{2n}}.$$

Proof. Writing $g(z) = M^{-1}\widetilde{g}(z)$, we have by the definition of the Rademacher complexity and Lemma A.9 that

$$\begin{aligned}
\mathfrak{R}(\mathcal{G}) &= \mathbb{E}\left[\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_\sigma\left[\sup_{\widetilde{g}\in\widetilde{\mathcal{G}}} \frac{1}{n}\sum_{i=1}^n \sigma_i M^{-1}\widetilde{g}(Z_i)\right]\right] \\
&= M^{-1}\mathbb{E}\left[\mathbb{E}_\sigma\left[\sup_{\widetilde{g}\in\widetilde{\mathcal{G}}} \frac{1}{n}\sum_{i=1}^n \sigma_i \widetilde{g}(Z_i)\right]\right] = M^{-1}\mathfrak{R}(\widetilde{\mathcal{G}})
\end{aligned}$$

and so

$$\begin{aligned}
M^{-1}\mathbb{E}[\widetilde{g}(Z)] &\leq M^{-1}\frac{1}{n}\sum_{i=1}^n \widetilde{g}(Z_i) + 2M^{-1}\mathfrak{R}(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2n}}. \\
\implies \mathbb{E}[\widetilde{g}(Z)] &\leq \frac{1}{n}\sum_{i=1}^n \widetilde{g}(Z_i) + 2\mathfrak{R}(\mathcal{G}) + M\sqrt{\frac{\log 1/\delta}{2n}}.
\end{aligned}$$

□

Corollary A.11 (Corollary 2.5 in the main body)**.** *Suppose that ℓ is a loss function bounded in $[0,M]$ and $w(X)$ is a nonnegative reweighing function bounded in $[0, w_{\max}]$. Then, for any $\lambda > 0$, with probability at least $1-\delta$ over the draw of the training data $(X_i, T_i, Y_i)_{i=1}^n$, for all $h \in \mathcal{H}$ and $\nu \in \mathcal{H}_\nu$,*

$$\begin{aligned}
\mathbb{E}[\ell(Y^a, h(X))] \leq \frac{1}{n_{T=a}}\sum_{T_i=a} w(X_i)\ell(Y_i, h(X_i)) + \lambda\widehat{\Delta}_{T=a} + \frac{M^2}{16\lambda} \\
+ 2\mathfrak{R}(w\cdot\ell\circ\mathcal{H}) + 2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu) + \left(Mw_{\max} + C(w_{\max})\sqrt{n_{T=a}/n}\right)\sqrt{\frac{\log 2/\delta}{2n_{T=a}}}
\end{aligned}$$

where

$$\widehat{\Delta}_{T=a} := \frac{2\lambda}{n}\sum_{i=1}^n \left(\frac{w(X_i)}{\mathbb{P}[T=a]}\right)^2 (\nu(X_i) - \mathbb{1}[T_i=a])^2 + \frac{2\lambda}{n}\sum_{i=1}^n \left(w(X_i)\frac{\mathbb{1}[T_i=a]}{\mathbb{P}[T=a]} - 1\right)^2,$$

$\mathfrak{R}(w\cdot\ell\circ\mathcal{H})$ and $\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu)$ are the Rademacher complexities of $\mathcal{H}$ and $\mathcal{H}_\nu$ composed with their respective loss functions/means and $C(w_{\max})$ is a constant nonnegative quantitity defined in the proof.

Proof. By Theorems 2.3 and 2.4,

$$\mathbb{E}[\ell(Y^a, h(X))] \leq \mathbb{E}[w(X_i)\ell(Y_i, h(X_i))|T=a] + 2\lambda\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2\right] + 2\lambda\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right] + \frac{\mathrm{Var}[\ell(Y^a, h(X))]}{4\lambda}.$$

By Popoviciu's inequality, $\sigma^2 \leq M^2/4$ and so

$$\mathbb{E}[\ell(Y^a, h(X))] \leq \mathbb{E}[w(X_i)\ell(Y_i, h(X_i))|T=a] + 2\lambda\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2\right] + 2\lambda\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right] + \frac{M^2}{16\lambda}.$$

Note that for all x, y, it holds that $w(X)\ell(Y, h(X)) \leq w_{\max}M$. By Lemma A.10 along with Lemma A.8, with probability of at least $1-\delta/2$,

$$\mathbb{E}[w(X)\ell(Y, h(X))|T=a] \leq \frac{1}{n_{T=a}}\sum_{T_i=a} w(X_i)\ell(Y_i, h(X_i)) + 2\mathfrak{R}(w\cdot\ell\circ\mathcal{H}) + Mw_{\max}\sqrt{\frac{\log 2/\delta}{2n_{T=a}}}.$$

And, also with probability of at least $1-\delta/2$,

$$\begin{aligned}
&2\lambda\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2\right] + 2\lambda\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right] \\
&= 2\lambda\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2 + \left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right] \\
&\leq \frac{2\lambda}{n}\sum_{i=1}^n\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2 + \frac{2\lambda}{n}\sum_{i=1}^n\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2 \\
&\qquad + 2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu) + \left(\left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2 + \max\left\{1, \left(\frac{w_{\max}}{\mathbb{P}[T=a]}-1\right)^2\right\}\right)\sqrt{\frac{\log 2/\delta}{2n}}.
\end{aligned}$$

Therefore, by an union bound, with probability of at least $1-\delta$,

$$\begin{aligned}
\mathbb{E}[\ell(Y^a, h(X))] &\leq \frac{1}{n_{T=a}}\sum_{T_i=a} w(X_i)\ell(Y_i, h(X_i)) + \frac{2\lambda}{n}\sum_{i=1}^n\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2 \\
&+ \frac{2\lambda}{n}\sum_{i=1}^n\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2 + \frac{M^2}{16\lambda} + 2\mathfrak{R}(w\cdot\ell\circ\mathcal{H}) + 2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu) \\
&+ Mw_{\max}\sqrt{\frac{\log 2/\delta}{2n_{T=a}}} + \underbrace{\left(\left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2 + \max\left\{1, \left(\frac{w_{\max}}{\mathbb{P}[T=a]}-1\right)^2\right\}\right)}_{C(w_{\max})}\sqrt{\frac{\log 2/\delta}{2n}}.
\end{aligned}$$

And we conclude by rearranging and observing that

$$\begin{aligned}
&Mw_{\max}\sqrt{\frac{\log 2/\delta}{2n_{T=a}}} + C(w_{\max})\sqrt{\frac{\log 2/\delta}{2n}} = Mw_{\max}\sqrt{\frac{\log 2/\delta}{2n_{T=a}}} + C(w_{\max})\sqrt{\frac{n_{T=a}}{n}}\sqrt{\frac{\log 2/\delta}{2n_{T=a}}} \\
&= \left(w_{\max}M + C(w_{\max})\sqrt{\frac{n_{T=a}}{n}}\right)\sqrt{\frac{\log 2/\delta}{2n_{T=a}}}.
\end{aligned}$$

□

Corollary A.12 (Corollary 2.7 in the main body)**.** *Let ℓ be a loss function bounded in $[0, M]$ satisfying Assumption 2.6 and let $w^1(X), w^0(X)$ be nonnegative reweighing functions bounded in $[0, w_{\max}]$. Then, for any $\lambda_1, \lambda_2 > 0$, with probability at least $1-\delta$ over the draw of the training data $(X_i, T_i, Y_i)_{i=1}^n$, for all $h^1, h^0 \in \mathcal{H}$ and $\nu \in \mathcal{H}_\nu$,*

$$\begin{aligned}\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))] \leq \frac{1}{n_{T=1}} \sum_{T_i=1} w(X_i)\ell(Y_i, h^1(X_i)) + \frac{1}{n_{T=0}} \sum_{T_i=0} w(X_i)\ell(Y_i, h^0(X_i)) \\ + \lambda_1 \widehat{\Delta}_{T=1} + \lambda_0 \widehat{\Delta}_{T=0} + \frac{M^2}{16\lambda_1} + \frac{M^2}{16\lambda_0} \\ + 2\mathfrak{R}(w^1 \cdot \ell \circ \mathcal{H}) + 2\mathfrak{R}(w^0 \cdot \ell \circ \mathcal{H}) + 2\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu) \\ + \left(cMw_{\max} + C(w_{\max})\sqrt{n_{T=\min}/n}\right)\sqrt{\frac{\log 3/\delta}{2n_{T=\min}}}\end{aligned}$$

where

$$\widehat{\Delta}_{T=a} := \frac{2\lambda}{n} \sum_{i=1}^n \left(\frac{w(X_i)}{\mathbb{P}[T=a]}\right)^2 (\nu(X_i) - \mathbb{1}[T_i = a])^2 + \frac{2\lambda}{n} \sum_{i=1}^n \left(w(X_i)\frac{\mathbb{1}[T_i = a]}{\mathbb{P}[T=a]} - 1\right)^2,$$

$n_{T=\min} = \min\{n_{T=1}, n_{T=0}\}$, $\mathfrak{R}(w^a \cdot \ell \circ \mathcal{H})$ and $\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu)$ are the Rademacher complexities of $\mathcal{H}$ and $\mathcal{H}_\nu$ composed with their respective loss functions/means, and c and $C(w_{\max})$ are constant nonnegative quantities defined in the proof, with $1 \leq c \leq 2$.

Proof. By Propositions 2.7,

$$\begin{aligned}\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))] \leq C\Big(&\mathbb{E}[w^1(X)\ell(Y, h^1(X))|T=1] + \mathbb{E}[w^0(X)\ell(Y, h^0(X))|T=0] \\ &+ \frac{2\lambda_1}{\mathbb{P}[T=1]^2}\mathbb{E}\left[w^1(X)^2\left(\nu(X) - \mathbb{1}[T=1]\right)^2\right] + 2\lambda_1\mathbb{E}\left[\left(w^1(X)\frac{\mathbb{1}[T=1]}{\mathbb{P}[T=1]} - 1\right)^2\right] \\ &+ \frac{2\lambda_0}{\mathbb{P}[T=0]^2}\mathbb{E}\left[w^0(X)^2\left(\nu(X) - \mathbb{1}[T=0]\right)^2\right] + 2\lambda_0\mathbb{E}\left[\left(w^0(X)\frac{\mathbb{1}[T=0]}{\mathbb{P}[T=0]} - 1\right)^2\right] \\ &+ \frac{1}{4\lambda_1}\operatorname{Var}[\ell(Y^1, h^1(X))] + \frac{1}{4\lambda_0}\operatorname{Var}[\ell(Y^0, h^0(X))]\Big).\end{aligned}$$

By Popoviciu's inequality, $\sigma^2 \leq M^2/4$ and so

$$\begin{aligned}\mathbb{E}[\ell(Y^1 - Y^0, h^1(X) - h^0(X))] \leq C\Big(&\mathbb{E}[w^1(X)\ell(Y, h^1(X))|T=1] + \mathbb{E}[w^0(X)\ell(Y, h^0(X))|T=0] \\ &+ \frac{2\lambda_1}{\mathbb{P}[T=1]^2}\mathbb{E}\left[w^1(X)^2\left(\nu(X) - \mathbb{1}[T=1]\right)^2\right] + 2\lambda_1\mathbb{E}\left[\left(w^1(X)\frac{\mathbb{1}[T=1]}{\mathbb{P}[T=1]} - 1\right)^2\right] \\ &+ \frac{2\lambda_0}{\mathbb{P}[T=0]^2}\mathbb{E}\left[w^0(X)^2\left(\nu(X) - \mathbb{1}[T=0]\right)^2\right] + 2\lambda_0\mathbb{E}\left[\left(w^0(X)\frac{\mathbb{1}[T=0]}{\mathbb{P}[T=0]} - 1\right)^2\right] \\ &+ \frac{M^2}{16\lambda_1} + \frac{M^2}{16\lambda_0}\Big).\end{aligned}$$

By Lemma A.10 along with Lemma A.8, with probability of at least $1 - \delta/3$,

$$\mathbb{E}[w^a(X)\ell(Y, h^a(X))|T=a] \leq \frac{1}{n_{T=a}} \sum_{T_i=a} w^a(X_i)\ell(Y_i, h^a(X_i)) + 2\mathfrak{R}(w^a \cdot \ell \circ \mathcal{H}) + Mw_{\max}\sqrt{\frac{\log 3/\delta}{2n_{T=a}}}.$$

And, also with probability of at least $1-\delta/3$,

$$\sum_{a\in\{0,1\}}\left(2\lambda\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2\right]+2\lambda\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right]\right)$$
$$=2\lambda\mathbb{E}\left[\sum_{a\in\{0,1\}}\left(\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2+\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right)\right]$$
$$\leq\sum_{a\in\{0,1\}}\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2+\frac{2\lambda}{n}\sum_{i=1}^{n}\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right)$$
$$+2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu)+\left(\sum_{a\in\{0,1\}}\left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2+\sum_{a\in\{0,1\}}\max\left\{1,\left(\frac{w_{\max}}{\mathbb{P}[T=a]}-1\right)^2\right\}\right)\sqrt{\frac{\log 3/\delta}{2n}}.$$

Therefore, by an union bound, with probability of at least $1-\delta$,

$$\mathbb{E}[\ell(Y^a,h^1(X)-h^0(X))]\leq\frac{1}{n_{T=1}}\sum_{T_i=1}w^1(X_i)\ell(Y_i,h^1(X_i))+\frac{1}{n_{T=0}}\sum_{T_i=0}w^0(X_i)\ell(Y_i,h^0(X_i))$$
$$+\frac{2\lambda}{n}\sum_{i=1}^{n}\left(\frac{w^1(X)}{\mathbb{P}[T=1]}\right)^2(\nu(X)-\mathbb{1}[T=1])^2+\frac{2\lambda}{n}\sum_{i=1}^{n}\left(w^1(X)\frac{\mathbb{1}[T=1]}{\mathbb{P}[T=1]}-1\right)^2$$
$$+\frac{2\lambda}{n}\sum_{i=1}^{n}\left(\frac{w^0(X)}{\mathbb{P}[T=0]}\right)^2(\nu(X)-\mathbb{1}[T=0])^2+\frac{2\lambda}{n}\sum_{i=1}^{n}\left(w^0(X)\frac{\mathbb{1}[T=0]}{\mathbb{P}[T=0]}-1\right)^2$$
$$+\frac{M^2}{16\lambda_1}+\frac{M^2}{16\lambda_2}+2\mathfrak{R}(w^1\cdot\ell\circ\mathcal{H})+2\mathfrak{R}(w^0\cdot\ell\circ\mathcal{H})+2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu)$$
$$+Mw_{\max}\sqrt{\frac{\log 3/\delta}{2n_{T=1}}}+Mw_{\max}\sqrt{\frac{\log 3/\delta}{2n_{T=0}}}$$
$$+\underbrace{\left(\sum_{a\in\{0,1\}}\left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2+\sum_{a\in\{0,1\}}\max\left\{1,\left(\frac{w_{\max}}{\mathbb{P}[T=a]}-1\right)^2\right\}\right)}_{C(w_{\max})}\sqrt{\frac{\log 3/\delta}{2n}}.$$

And we conclude by rearranging and observing that, assuming without loss of generality that $n_{T=1}=\min\{n_{T=1},n_{T=0}\}$,

$$Mw_{\max}\sqrt{\frac{\log 3/\delta}{2n_{T=1}}}+Mw_{\max}\sqrt{\frac{\log 3/\delta}{2n_{T=0}}}+C(w_{\max})\sqrt{\frac{\log 3/\delta}{2n}}$$
$$=Mw_{\max}\sqrt{\frac{\log 3/\delta}{2n_{T=\min}}}+Mw_{\max}\sqrt{\frac{n_{T=\min}}{n_{T=0}}}\sqrt{\frac{\log 3/\delta}{2n_{T=\min}}}+C(w_{\max})\sqrt{\frac{n_{T=\min}}{n}}\sqrt{\frac{\log 3/\delta}{2n_{T=\min}}}$$
$$=\left(Mw_{\max}\underbrace{\left(1+\sqrt{\frac{n_{T=\min}}{n_{T=0}}}\right)}_{c}+C(w_{\max})\sqrt{\frac{n_{T=\min}}{n}}\right)\sqrt{\frac{\log 3/\delta}{2n_{T=\min}}}$$

and that $0\leq\sqrt{n_{T=\min}/n_{T=0}}\leq 1$. □

Corollary A.13. *Let ℓ be a loss function bounded in $[0,M]$ satisfying Assumption 2.6 and let $w^1(X),w^0(X)$ be nonnegative reweighing functions bounded in $[0,w_{\max}]$. Then, for any $\lambda_1,\lambda_0>0$, with probability at least $1-\delta$ over the draw of the*

training data $(X_i, T_i, Y_i)_{i=1}^n$*, for all* $h^1, h^0, \tau^1, \tau^0 \in \mathcal{H}$ *and* $\nu \in \mathcal{H}_\nu$*,*

$$\begin{aligned}\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + \bar{e}(X)\tau^0(X))] \leq C^2 \Bigg(& \frac{1}{n_{T=1}} \sum_{T_i=1} w(X_i)\ell_{\bar{e}}(Y_i, h^1(X_i)) + \frac{1}{n_{T=0}} \sum_{T_i=0} w(X_i)\ell_e(Y_i, h^0(X_i)) \\ & + \frac{1}{n_{T=1}} \sum_{T_i=1} w(X_i)\ell_e(Y_i - h^0(X_i), \tau^1(X_i)) + \frac{1}{n_{T=0}} \sum_{T_i=0} w(X_i)\ell_{\bar{e}}(h^1(X_i) - Y_i, \tau^0(X_i)) \\ & + \lambda_1 \widehat{\Delta}_{T=1} + \lambda_0 \widehat{\Delta}_{T=0} + \frac{M^2}{4\lambda_1} + \frac{M^2}{4\lambda_0} + 4\mathfrak{R}(w^1 \cdot \ell \circ \mathcal{H}) + 4\mathfrak{R}(w^0 \cdot \ell \circ \mathcal{H}) + 2\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu) \\ & + \left(2cMw_{\max} + C(w_{\max})\sqrt{n_{T=\min}/n}\right) \sqrt{\frac{\log 5/\delta}{2n_{T=\min}}} \Bigg)\end{aligned}$$

where

$$\widehat{\Delta}_{T=a} := \frac{2\lambda}{n} \sum_{i=1}^n \left(\frac{w(X_i)}{\mathbb{P}[T=a]} \right)^2 (\nu(X_i) - \mathbb{1}[T_i = a])^2 + \frac{2\lambda}{n} \sum_{i=1}^n \left(w(X_i) \frac{\mathbb{1}[T_i = a]}{\mathbb{P}[T=a]} - 1 \right)^2,$$

$n_{T=\min} = \min\{n_{T=1}, n_{T=0}\}$*,* $\mathfrak{R}(w^a \cdot \ell \circ \mathcal{H})$ *and* $\mathfrak{R}(\widehat{\Delta} \circ \mathcal{H}_\nu)$ *are the Rademacher complexities of* $\mathcal{H}$ *and* $\mathcal{H}_\nu$ *composed with their respective loss functions/means, and* c *and* $C(w_{\max})$ *are constant nonnegative quantities defined in the proof, with* $1 \leq c \leq 2$*.*

Proof. By Proposition 2.9,

$$\begin{aligned}\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + \bar{e}(X)\tau^0(X))] \leq C^2 \Big(& \mathbb{E}[w^1(X)\ell_{\bar{e}}(Y, h^1(X))|T=1] + \mathbb{E}[w^0(X)\ell_e(Y, h^0(X))|T=0] \\ & + \mathbb{E}[w^1(X)\ell_e(Y^1 - h^0(X), \tau^1(X))|T=1] + \mathbb{E}[w^0(X)\ell_{\bar{e}}(h^1(X) - Y^0, \tau^0(X))|T=0] \\ & + \left(\frac{\lambda_1}{2} + \frac{\lambda_1}{2}\right)\Delta_{T=1} + \left(\frac{\lambda_0}{2} + \frac{\lambda_0}{2}\right)\Delta_{T=0} + \frac{\sigma^2_{T=1}}{4\lambda_1/2} + \frac{\sigma^2_{T=0}}{4\lambda_0/2} + \frac{\sigma^2_{1,0}}{4\lambda_1/2} + \frac{\sigma^2_{0,1}}{4\lambda_0/2} \Big),\end{aligned}$$

By Popoviciu's inequality, $\sigma^2 \leq M^2/4$ and so

$$\begin{aligned}\mathbb{E}[\ell(Y^1 - Y^0, e(X)\tau^1(X) + \bar{e}(X)\tau^0(X))] \leq C^2 \Big(& \mathbb{E}[w^1(X)\ell_{\bar{e}}(Y, h^1(X))|T=1] + \mathbb{E}[w^0(X)\ell_e(Y, h^0(X))|T=0] \\ & + \mathbb{E}[w^1(X)\ell_e(Y^1 - h^0(X), \tau^1(X))|T=1] + \mathbb{E}[w^0(X)\ell_{\bar{e}}(h^1(X) - Y^0, \tau^0(X))|T=0] \\ & + \lambda_1\Delta_{T=1} + \lambda_0\Delta_{T=0} + \frac{M^2}{4\lambda_1} + \frac{M^2}{4\lambda_0} \Big),\end{aligned}$$

By Lemma A.10 along with Lemma A.8, with probability of at least $1 - \delta/5$,

$$\mathbb{E}[w^1(X)\ell_{\bar{e}}(Y, h^1(X))|T=1] \leq \frac{1}{n_{T=1}} \sum_{T_i=1} w^1(X_i)\ell_{\bar{e}}(Y_i, h^1(X_i)) + 2\mathfrak{R}(w^1 \cdot \ell \circ \mathcal{H}) + Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=1}}}.$$

$$\mathbb{E}[w^0(X)\ell_e(Y, h^0(X))|T=0] \leq \frac{1}{n_{T=0}} \sum_{T_i=0} w^0(X_i)\ell_e(Y_i, h^0(X_i)) + 2\mathfrak{R}(w^0 \cdot \ell \circ \mathcal{H}) + Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=0}}}.$$

Also with probability of at least $1-\delta/5$,

$$\mathbb{E}[w^1(X)\ell_e(Y^1-h^0(X),\tau^1(X))|T=1] \leq \frac{1}{n_{T=1}}\sum_{T_i=1} w^1(X_i)\ell_e(Y^1-h^0(X),\tau^1(X)) + 2\mathfrak{R}(w^1\cdot\ell\circ\mathcal{H}) + Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=1}}}.$$

$$\mathbb{E}[w^0(X)\ell_{\bar{e}}(h^1(X)-Y^0,\tau^0(X))|T=0] \leq \frac{1}{n_{T=0}}\sum_{T_i=0} w^0(X_i)\ell_{\bar{e}}(h^1(X)-Y^0,\tau^0(X)) + 2\mathfrak{R}(w^0\cdot\ell\circ\mathcal{H}) + Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=0}}}.$$

And, likewise,

$$\begin{aligned}
&\sum_{a\in\{0,1\}}\left(2\lambda\mathbb{E}\left[\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2\right]+2\lambda\mathbb{E}\left[\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right]\right)\\
&=2\lambda\mathbb{E}\left[\sum_{a\in\{0,1\}}\left(\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2+\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right)\right]\\
&\leq\sum_{a\in\{0,1\}}\left(\frac{2\lambda}{n}\sum_{i=1}^n\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2+\frac{2\lambda}{n}\sum_{i=1}^n\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right)\\
&\qquad+2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu)+\left(\sum_{a\in\{0,1\}}\left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2+\sum_{a\in\{0,1\}}\max\left\{1,\left(\frac{w_{\max}}{\mathbb{P}[T=a]}-1\right)^2\right\}\right)\sqrt{\frac{\log 5/\delta}{2n}}.
\end{aligned}$$

Therefore, by an union bound, with probability of at least $1-\delta$,

$$\begin{aligned}
\mathbb{E}[\ell(Y^a,e(X)\tau^1(X)+\bar{e}(X)\tau^0(X))] &\leq \frac{1}{n_{T=1}}\sum_{T_i=1} w^1(X_i)\ell_{\bar{e}}(Y_i,h^1(X_i)) + \frac{1}{n_{T=0}}\sum_{T_i=0} w^0(X_i)\ell_e(Y_i,h^0(X_i))\\
&+\frac{1}{n_{T=1}}\sum_{T_i=1} w^1(X_i)\ell_e(Y^1-h^0(X),\tau^1(X)) + \frac{1}{n_{T=0}}\sum_{T_i=0} w^0(X_i)\ell_{\bar{e}}(h^1(X)-Y^0,\tau^0(X))\\
&+\sum_{a\in\{0,1\}}\left(\frac{2\lambda}{n}\sum_{i=1}^n\left(\frac{w(X)}{\mathbb{P}[T=a]}\right)^2(\nu(X)-\mathbb{1}[T=a])^2+\frac{2\lambda}{n}\sum_{i=1}^n\left(w(X)\frac{\mathbb{1}[T=a]}{\mathbb{P}[T=a]}-1\right)^2\right)\\
&+4\mathfrak{R}(w^1\cdot\ell\circ\mathcal{H})+4\mathfrak{R}(w^0\cdot\ell\circ\mathcal{H})+2\mathfrak{R}(\widehat{\Delta}\circ\mathcal{H}_\nu)\\
&+2Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=1}}}+2Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=0}}}\\
&+\underbrace{\left(\sum_{a\in\{0,1\}}\left(\frac{w_{\max}}{\mathbb{P}[T=a]}\right)^2+\sum_{a\in\{0,1\}}\max\left\{1,\left(\frac{w_{\max}}{\mathbb{P}[T=a]}-1\right)^2\right\}\right)}_{C(w_{\max})}\sqrt{\frac{\log 5/\delta}{2n}}.
\end{aligned}$$

And we conclude by rearranging and observing that, assuming without loss of generality that $n_{T=1} = \min\{n_{T=1}, n_{T=0}\}$,

$$\begin{aligned}
&2Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=1}}} + 2Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=0}}} + C(w_{\max})\sqrt{\frac{\log 5/\delta}{2n}} \\
&= 2Mw_{\max}\sqrt{\frac{\log 5/\delta}{2n_{T=\min}}} + 2Mw_{\max}\sqrt{\frac{n_{T=\min}}{n_{T=0}}}\sqrt{\frac{\log 5/\delta}{2n_{T=\min}}} + C(w_{\max})\sqrt{\frac{n_{T=\min}}{n}}\sqrt{\frac{\log 5/\delta}{2n_{T=\min}}} \\
&= \left(2Mw_{\max}\underbrace{\left(1+\sqrt{\frac{n_{T=\min}}{n_{T=0}}}\right)}_{c} + C(w_{\max})\sqrt{\frac{n_{T=\min}}{n}}\right)\sqrt{\frac{\log 5/\delta}{2n_{T=\min}}}
\end{aligned}$$

and that $0 \leq \sqrt{n_{T=\min}/n_{T=0}} \leq 1$. □

B. Details About the Experiments

B.1. Details About the Simulated Data

B.1.1. Learned IHDP

The IHDP dataset (Hill, 2011) are the results of a randomized control trial. In order to be able to simulate the potential outcomes, we train generative models on it and use these generative models as our data generating process.

First, we train a Forest Diffusion (Jolicoeur-Martineau et al., 2023) model to generate the covariates X. We then train two Gaussian Processes: one to predict Y^1 from X based on samples from $Y|X, T = 1$, and another to predict Y^0 from X based on samples from $Y|X, T = 0$. The use of Gaussian Processes here allows us to sample from the predictive distribution, introducing variability into the potential outcomes, an essential element of real data. Finally, a calibrated random forest model is trained to predict the treatment assignments from X.

Such data is guaranteed to satisfy ignorability, since the treatment assignment is determined independenly from the potential outcomes, given the covariates X. Moreover, since the original data is an RCT, it is expected that the data will roughly resemble an RCT (but not exactly, since the treatment assignment is allowed to vary over the X).

B.1.2. ACIC16

This is the data from the 2016 edition of the Atlantic Causal Inference Competition (Dorie et al., 2017), widely used in previous works.

It is synthetic, providing even the potential outcomes, but crafted to resemble real data. Nevertheless, it is guaranteed to satisfy the standard ignorability (i.e., no hidden confounding) and positivity assumptions.

B.1.3. Confounded ACIC16

This is built on top of the data from ACIC16. Having generated X, T, Y^1, Y^0 as in ACIC16, we now modify the potential outcomes in order to violate positivity and ignorability by making it so that when $T = 0$, Y^1 is offset by -20. This makes it so that T (which is not part of X) becomes a hidden confounder, and positivity is violated (since certain outcomes happen only in the counterfactuals). Finally, this modification makes it so that the true treatment propensities (accounting for the unobserved confounder) are exactly equal to T.

B.2. Learning reweightings and the ν

B.2.1. Learning to reweight

We consider two options to learn the reweighting functions $w(X)$. One is to simply use the constant reweighting $w \equiv 1$, which is a surprisingly strong option.

Another option is to approximate the "optimal" weights given by

$$w^{a\star}(X) = \frac{\mathbb{P}[T = a]}{\mathbb{P}[T = a|X]}.$$

Were we to use these precise weights under no hidden confounding and positivity, we would fully eliminate the gap between the observed and complete distributions, since for any $\phi(X)$,

$$\mathbb{E}[w^{a\star}(X) \cdot \phi(X)|T = a] = \mathbb{E}[\phi(X)].$$

To approximate $w^{a\star}$, we first train a classifier $\widehat{e}(X)$ to estimate $\mathbb{P}[T = a|X]$ and estimate the probability $\mathbb{P}[T = a]$ via its sample mean as $\widehat{p}_{T=a}$, and produce the following intermediate unnormalized approximation $\widetilde{w}^a(X)$ for the weights:

$$\widetilde{w}^a(X) = \widehat{p}_{T=a}/\widehat{e}(X).$$

To ensure that $\mathbb{E}[\widehat{w}^a(X)|T = a] = 1$ (as required by our bounds), we normalize $\widetilde{w}^a(X)$ by its sample mean M:

$$\widehat{w}^a(X) = \widetilde{w}^a(X)/\widehat{M}, \qquad \widehat{M} = \widehat{\mathbb{E}}[\widetilde{w}^a(X)].$$

B.2.2. LEARNING A ν

The $\nu(X)$ is a probabilistic classification model for the treatment assignment (possibly the very same one used for weight estimation), with $\nu(X)$ being the predicted probability of $T = a$ given X (i.e., $\nu(X) \approx \mathbb{P}[T = a|X]$). Since ν appears in the bound within a classification loss of the model for predicting T (the Brier score), a better classification model (i.e., a better ν) means a better bound.

B.3. Details about the figures

In Figure 1, all means were estimated via the standard empirical mean with an adequately high number of samples. The underlying models used were Random Forests as per Scikit-Learn's implementation (Pedregosa et al., 2011) with the default hyperparameters. For the prior work bound of (Johansson et al., 2022), the estimation of the Wasserstein distance (which is highly nontrivial) was done with `GeomLoss`, which implements the method from (Feydy et al., 2018).

Figure 2 was computed on the Confounded ACIC16 dataset on the loss of outcome regression of Y^1, with the relevant means being estimated via the standard empirical mean from a sufficiently large number of samples.

In Figure 3, none of the models use sample reweighting. An alternate version of the figure including such models can be found in Section C. The loss used in the figure is the mean squared loss.

C. More figures

In what follows:

- Quantile loss refers to the quantile loss with $\alpha = 0.8$;
- 0-1 loss refers to the 0-1 loss for predicting whether the target is above its median value.

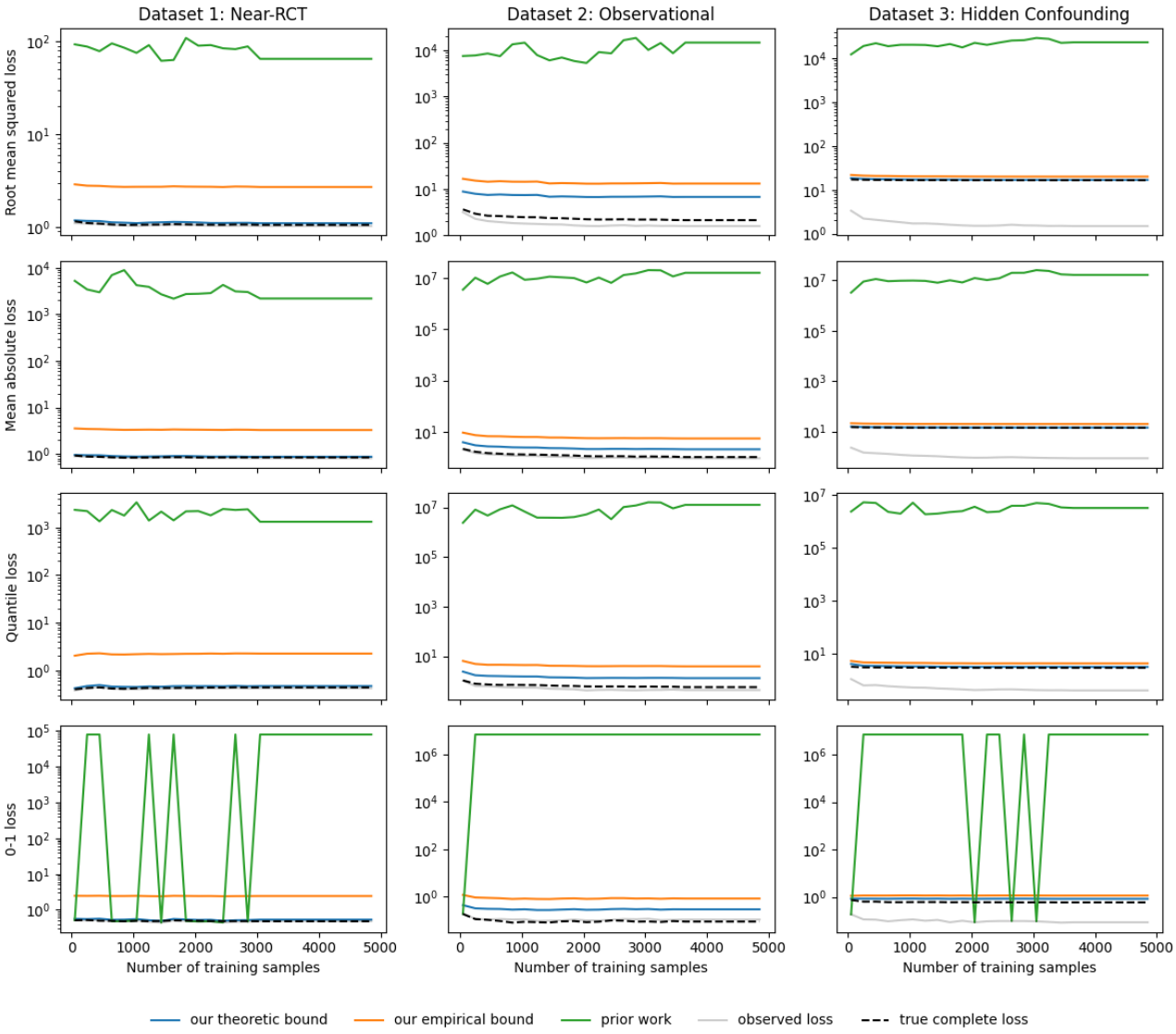


Figure 4. **Alternate version of Figure 1 for outcome regression of Y^1 and including more losses.**

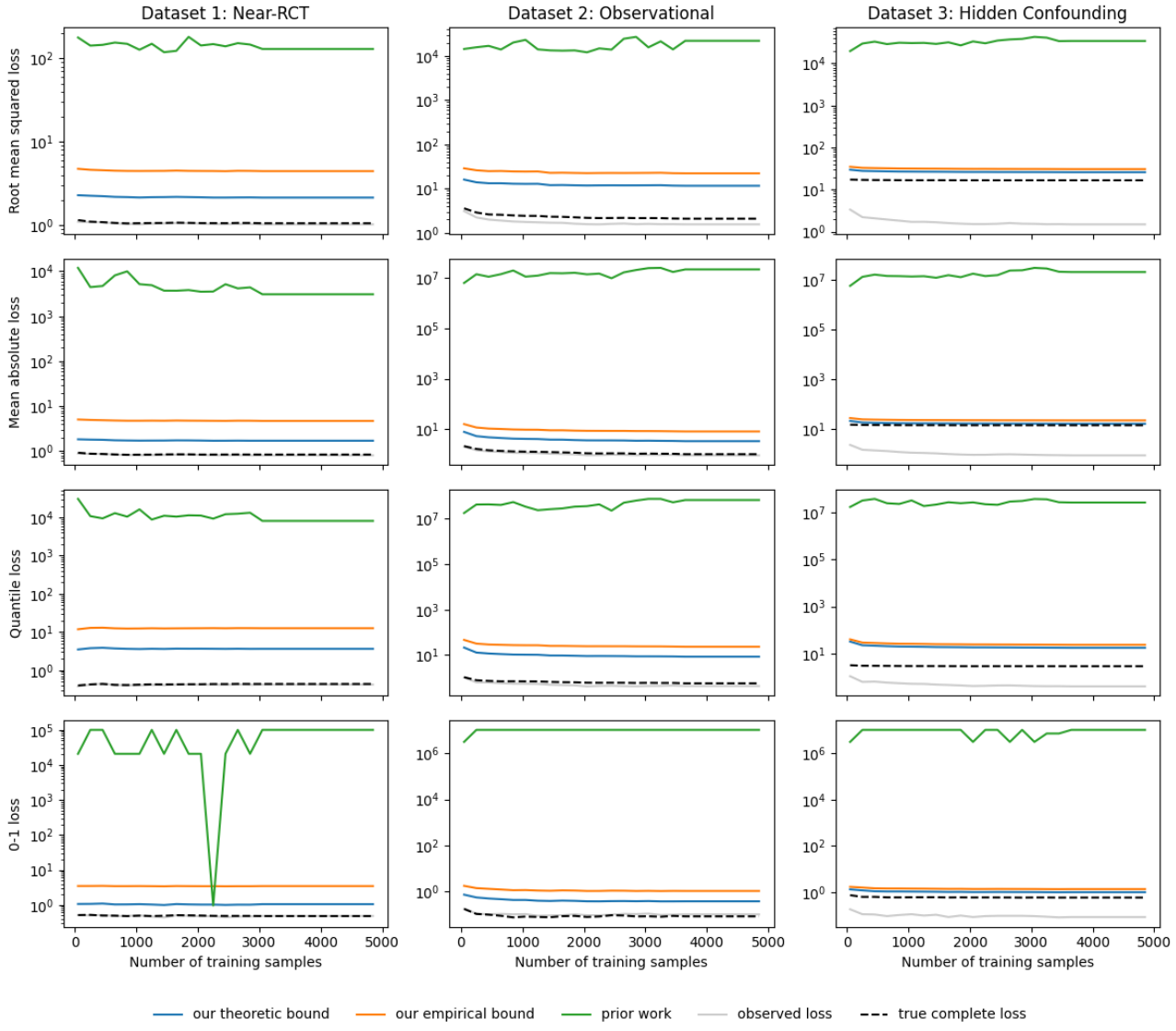
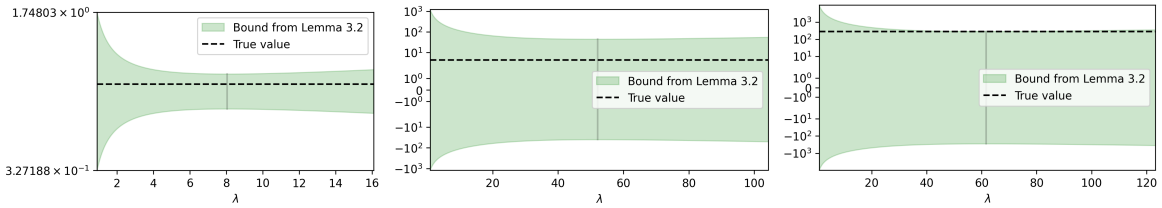


Figure 5. **Alternate version of Figure 1 for T-learners and including more losses.**



(a) Near-RCT dataset. (b) Observational dataset. (c) Hidden Confounding dataset.

Figure 6. **Alternate versions of Figure 2.**

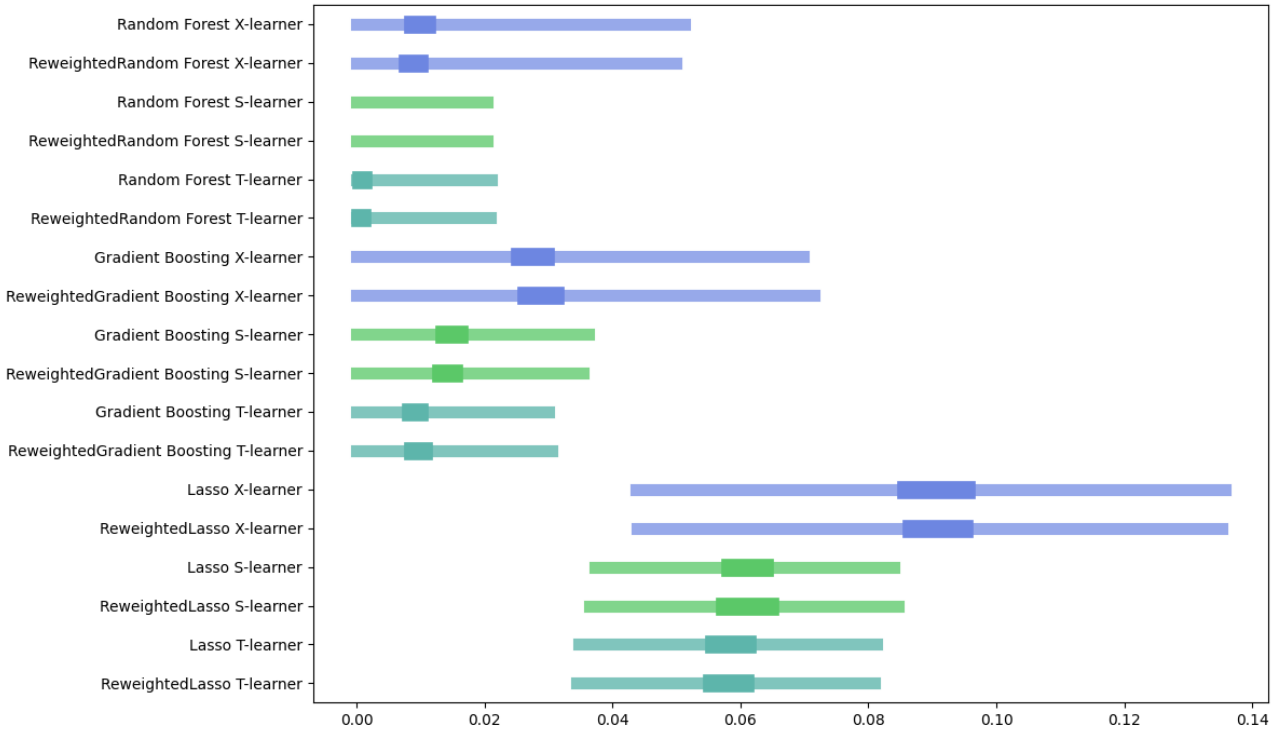


Figure 7. **Alternate version of Figure 3 which includes models using sample reweighting.**