Sensitivity Analysis under the $f$-Sensitivity Models: Definition, Estimation and Inference

Ying Jin$^1$, Zhimei Ren$^2$, and Zhengyuan Zhou$^3$

1Department of Statistics, Stanford University
2Department of Statistics, University of Chicago
3Stern School of Business, New York University

Abstract

The digitization of the economy has witnessed an explosive growth of available observational data across a variety of domains, which has provided an exciting opportunity for causal inference that aims to discern the effectiveness of different treatment options. However, unlike in experimental data – where the designer has the full control to randomize treatment choices based on observable covariates – causal inference using observational data faces the key challenge of confoundedness: there might exist unobservable covariates that affect both the treatment choice and the outcome.

In this paper, we aim to add to the emerging line of tools for sensitivity analysis that assesses the robustness of causal conclusions against potential unmeasured confounding. We propose a new sensitivity model where, in contrast to uniform bounds on the selection bias in the literature, we assume the selection bias is bounded “on average”. It generalizes the widely used $\Gamma$-marginal selection model of Tan (2006) and is particularly suitable for situations where the selection bias grows unbounded locally but, due to the impact of such region being small overall, is controlled at the population level. We study the partial identification of treatment effects under the new sensitivity model. From a distributional-shift perspective, we represent the bounds on counterfactual means via certain distributionally robust optimization programs. We then design procedures to estimate these bounds, and show that our procedure is doubly robust to the estimation error of nuisance components and remains valid even when the optimization step is off. In addition, we establish the Wald-type inference guarantee of our procedure that is again robust to the optimization step. Finally, we demonstrate our method and verify its validity with numerical experiments.

1 Introduction

Causal inference using observational data is a necessity in many fields for several reasons. First, experimental data is often limited and costly to obtain. Second, modern advances in digitization technology have brought forth a growing availability of observational data across a variety of domains. Third, even in settings where experimentation is an economic possibility, observational data can help support initial inquiries and inform effective designs of experiments, by, for instance, identifying the unknown/least confident regimes that require further experimental probing. As such, developing efficient – both statistically and computationally – causal inference schemes using observational data has immense practical value.

In contrast to experimental data where the experimenter has the luxury to randomize treatment options, the key difficulty with observational data is the (untestable) presence of unobserved confounding covariates (also known as confounders). A confounder is a variable $U$ that affects both the treatment $T$ and the outcome $Y$, and whose presence could invalidate the causal relationship drawn from the inference. For instance, one might observe that patients that are sent to ICUs (intensive care units) have a high death rate. However, one would be wrong to conclude that admission to ICUs (a treatment option) causes high likelihood of death, because the patients who are sent to ICUs are those who are already very sick. As such, in this example,
the patient condition is a confounder, which impacts both the treatment selected (ICU/non-ICU) and the medical outcome (death/recovery): ignoring this confounder will lead to faulty conclusions.

To address this challenge, we start by describing the problem of causal inference with observational data and stating the assumptions that will be used throughout the paper. We follow the potential outcome model (Rosenbaum and Rubin, 1983b; Rubin, 1990) and posit an underlying data-generating distribution $\mathbb{P}$ on $(X, U, T, Y(1), Y(0))$, where $X \in \mathcal{X}$ is the observed covariate vector and $\mathcal{X}$ is compact, $U \in \mathcal{U}$ is the unobserved confounding vector, $T \in \{0, 1\}$ is the treatment option, and $Y(1) \in \mathbb{R}$ and $Y(0) \in \mathbb{R}$ are the two potential outcomes. We assume access to a dataset $\{(X_i, T_i, Y_i)\}_{i=1}^n$ of $n$ triples, where $X_i$ is unit $i$’s covariate vector, $T_i$ is the treatment applied to unit $i$ and $Y_i = Y_i(T_i)$ is the observed outcome for unit $i$ under treatment $T_i$. Note that although there are two potential outcomes $Y_i(0)$ and $Y_i(1)$, only $Y_i(T_i)$ – the potential outcome corresponding to the actual applied treatment $T_i$ – is observed. Note that without loss of generality (since $U$ is arbitrary), $\mathbb{P}$ satisfies

\[(Y(1), Y(0)) \perp T \mid X, U. \quad (1)\]

We are interested in estimating the following quantities:

1. Average treatment effect on the control (ATC): $\mathbb{E}[Y(1) - Y(0) \mid T = 0]$,
2. Average treatment effect on the treated (ATT): $\mathbb{E}[Y(1) - Y(0) \mid T = 1]$,
3. Average treatment effect (ATE): $\mathbb{E}[Y(1) - Y(0)]$,

where in all quantities the expectation is taken with respect to the underlying joint distribution. To make progress, we assume that there is sufficient exploration in the dataset, known as the overlap assumption in the literature. We define the observed propensity score $e(x) \equiv \mathbb{P}(T = 1 \mid X = x)$.

**Assumption 1** (Overlap). $0 < e(x) < 1$ for $\mathbb{P}$-almost all $x \in \mathcal{X}$.\(^4\)

Under the overlap assumption, classical causal inference (see Imbens (2004); Imbens and Rubin (2015) for two excellent surveys) has extensively treated the topic of estimating ATC and ATT (along with other treatment effect quantities) under the unconfoundedness assumption (a.k.a. strong ignorability (Rosenbaum and Rubin, 1983b)), where it is assumed that $(Y(1), Y(0)) \perp T \mid X$ (i.e. no confounders $U$ exist). As previously mentioned, this assumption is not testable and hard to justify in observational data, where the statistician often has little knowledge and/or control over how the dataset was collected. The key technical challenge arising therefrom is that the counterfactual distribution $\mathbb{P}_{X,Y\mid T=0}$ (and similarly $\mathbb{P}_{X,Y\mid T=1}$) is not identifiable using observed data. To address this problem, a prominent approach, dating back to Cornfield et al. (1959); Bross (1966), is sensitivity analysis, which studies how robust a certain causal conclusion drawn from observational data is against unmeasured confounding. In particular, a sensitivity model often makes assumptions on how much the treatment selection probability when conditioned on the confounders deviates from the treatment selection probability when unconditioned on the confounders. From such an assumption on the deviation, the statistician then proceeds to estimate the range – rather than a single value – of the underlying treatment effect quantity, where the width of the range depends on the deviation bounds. Such an approach is valuable in that it provides the statistician with a quantitative understanding how much the treatment effect is under varying severities of confoundedness.

Within sensitivity analysis, Rosenbaum’s $\Gamma$-selection sensitivity model (Rosenbaum, 1987) provides a pioneering model on the selection bias that has instantly become a classic. Subsequently, it was generalized by the marginal $\Gamma$-selection sensitivity model proposed in Tan (2006), which imposes weaker assumptions and is widely adopted. Based on Tan (2006), a series of work have developed various treatment effects estimation

---

\(^1\)Here we loosely refer to all patient medical characteristics as a confounder to illustrate that $U$ could affect both $T$ and $Y$. In practice, some of the patient medical characteristics are recorded and hence observable, while others are not. It is the unobserved confounding factors that pose challenges to estimation and inference using observational data.

\(^2\)We call $T = 1$ and $T = 0$ treat and control, respectively.

\(^3\)In using the notation $Y_i = Y_i(T_i)$, we are implicitly making the Stable Unit Treatment Value Assumption (SUTVA) that is standard in the literature. This assumption states that “the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units” Cox (1958).

\(^4\)Since $\mathcal{X}$ is compact, this is equivalent to assuming that $\eta \leq e(x) \leq 1 - \eta$ for some positive $\eta$, as used in certain versions of overlap in the literature.
and inference schemes (Zhao et al., 2017; Kallus et al., 2019; Lee et al., 2020; Dorn and Guo, 2021; Jin et al., 2021; Nie et al., 2021; Dorn et al., 2021). More specifically, Tan (2006) assumes uniform upper and lower bounds on selection bias, in terms of the odds ratio of receiving treatment conditional on both unmeasured confounders and observed covariates versus conditional only on observed covariates, specified as

\[ \frac{1}{\Gamma} \leq \frac{P(T = 1 | X = x, U = u)/P(T = 0 | X = x, U = u)}{P(T = 1 | X = x)/P(T = 0 | X = x)} \leq \Gamma \tag{2} \]

for \( P \)-almost all \( x \in \mathcal{X} \) and \( u \in \mathcal{U} \) for some \( \Gamma \geq 1 \). Note that when \( \Gamma = 1 \), this assumption recovers the unconfoundedness assumption, and the larger the \( \Gamma \), the more confoundedness the model tolerates.

1.1 Motivation of our approach

Despite being widely used, the marginal \( \Gamma \)-selection model (2) can be limited. We illustrate this point by a simple and natural parametric example, which further motivates our new sensitivity model.

Example 1. Let us consider a simple example where there are no covariates, and the statistician would like to infer the counterfactual mean of \( Y(1) \) in the control group (i.e. \( E[Y(1) | T = 0] \)). We assume the observed probability of treatment is \( P(T = 1) = 1/2 \). Instead of believing in the strong ignorability condition (that is, all units receive treatments with the same probability in our example), the statistician would like to estimate the range of \( E[Y(1) | T = 0] \) if the observational data is confounded to some extent. She might think of a logistic model \( T \mid U \sim \text{Bern}(\exp(\delta U)/1 + \exp(\delta U)) \) for some \( \delta \in (0, 1) \), where the confounder \( U \sim N(0, 1) \); hence,

\[ P(T = 1) = \mathbb{E}[P(T = 1 | U)] = \mathbb{E}\left[ \frac{\exp(\delta U)}{1 + \exp(\delta U)} \right] = \mathbb{E}\left[ \frac{\exp(-\delta U)}{1 + \exp(-\delta U)} \right] = \frac{1}{2}, \]

where the second-to-last inequality follows by symmetry of \( U \), and agrees with the observed probabilities. For simplicity, we also assume that \( Y(1) = g(U, \delta) \) for some measurable function \( g \), so that \( Y(1) | T = 1 \) agrees with the observations regardless of the choice of \( \delta \) (and hence also \( Y(1) \perp T | U \), satisfying (1)).

By construction, the odds ratio characterizing the selection bias caused by \( U \) is

\[ \text{OR}(U) := \frac{P(T = 0 | U)}{P(T = 1 | U)} \cdot \frac{P(T = 1)}{P(T = 0)} = e^{-\delta U}. \]

With an unbounded \( U \), the above odds ratio can be unbounded and hence the marginal \( \Gamma \)-selection assumption (2) does not hold. Consequently, estimating the range of counterfactual means under a hypothesized uniform bound on \( \text{OR}(U) \) is problematic even in this simple stylized example. More generally, this example highlights a critically inadequate aspect in the marginal \( \Gamma \)-selection model: it may well be that some region of the sample space could be drastically confounded; considering such a possibility, the statistician might be reluctant to imagine a uniform bound on the selection bias. However, one could still imagine that even though the selection bias can have a huge scale in some region, the impact of such region might still be small if it only takes a tiny proportion of the whole population. As such, the locally-severe confounding might not have an overall big impact at the population level. A desirable sensitivity model should still allow us to conduct estimation and inference in such scenarios.

We do mention that the earlier work Imbens (2003) postulates a parametric logistic model for treatment that includes the confounders, which contains the simple example discussed here as a special case and hence allows for unbounded local confounded effects. However, its key limitation is that the proposed confounding model is highly specialized to the logistic form, whereas the marginal \( \Gamma \)-selection criterion provides a non-parametric model that is quite general, which is an important merit. As such, motivated by achieving the best of both worlds, we are naturally led to the following research questions:

Can we formulate a general, non-parametric sensitivity model that can deal with unbounded local confounding effects with an overall bounded impact at the population level? If so, can we efficiently—both statistically and computationally—conduct estimation and inference on average treatment effects?
1.2 Our proposed \(f\)-sensitivity model

We answer the above questions in the affirmative in this paper. In this subsection, our focus is on providing a quick answer to the first question, where we briefly describe our proposed \(f\)-sensitivity model, characterized by the following \((f, \rho)\)-selection condition.

**Definition 1** (The \((f, \rho)\)-selection condition). Suppose \(f : \mathbb{R}_+ \mapsto \mathbb{R}\) is a convex function such that \(f(1) = 0\). \(\mathbb{P}\) satisfies the \((f, \rho)\)-selection condition if

\[
d_f(\mathbb{P}) := \max \left\{ \int f \left( \frac{P(T = 0 | X = x, U) \ P(T = 1 | X = x)}{P(T = 1 | X = x, U) \ P(T = 0 | X = x)} \right) d\mathbb{P}_{(X=x,T=1)} , \int f \left( \frac{P(T = 1 | X = x, U) \ P(T = 0 | X = x)}{P(T = 0 | X = x, U) \ P(T = 1 | X = x)} \right) d\mathbb{P}_{(X=x,T=0)} \right\} \leq \rho ,
\]

for \(\mathbb{P}\)-almost all \(x\).

**Remark 1.** If \(\mathbb{P}\) satisfies the marginal \(\Gamma\)-selection condition (2), then it automatically satisfies the \((f, \rho)\)-selection condition with any qualified \(f\) and \(\rho = \max \{ f(1/\Gamma), f(\Gamma) \}\). This can easily be checked by noting that (2) implies

\[
f \left( \frac{P(T = 1 | X = x, U) \ P(T = 0 | X = x)}{P(T = 0 | X = x, U) \ P(T = 1 | X = x)} \right) \leq \max \{ f(1/\Gamma), f(\Gamma) \},
\]

by convexity of \(f\), thereby leading to the bound on the second integral in \(d_f(\mathbb{P})\):

\[
\int f \left( \frac{P(T = 1 | X = x, U) \ P(T = 0 | X = x)}{P(T = 0 | X = x, U) \ P(T = 1 | X = x)} \right) d\mathbb{P}_{(X=x,T=0)} \leq \max \{ f(1/\Gamma), f(\Gamma) \}.
\]

The first integral in \(d_f(\mathbb{P})\) can be bounded by exactly the same way by the symmetry of (2).

**Remark 2.** In the definition of the \((f, \rho)\)-selection condition, we take the maximum of two integrals, each from one direction. This is mainly to keep the condition symmetric with regards to the choice of the treated or control groups, and align with the convention in the sensitivity models in the literature (Tan, 2006). That said, as we would see shortly in Section 2, it might be more natural to only work with one of them (i.e. assume one of them be bounded by \(\rho\)) when our primary interest is to estimate and perform inference on one of the counterfactuals.

We now show that the proposed \((f, \rho)\)-selection model addresses the unbounded confoundness in Example 1: even though the selection bias (the odds ratio) is not uniformly bounded, it is controlled overall; in this case, our proposed model can be a better characterization of the selection bias.

**Example 1** (Continued). We take \(f(t) = t \log t\), a convex function with \(f(1) = 0\). Continuing the computation in Example 1, the first term in \(d_f(\mathbb{P})\) of Definition 1 can be computed as

\[
\int f(\text{OR}(U)) d\mathbb{P} = \int -\delta U \cdot e^{-\delta U} d\mathbb{P}_{(T=1)} = -\delta \int u \cdot e^{-\delta u} \cdot \frac{2e^{\delta u}}{\sqrt{2\pi}} \cdot \frac{1}{e^{-u^2/2}} du < \infty . \quad (3)
\]

The right-handed side is approximately 0.2 if we take \(\delta = 1\) and 0.6 if we take \(\delta = 2\). Note that \(\int f(\text{OR}(U)) d\mathbb{P}_{(T=1)}\) can be interpreted as the overall deviation of \(\text{OR}(U)\) from 1 in the treated (observed) group, which is bounded, even though \(\text{OR}(U) \to \infty\) when \(U \to -\infty\).

While the uniform bound on selection bias is infeasible in this case, the procedures proposed in this paper still yield meaningful bounds on the counterfactual mean \(E[Y(1) | T = 0]\). We apply our procedure to Example 1 for a sequence of \(\delta\) and the corresponding upper bounds in (3). The results are summarized in Figure 1, where we assume \(Y(1) | T = 1 \sim \mathcal{N}(0, 1)\) for simplicity.
Our contributions are threefold. First, we propose the $f$-sensitivity model, a general non-parametric model that allows for unbounded local confounding with a limited overall impact at the population level. In Section 2, we show a distributional shift interpretation of our $f$-sensitivity model that is perhaps more illuminating: under the $(f, \rho)$-selection condition, the (unobservable) counterfactual distribution $P_{Y(1) | X=x, T=0}$ is close to the observable distribution $P_{Y(1) | X=x, T=1}$ under the $f$-divergence measure, whose precise definition is provided shortly. This allows us to represent the partial identification lower (resp. upper) bound on the unidentifiable $E[Y(1) | T=0]$ based on the observable (identifiable) distribution $P_{Y(1) | X=x, T=1}$ with a distributionally robust optimization problem we provide in Section 3.1, where we infer the smallest (or largest) objective within the $f$-divergence ball of radius $\rho$ centered at $P_{Y(1) | X=x, T=1}$.

Building on this distributional shift perspective, we provide in Section 3 an estimation procedure for the bounds on counterfactual means and treatment effects. We first transform the distributionally-robust optimization problem into an equivalent dual problem, through which the bounds are in fact connected to the optimal objective of a risk minimization problem. We then utilize an empirical risk minimization (ERM) procedure to solve for the risk minimizers. Finally, we propose a cross-fitted estimator for the bounds, where we adjust for the estimation error of the plugged-in empirical risk minimizer and other nuisance components.

Third, in Section 4, we establish the estimation and inference guarantees of our procedures for counterfactual means, demonstrating certain robustness properties. We show that our estimators are doubly robust to the nuisance estimation when the ERM step is consistent, and remain valid even when the ERM step is misspecified. Furthermore, we show that our cross-fitted estimator yields Wald-type inference guarantees even when all the nuisance components can only be estimated at a slower-than-parametric rate; such inferential guarantee is again robust to the misspecification of the ERM step. Based on the the bounds on counterfactual means, we provide confidence intervals for the treatment effects.

1.4 Related work

In addition to the sensitivity analysis literature mentioned before, several other works exist that take the approach of modelling the impact of unmeasured confounding is through bounds on selection bias. Remarkably, Rosenbaum and Rubin (1983a) studies the impact of selection bias among matched pairs, which is further extended by a series of works (Rosenbaum, 1987; Gastwirth et al., 1998; Rosenbaum, 2002b,a) to sensitivity models that uniformly bounds the selection bias among samples with matching covariates.

Also related to sensitivity analysis under uniformly bounded selection bias is Yadlowsky et al. (2018), which works under an extension of Rosenbaum’s sensitivity model that is similar to Tan (2006). Besides...
modelling the selection bias among treatment, observed covariates and unmeasured confounders, we mention in passing that Ding and VanderWeele (2016) considers the sensitivity analysis of other metrics than treatment effects.

Our novel sensitivity models provide a perspective of studying the distributional shift from observations to the counterfactuals, which echo the ideas of several previous works (Jin et al., 2021; Yadlowsky et al., 2018; Dorn et al., 2021). In particular, we formulate the estimand via an optimization problem under constraints on distributional shifts, similar to Yadlowsky et al. (2018); Dorn et al. (2021). However, as they work under different sensitivity models, the techniques employed in our works are in sharp distinction from theirs.

The $f$-divergence bounds derived from our sensitivity models are related to a series of works that characterize discrepancy between distributions using $f$-divergence (Rényi, 1961; Morimoto, 1963; Csiszár, 1964; Liese and Vajda, 2006). For example, Christensen and Connault (2019) places bounds on the marginal distributions of hidden variables in structural equation models; Andrews et al. (2020) studies parameter estimation when the joint distributions of variables are within an $f$-divergence ball; Rahman (2016) estimates the $f$-divergence between conditional and unconditional distributions to measure variable importance; we differ from them both in terms of the target (bounding selection bias), the estimand (estimation of treatment effects) and the resulting formulation ($f$-divergence between conditional distributions, while the shifts between $X$-distribution are identifiable). In particular, these existing $f$-divergence bounds are not sufficient to capture the impact of unmeasured confounding in our problem setting.

Our work also finds connection to a line of work estimation, inference and learning under various types of distributional shifts, e.g., $f$-divergence, outside the task of inferring causal effects, including Duchi and Namkoong (2021) which studies the empirical risk minimization problem when the joint distribution of $(X, Y)$ shifts within an $f$-divergence ball. Si et al. (2020) which studies the policy learning for contextual bandits under unknown marginal distribution shifts, Gupta and Rothenhäusler (2021) that studies the estimation and inference of statistical parameters under distributional shifts in certain directions, etc. As we work under distinct type of (conditional) distributional shifts, we need a different set of tools than theirs.

Finally, the post-hoc adjustment method we propose to deal with nuisance components is also generally connected to a vast body of missing data literature with unknown missing mechanisms, although in different contexts and with different details. In particular, we use the cross-fitting (Schick, 1986; Zheng and Laan, 2011; Chernozhukov et al., 2018) technique to mitigate the error in nuisance component estimation; the technique of bias correction using a different dataset under covariate shift is also related to and inspired by Jin and Rothenhäusler (2021).

2 Distributional Shifts under the $f$-Sensitivity Model

The $f$-sensitivity model leads to constraints on the distributional shifts from the observations to the counterfactuals; this identifies a new class of robust inference problems that are new to the literature. We can thus view sensitivity analysis through the lens of distributionally-robust inference. Take the counterfactual inference on $Y(1)$ for $T = 0$ as an example (the case of $Y(0)$ is completely symmetric): since we are able to observe $Y(1)$ in the treated group, the distribution of $(X, Y(1)) | T = 1$ is identifiable, and decomposes as

$$P_{X,Y(1)|T=1} = P_X|T=1 \times P_{Y(1)|X,T=1}.$$  

The counterfactual distribution of $(X, Y(1))$ in the control group is

$$P_{X,Y(1)|T=0} = P_X|T=0 \times P_{Y(1)|X,T=0};$$

where part (a) is identifiable but part (b) is not without the strong ignorability condition. Put another way, the distributional shift from $P_{X,Y(1)|T=1}$ to $P_{X,Y(1)|T=0}$ has an identifiable component as a function of $X$, and an unidentifiable one concerning the conditional distribution of $Y|X$; the latter, however, is bounded under the $(f, \rho)$-selection condition in terms of $f$-divergence defined as follows.

**Definition 2 ($f$-divergence).** Let $P$ and $Q$ be two probability distributions over a space $\Omega$ such that $P$ is absolutely continuous with respect to $Q$. For a convex function $f$ such that $f(1) = 0$, the $f$-divergence of $P$ from $Q$ is defined as $D_f(P \parallel Q) = E_Q[f\left(\frac{dP}{dQ}\right)]$, where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative.
Proof of Lemma 1. **Under the \((f,\rho)\)-selection condition, we have**

\[
D_f(\mathbb{P}_{Y(1)} | X=x,T=0 \parallel \mathbb{P}_{Y(1)} | X=x,T=1) \leq \rho
\]

for \(\mathbb{P}_{X|T=1}\)-almost all \(x\); that is, the \(f\)-divergence between the conditional distributions in the two groups are bounded by \(\rho\) for almost all \(X\) in group \(T = 1\).

**Proof of Lemma 1.** Suppose a distribution \(\mathbb{P}\) over the \((X,U,T,Y(0),Y(1))\) satisfies the \((f,\rho)\)-selection condition. By the data-processing inequality,

\[
D_f(\mathbb{P}_{Y(1)} | X=x,T=0 \parallel \mathbb{P}_{Y(1)} | X=x,T=1) \leq D_f(\mathbb{P}_{Y(1)|U} | X=x,T=0 \parallel \mathbb{P}_{Y(1)|U} | X=x,T=1)
\]

\[
=E_{Y(1),U} [f\left(\frac{d\mathbb{P}_{Y(1)|U} | X=x,T=0}{d\mathbb{P}_{Y(1)|U} | X=x,T=1} \right)]
\]

Above, the likelihood ratio can be decomposed as

\[
\frac{d\mathbb{P}_{Y(1)|U} | X=x,T=0}{d\mathbb{P}_{Y(1)|U} | X=x,T=1} = \frac{d\mathbb{P}_{Y(1)} | U,X=x,T=0}{d\mathbb{P}_{Y(1)} | U,X=x,T=1} \cdot \frac{d\mathbb{P}_{U} | X=x,T=0}{d\mathbb{P}_{U} | X=x,T=1}
\]

\[
\leq \mathbb{P}(T = 0 | X = x, U) \cdot \mathbb{P}(T = 1 | X = x) / \mathbb{P}(T = 1 | X = x, U) \cdot \mathbb{P}(T = 0 | X = x),
\]

where step (a) is due to condition (1). Combining the above, we have

\[
D_f(\mathbb{P}_{Y(1)} | X,T=0 \parallel \mathbb{P}_{Y(1)} | X,T=1) \leq E_{U,X,T=1} [f\left(\frac{\mathbb{P}(T = 0 | X, U)}{\mathbb{P}(T = 1 | X, U)} \cdot \frac{\mathbb{P}(T = 1 | X)}{\mathbb{P}(T = 0 | X)}\right)]
\]

\[
=E_{U,X,T=1} [f\left(\frac{\mathbb{P}(T = 0 | X, U)}{\mathbb{P}(T = 1 | X, U)} \cdot \frac{\mathbb{P}(T = 1 | X)}{\mathbb{P}(T = 0 | X)}\right)] \leq \rho,
\]

where the last inequality is due to the \((f,\rho)\)-selection condition. 

In Proposition 1, we characterize the counterfactual distributions induced by all super-populations that agrees with the observables and satisfies our sensitivity models.

**Proposition 1.** Let \(\mathbb{P}_{\text{sup}}\) be the true unknown super-population over \((X,U,T,Y(0),Y(1))\) and let \(\mathcal{P}\) be the set of all distributions over \((X,U,T,Y(0),Y(1))\). Let \(\mathbb{P}_{X,Y,T}^{\text{obs}}\) be the joint distribution of all observable random variables \((X,Y,T)\). Define \(\mathcal{Q}_{1,0}\) to be the set of all counterfactual distributions that agrees with the observables and satisfies the \((f,\rho)\) selection condition, i.e.,

\[
\mathcal{Q}_{1,0} = \{P_{X,Y,T} | \mathcal{Q}_{1,0} \subset \mathcal{Q} : \frac{d\mathbb{P}_{X,Y,T}}{d\mathbb{P}_{X,Y,T}} | T=1, r_{1,0}(x), D_f(\mathbb{P}_{Y|X=x} | P_{X,Y,T}^{\text{obs}} | X=x,T=1) \leq \rho, \text{ for } \mathbb{P}_{X,Y,T}^{\text{obs}} | T=1 \text{-almost all } x \},
\]

where \(r_{1,0}(x) = \frac{(1-e(x))p_1}{e(x)T_1} \cdot \mathbb{P}(T = 1 | X = x), p_1 = \mathbb{P}(T = 1)\).

We defer the proof of Proposition 1 to Appendix A, where we prove a stronger version of Proposition 1 that gives a tight characterization of \(\mathcal{Q}_{1,0}\). Symmetrically, we can also define \(\mathcal{Q}_{0,1}\) as the identification set of \(\mathbb{P}_{X,Y(0)|T=1}\); the tight characterization of \(\mathcal{Q}_{0,1}\) is also given in Appendix A. Throughout this paper, we
mainly work under the ambiguity sets in Proposition 1 to emphasize the more general distributionally-robust estimation aspect of this problem.

Proposition 1 identifies a new class of robust inference problems, where the target distribution has an identifiable $X$-shift and the unidentifiable conditional distribution is restricted in $f$-divergence ball; it is similar to the ambiguity set studied in Jin et al. (2021). This model is closely related to, but quite distinct from other ambiguity sets considered in the literature that involves $f$-divergence (Duchi and Namkoong, 2021; Si et al., 2020; Andrews et al., 2020) that only concern marginal distributions.

Remark 3 (Relation to other $f$-divergence bounds). Previous works in the literature (see Section 1.4 for a summary) often work under the $f$-divergence ball around the marginal distribution $P_{X,Y}$, characterized by

$$\tilde{Q} = \{Q: D_f(Q_{X,Y} || P_{X,Y}) \leq \rho\}. \quad (4)$$

While in our formulation, the ambiguity sets take the form

$$Q = \{Q: \frac{dQ}{dP_{X,Y}}(x) = r(x), \ D_f(Q_{Y|X} || P_{Y|X}) \leq \rho\}, \quad (5)$$

where $r(x)$ is a known or identifiable function. Such distinction is similar to Remark 3 of Jin et al. (2021). Instead of bounding the overall shift as in (4), the constraint in (5) actually allows freedom in the shift of $X$: the sets in (5) can be small as long as $\rho$ is small. For counterfactual inference under the strong ignorability condition, the set (5) is a singleton even if $P_X$ and $\bar{P}_X$ are drastically different, while (4) would require a large $\rho$ to hold. More generally, when there is a known (or estimable) large shift in $P_X$ but a relatively small shift in $P_{Y|X}$, (5) provides a tighter range of the target distributions, and the methods we develop in this paper can be directly applied.

3 Estimating Treatment Effects under the $f$-Sensitivity Model

3.1 Bounding treatment effects

We first focus on the bounds on counterfactual means, which serve as building blocks for bounding various treatment effects. For example, $E[Y(1)|T = 0]$ can be written as $E[Y(1)w(X,Y(1))|T = 1]$, where $w(x,y) := \frac{dP_{X,Y(1)|T=0}}{dP_{X,Y(1)|T=1}}$ is the likelihood ratio that lies within the ambiguity set in Proposition 1, and $P_{X,Y(1)|T=1}$ is a distribution that one has access to. Proposition 2 explicitly characterizes the bounds on $E[Y(1)|T = 0]$ under the $(f,\rho)$-selection condition; they are provided as solutions to convex optimization problems that only involve identifiable quantities.

Proposition 2. Let $\mu_{1,0}^-$ (resp. $\mu_{1,0}^+$) be the optimal objective function of the convex optimization problem

$$\min_{\text{resp. max}} E[Y(1)L(X)|T = 1]$$

s.t. $E[L(x)|X = x, T = 1] = r_{1,0}(x)$

$$E[f(L(x)/r_{1,0}(x))|X = x, T = 1] \leq \rho, \ for \ almost \ all \ x,$$

where all the expectations are induced by the observed distribution. Then $\mu_{1,0}^- \leq E[Y(1)|T = 0] \leq \mu_{1,0}^+$ under the $(f,\rho)$-selection condition.

Proposition 2 immediately implies bounds on the ATC: denote the observable group-wise means as $\mu_{t}^{\text{obs}} := E[Y(t)|T = t]$ for $t \in \{0,1\}$; then under the $(f,\rho)$-selection condition, the ATC is bounded as

$$\mu_{1,0}^- - \mu_{0,1}^{\text{obs}} \leq E[Y(1) - Y(0)|T = 0] \leq \mu_{1,0}^+ - \mu_{0,1}^{\text{obs}}. \quad (7)$$

Switching the role of 1 and 0 in Proposition 2, one can obtain bounds on $E[Y(0)|T = 1]$; let $\mu_{0,1}^+$ and $\mu_{0,1}^-$ denote the upper and lower bound on $E[Y(0)|T = 1]$, respectively, we then get bounds on the ATT:

$$\mu_{1,0}^{\text{obs}} - \mu_{0,1}^- \leq E[Y(1) - Y(0)|T = 1] \leq \mu_{1,0}^{\text{obs}} - \mu_{0,1}^+.$$
By the decomposition of ATE (average treatment effects), we also have the representation of lower and upper bounds for \( E[Y(1) - Y(0)] \). Under the \((f, \rho)\)-selection condition, we have
\[
\mu_1^{\text{obs}} - \mu_{0,1}^+ + p_0(\mu_{1,0} - \mu_0^{\text{obs}}) \leq E[Y(1) - Y(0)] \leq \mu_1^{\text{obs}} - \mu_{0,1}^- + p_0(\mu_{1,0} + \mu_0^{\text{obs}}).
\]
These bounds thus boil down to the estimation of \( \mu_{1,1}^- \) under our sensitivity models, the optimal objective value of the convex optimization problems in Proposition 2.

**Remark 4.** As we mentioned before, optimal objectives of the problems in Proposition 2 are not necessarily tight bounds for counterfactual means. To align with the literature and keep a relatively clean formulation of dual problems, we only account for the direction \( D_f(P_{Y(1)}|X=x,T=0) \) \( \| P_{Y(1)}|X,T=1 \) when considering \( E[Y(1)|T=0] \). For completeness, we include the discussion of the tight bounds on counterfactual means, hence ATT and ATC, in Section 6. We also note that combining sharp bounds on ATT and ATC does not necessarily lead to sharp bounds on ATE, as they might be attained by different super-populations. We leave the investigation of sharp bounds on ATEs for future pursuit.

### 3.2 From the primal to the dual

The infinite-dimensional optimization problem (6) is hard to solve directly. We address this issue by transiting to its dual form, which is easier to tackle with. In the following, we primarily focus on \( \mu_{1,0}^- \), the lower bound on \( E[Y(1)|T=0] \), and the same idea carries over to other quantities. Proposition 3 represents \( \mu_{1,0}^- \) with a dual formulation, whose proof is in Appendix C.1.

**Proposition 3.** The optimal objective of (6) is given by
\[
\mu_{1,0}^- = \inf_{\alpha \geq 0, \eta \in \mathbb{R}} \mathbb{E}_{r_{1,0}(X)} \left\{ \alpha(X)f^\ast \left( \frac{Y(1) + \eta(X)}{-\alpha(X)} \right) + \eta(X) + \alpha(X)\rho \right\} | T = 1,
\]
where \( f^\ast(s) = \sup_{t \geq 0} \{ st - f(t) \} \) is the conjugate function of \( f \). In particular, denoting \( \ell(\alpha, \eta, x, y) = \alpha f^\ast \left( \frac{Y + \eta}{\alpha} \right) + \eta + \alpha \rho \) for \( (\alpha, \eta) \in \mathbb{R}^+ \times \mathbb{R} \), we have \( \mu_{1,0}^+ = -\mathbb{E}\left[ \ell(\alpha^\ast(X), \eta^\ast(X), X, Y(1)) \right] | T = 1 \), where for \( \mathbb{P}_{X|T=1}\text{-almost all } x \),
\[
(\alpha^\ast(x), \eta^\ast(x)) \in \operatorname{argmin}_{\alpha \geq 0, \eta \in \mathbb{R}} \mathbb{E}\left[ \alpha f^\ast \left( \frac{Y(1) + \eta}{-\alpha} \right) + \eta + \alpha \rho \right] | X = x, T = 1.
\]

The formulation (9) can be viewed as a risk minimization problem, where the minimizer \((\alpha^\ast(x), \eta^\ast(x))\) is also the minimizer of per-x conditional risk; as a result, we might employ tools in empirical risk minimization to estimate \( \alpha^\ast(\cdot), \eta^\ast(\cdot) \), and then estimate \( \mu_{1,0}^+ \) by plugging in the empirical risk minimizer.

However, when it involves functions \( \alpha(\cdot), \eta(\cdot) \), the estimation might not converge at a parametric rate; furthermore, another particular challenge is the nuisance component \( r_{1,0}(x) \) that relies on \( e(x) = \mathbb{P}(T = 1|X = x) \), which typically needs to be estimated in observational studies. Similar to the situations for treatment effect estimation without confounding, naive plug-in estimator might suffer from slow rate of convergence. We are to show that, with an adjustment term to the naive plug-in estimator that utilizes covariates in the control group, we can still obtain Wald-type inference when the nuisance components are estimated at a lower-than-parametric rate.

Before proceeding to our procedures, we show a result on the behavior of the optimizer \( \alpha^\ast(x) \): it is positive as long as \( \mathbb{P}_{Y(1)|X,T=1} \) does not have a large point mass at its essential infimum and the function \( f \) satisfy some regularity conditions in the limit. The proof of Proposition 4 is deferred to Appendix B.1.

**Proposition 4.** Define \( y(x) = \sup \left\{ t : \mathbb{P}(Y(1) < t | X = x, T = 1 = 0) \right\} \) and \( \bar{p}(x) = \mathbb{P}(Y(1) = y(x) | X = x, T = 1) \). We assume that \( \bar{p}(x)f(1/\bar{p}(x)) + (1 - \bar{p}(x))f(0) > \rho \) for \( \mathbb{P}_{X|T=1}\text{-almost all } x \). Also suppose there exist constants \( L \) and \( U \) such that \( f(x) \geq L \) for \( x \in \mathbb{R} \), \( f(x) \leq U \) for \( x \leq 0 \), \( \lim_{x \to -\infty} f(x)/x = 0 \) and \( \lim_{x \to \infty} f(x)/x = \infty \). Then the solution to (9) satisfies \( \alpha^\ast(x) > 0 \) for \( \mathbb{P}_{X|T=1}\text{-almost all } x \).

In particular, the conditions on \( f \) hold for a large variety of functions; concrete examples include KL divergence, where \( f(x) = x \log x \) and \( f^\ast(x) = e^{(x-1)} \), as well as \( \chi^2 \)-divergence, where \( f(x) = (x-1)^2 \) and \( f^\ast(x) = \frac{1}{2}((x+2)^2 - 1) \), etc.
In the following, we assume throughout that the conditions of Proposition 4 hold, hence there exists some $\epsilon > 0$ such that $\alpha^*(x) > \epsilon$ for $\mathbb{P}_X \mid_{\tau=1}$-almost all $x$ by the compactness of $\mathcal{X}$. An important implication is that, as $\alpha^*(x)$ lies in the interior of $[0, \infty)$, the gradient of the risk function is typically mean-zero at $(\alpha^*(x), \eta^*(x))$; this would play an important role in our inferential guarantees.

### 3.3 The estimation procedure

We start with splitting the treated and control groups into three folds, denoted as $\mathcal{I}^{(j)}_1$, $\mathcal{I}^{(j)}_0$, $j = 1, 2, 3$, respectively. For each $j = 1, 2, 3$, we use the fold $\mathcal{I}^{(j+1)}_1$ and $\mathcal{I}^{(j+1)}_0$ to obtain an estimator $\hat{r}^{(j)}$ for $r_{1,0}$, and run an empirical risk minimization problem to obtain $\hat{\alpha}^{(j)}$ and $\hat{\eta}^{(j)}$ for $(\alpha^*, \eta^*)$ on the same fold (the estimation details are to be discussed right after the introduction of the procedure). In particular, the optimization of $\hat{\alpha}^{(j)}$ and $\hat{\eta}^{(j)}$ does not require the knowledge of $r_{1,0}$ here. We then define the function

$$\hat{H}^{(j)}(x, y) = \hat{\alpha}^{(j)}(x)f^*(\frac{y + \hat{\eta}^{(j)}(x)}{-\hat{\alpha}^{(j)}(x)}) + \hat{\eta}^{(j)}(x) + \hat{\alpha}^{(j)}(x)\rho$$

Then on fold $\mathcal{I}^{(j+2)}_1$, we run a regression algorithm to obtain an estimator $\hat{h}^{(j)}$ for $\hat{h}^{(j)}(x) := \mathbb{E} [\hat{H}^{(j)}(x, Y(1)) \mid X = x, T = 1, \mathcal{I}^{(j+1)}_1]$, where we view $\hat{\alpha}^{(j)}$, $\hat{\eta}^{(j)}$ hence $\hat{H}^{(j)}$ as fixed functions. Finally, we define the estimator

$$\hat{\mu}_{1,0} = \frac{1}{|\mathcal{I}^{(j)}_1|} \sum_{i \in \mathcal{I}^{(j)}_1} \hat{r}^{(j)}(X_i)(\hat{H}^{(j)}(X_i, Y_i) - \hat{h}^{(j)}(X_i)) + \frac{1}{|\mathcal{I}^{(j)}_0|} \sum_{i \in \mathcal{I}^{(j)}_0} \hat{h}^{(j)}(X_i).$$

The above procedure is repeated for each $j = 1, 2, 3$, and we average the three estimators to obtain

$$\bar{\hat{\mu}}_{1,0} = \frac{1}{3} \sum_{j=1}^3 \hat{\mu}_{1,0}.$$
that is convex in $\theta = (\alpha, \eta)$. The empirical risk is correspondingly (recall that we run with fold $Z_1^{(j+1)}$)

$$
\hat{\mathbb{E}}_n[\ell(\theta, X, Y(1))] = \frac{1}{|Z_1^{(j+1)}|} \sum_{i \in Z_1^{(j+1)}} \ell(\theta, X_i, Y_i).
$$

We can thus consider a function class $\Theta$, and solve for the empirical risk minimization (ERM) problem. While the downstream inference is different, this approach is similar to Yadlowsky et al. (2018), where they express the bounds of conditional expectations of counterfactuals themselves as empirical risk minimizers. To be specific, we use the method of sieves (Geman and Hwang, 1982); we consider an increasing sequence $\Theta_1 \subset \Theta_2 \subset \cdots$ of spaces of smooth functions, and let

$$
\hat{\theta}^{(j)} = \arg\min_{\theta \in \Theta_n} \hat{\mathbb{E}}_n[\ell(\theta, X, Y(1))].
$$

We consider the following two examples inspired by Yadlowsky et al. (2018), where we also truncate the functions away from zero for $\alpha(x)$: we note that, if $\alpha^*(x)$ is always positive (implied by the minimality of the risk function) and continuous (satisfied if $\mathbb{P}_Y(Y) | X = x, T = 1$ is smooth in $x$) and $\mathcal{X}$ is a compact set, then there exists a positive $\epsilon > 0$ such that $\inf_{x \in \mathcal{X}} \alpha^*(x) \geq \epsilon$. In practice, we can set $\epsilon$ to be small enough, or let $\epsilon = \epsilon_n$ decays slowly to zero; this does not hurt the capability of function class or the convergence.

**Example 2** (Polynomials). Let $\text{Pol}(J)$ be the space of $J$-th order polynomials on $[0, 1]$:

$$
\text{Pol}(J, \epsilon) = \left\{ x \mapsto \sum_{k=0}^{J} a_k x^k : a_k \in \mathbb{R} \right\},
$$

and let $\text{Pol}(J, \epsilon)$ be the space of $J$-th order polynomials on $[0, 1]$ truncated at $\epsilon > 0$:

$$
\text{Pol}(J, \epsilon) = \left\{ x \mapsto \max\{\epsilon, \sum_{k=0}^{J} a_k x^k \} : a_k \in \mathbb{R} \right\}.
$$

Then we define the sieve $\Theta_n = \Theta_n^0 \times \Theta_n^0$, where $\Theta_n^0 = \{ x \mapsto \prod_{k=1}^{d} f_k(x_k) : f_k \in \text{Pol}(J_n, 0), k = 1, \ldots, d \}$ and $\Theta_n^0 = \{ x \mapsto \prod_{k=1}^{d} f_k(x_k) : f_k \in \text{Pol}(J_n, \epsilon), k = 1, \ldots, d \}$ for $J_n \to \infty$.

**Example 3** (Splines). Let $0 = t_0 < \cdots < t_{J+1} = 1$ be knots that satisfy $\frac{\max_{0 \leq j \leq J(t_{j+1} - t_j)}}{\min_{0 \leq j \leq J(t_{j+1} - t_j)}} \leq c$ for some $c > 0$. We define the space for $r$-th order splines with $J$ knots as

$$
\text{Spl}(r, J) = \left\{ x \mapsto \sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J} b_j (x - t_j)^{r-1} : a_k, b_k \in \mathbb{R} \right\}
$$

and the truncated space for $r$-th order splines with $J$ knots as

$$
\text{Spl}(r, J) = \left\{ x \mapsto \max\{\epsilon, \sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J} b_j (x - t_j)^{r-1} \} : a_k, b_k \in \mathbb{R} \right\}
$$

Then we define the sieve $\Theta_n = \Theta_n^0 \times \Theta_n^0$, where $\Theta_n^0 = \{ x \mapsto \prod_{k=1}^{d} f_k(x_k) : f_k \in \text{Spl}(J_n, 0), k = 1, \ldots, d \}$ and $\Theta_n^0 = \{ x \mapsto \prod_{k=1}^{d} f_k(x_k) : f_k \in \text{Spl}(J_n, \epsilon), k = 1, \ldots, d \}$ for $J_n \to \infty$.

We consider the classes of sufficiently smooth functions; for $p_1 = \lceil p \rceil - 1$ and $p_2 = p - p_1$, we define

$$
\Lambda_p^\alpha = \left\{ h \in C^{p_1}(\mathcal{X}) : \sup_{x \in \mathcal{X}} \left| D^\alpha h(x) \right| + \sup_{x \neq x' \in \mathcal{X} \atop \sum_{l=1}^{d} \alpha_l < p_1} \frac{|D^\beta h(x) - D^\beta h(x')|}{\|x - x'|^{p_2}} \leq c \right\}
$$

To ensure non-negativeness, we also define the truncated function class $\Lambda_p^\alpha(\mathcal{X}, \epsilon) := \{ x \mapsto \max\{f(x), \epsilon\} : f \in \Lambda_p^\alpha(\mathcal{X}) \}$, obtained by thresholding $\Lambda_p^\alpha(\mathcal{X})$ away from zero. To obtain convergence of the estimators, we impose the following assumptions on the true optimizer and regularity conditions of the loss function.
Assumption 1. Suppose \( \mathcal{X} = \prod_{k=1}^{d} \mathcal{X}_d \) is the Cartesian product of compact intervals, and \( \theta^* \in \Theta = \mathcal{L}_p(\mathcal{X}, \epsilon) \times \mathcal{L}_p(\mathcal{X}) \) for some \( c > 0 \). Suppose \( \mathbb{P}_X \mid T = 1 \) has positive density on \( \mathcal{X} \). We assume the function \( \mathbb{E} \ell(\theta(x), y) \mid X = x \) is \( \ell \)-strongly convex at \( (a, b) = \theta^*(x) \) for all \( x \in \mathcal{X} \). Also, \( |\ell(\theta, x, y) - \ell(\theta^*, x, y)| \leq \ell(x, y) - \theta^*(x) \) for \( \|\theta(x) - \theta^*(x)\|_2 < \epsilon \) for sufficiently small \( \epsilon > 0 \), where \( \|\cdot\|_2 \) is the Euclidean norm, and \( \sup_{x \in \mathcal{X}} \mathbb{E}[\|\ell(x, Y(x))\|_2^2 \mid X = x, T = 1] < M \) for some constant \( M > 0 \). Furthermore, there exists a constant \( C_1 \) such that \( \mathbb{E}[\ell(\theta, X, Y(1)) - \ell(\theta^*, X, Y(1)) \mid T = 1] \leq C_1 \|\theta - \theta^*\|_{L_2(P_{\mid T = 1})}^2 \) when \( \theta \in \mathcal{L}_p(\mathcal{X}) \) and \( \|\theta - \theta^*\|_{L_2(P_{\mid T = 1})} \) is sufficiently small.

We include a detailed discussion of Assumption 1 in Appendix B.2, where we provide concrete examples and justifications. In Assumption 1, we assume the true optimizer is sufficiently smooth, so that function approximator can learn it well. It can be satisfied if the conditional distribution \( \mathbb{P}_{Y(1) \mid X, T} \) is sufficiently “smooth” in \( x \). We require the strong convexity of the conditional risk function at its minimizer \( \theta^*(x) \); it is typically the case if \( Y(1) \) is not deterministic given \( X \). The stability condition at \( \theta^*(x) \) can be satisfied if \( Y \) is not heavy-tailed. We also assume that the population risk is stable in terms of \( L_2(P_{\mid T = 1}) \) norm of \( \theta \), which can be satisfied if \( \mathbb{E}[\ell(\theta, x, y) \mid X = x, T = 1] \) is smooth or have Lipschitz derivatives.

Under the above regularity conditions, we obtain convergence rates of the empirical risk minimizers \((\hat{\alpha}^{(j)}, \hat{\eta}^{(j)})\). The proof of the following theorem is in Appendix C.2.

Theorem 1. Suppose Assumption 1 holds. We set \( J_n = \left( \frac{\log n}{n} \right)^{1/(2p+d)} \) for the sieve estimators in Examples 2 and 3, and suppose \( \hat{\theta}^{(j)} \) satisfies \( \mathbb{E}_n[\ell(\hat{\theta}^{(j)}, X, Y(1))] \leq \inf_{\theta \in \Theta} \mathbb{E}_n[\ell(\theta, X, Y(1))] - O_P(\left( \frac{\log n}{n} \right)^p/(2p+d)) \).

Then, under the above regularity conditions on the loss function, the empirical risk minimizer \( \hat{\alpha}^{(j)}, \hat{\eta}^{(j)} \) converges to the truth. Besides the examples and guarantees we provide, similar results might be obtained for other function classes like wavelets (Daubechies, 1992), and the conditions in Assumption 1 might be weakened or modified to account for more generality. Such extension is beyond the scope of this work.

4 Theoretical Guarantees

In this section, we provide the theoretical guarantees of our procedure in Section 3.3. We first show the double robustness and one-side validity regarding consistency. We then present inferential guarantees: we achieve Wald-type inference for \( \mu_{-0} \) under slower-than-parametric convergence rates of the nuisance component estimation; moreover, even when the empirical risk minimization is not consistent to the optimum, our inference procedure can still be valid. Finally, we show how to leverage our procedure to construct bounds for treatment effects.

4.1 Double consistency and one-side validity

Our first result is regarding the consistency of our estimator: we show that \( \hat{\mu}_{-0} \) is doubly robust to nuisance estimation. Even more interestingly, our estimator is robust to the ERM step: given that either \( \hat{\sigma}^{(j)} \) or \( \hat{\beta}^{(j)} \) is consistent, it converges to the true bound if the ERM step is consistent; otherwise, our estimator converges to a conservative but still valid lower bound of \( \mu_{-0} \). We call this “one-side validity”.

We start with a mild assumption on the convergence of ERM step, note that we do not assume the convergence to the true minimizer \((\alpha^*, \eta^*)\). For notational convenience, we denote the risk function \( \ell(\theta, x, y) = a\theta^*(x) + b + a\theta \) for \( \theta = (a, b) \) as in Proposition 3. When there is no confusion, equivalently use \( \ell(\theta, x, y) = \ell((\alpha, \eta)(x), y) \) when \( \theta = (\alpha, \eta) \) is a function.

Assumption 2. Suppose for each \( j \), the empirical optimizer \((\hat{\alpha}^{(j)}, \hat{\eta}^{(j)})\) converges in sup-norm to some \((\alpha^*, \eta^*)\) such that for all \( x \in \mathcal{X} \), \( |\ell(\theta^*(x), y) - \ell((a, b), x, y)| \leq M(x, y) \|\theta^*(x) - (a, b)\|_2 \) for all \( \|\theta^*(x) - (a, b)\|_2 \leq \epsilon \) for some constant \( \epsilon > 0 \), and \( \mathbb{E}[M(x, y)|T = 1] \leq M \) for some constant \( M > 0 \). Assume that \( \hat{\sigma}^{(j)} \) are uniformly bounded, and \( \hat{H}^{(j)}, \hat{\beta}^{(j)} \) have uniformly bounded second moments.
In Assumption 2, we additionally assume a mild regularity condition on the first-order expansion at the limit; it is satisfied if the loss function $\ell$ is differentiable or locally Lipschitz. The second moment condition is also mild and standard. The following theorem shows the double robustness as well as one-side validity of our estimator, whose proof is in Appendix C.3.

**Theorem 2.** Suppose Assumption 2 holds for some fixed $\theta^0 = (\alpha^0, \eta^0)$. Assume either (i) $\|\hat{r}^{(j)} - r_{1,0}\|_{L_2(P_X | T = 1)} = o_P(1)$ or (ii) $\|\hat{h}^{(j)} - \hat{h}(\cdot)\|_{L_2(P_X | T = 1)} = o_P(1)$. Then the following holds: if $\theta^0 = \theta^*$, i.e., the ERM step is consistent, then $\hat{\mu}_{1,0} = \tilde{\mu}_{1,0} + o_P(1)$; otherwise, $\mu_{1,0} = \mu_{1,0}^* + o_P(1)$ for some constant $\mu_{1,0}^* \leq \hat{\mu}_{1,0}$.

A similar property of one-side validity has been documented by a recent work of Dorn et al. (2021), where they work under the marginal sensitivity model of Tan (2006) and develop a similar property based on an exact characterization of the worst-case scenario. However, in our setting, the one-side validity is a relatively straightforward consequence of the interplay between primal and dual problems. It would be interesting to find connections between our results; for example, whether their result can also be implied by the duality.

### 4.2 Wald-type inference for $\mu_{1,0}$

We now turn to inferential guarantees. We show that our procedure yields valid Wald-type inference under slow convergence rates of nuisance estimations. We begin with some regularity conditions on the risk function.

**Assumption 3.** Let $\theta^* = (\alpha^*, \eta^*)$ be the minimizer in (9). Suppose $\mathbb{E}[\nabla_{a,b}\ell(a, b, x, X(1)) | X = x, T = 1] = \nabla_{a,b}\ell(a, b, x, X(1))$ at $(a, b) = (\alpha^*(x), \eta^*(x))$ for $P_X | T = 1$-almost all $x$. Suppose $|\ell(a, b, x, y) - \ell(\theta^*, x, y) - \nabla_{a,b}\ell(\alpha^*(x), \eta^*(x), x, y)[\alpha^*(x) - a, \eta^*(x) - b]| \leq M(x, y)\|\alpha^*(x) - a, \eta^*(x) - b\|_2^2$ for some $(a, b)$ in some neighborhood of $(\alpha^*(x), \eta^*(x))$, where $\mathbb{E}[M(x, Y(1)) | X = x, T = 1] \leq M$ for some constant $M > 0$ for all $x \in X$. Furthermore, $\|\ell(\theta, X, Y(1)) - \ell(\theta^*, X, Y(1))\|_{L_2(P_X | T = 1)} = O(\|\theta - \theta^*\|_{L_2(P_X | T = 1)})$ for function $\theta$ in a small $L_2(P_{X | T = 1})$-neighborhood of $\theta^*$.

In Assumption 3, we require the risk function to be differentiable and admits a Taylor expansion near some optimizer, as well as a regularity condition on the exchangeability of differentiation and conditional expectation. These are mild conditions that are commonly adopted in the literature (Van der Vaart, 2000). The risk function is assumed to be stable, so that plugging in estimators of $\alpha^*, \eta^*$ won’t cause large errors, which is also a mild condition that can be satisfied under a first-order Taylor expansion condition.

We assume the following convergence rates, where we assume the ERM step is $o_P(n^{-1/4})$ consistent, and the nuisance estimation error of $\hat{\sigma}^{(j)}$ and $\hat{h}^{(j)}$ has a product of order $o_P(n^{-1/2})$.

**Assumption 4.** Suppose for each $j$, $\|\hat{r}^{(j)} - r_{1,0}\|_{L_2(P_X | T = 1)}$, $\|\hat{h}(\cdot) - \hat{h}^{(j)}\|_{L_2(P_X | T = 1)} = o_P(n^{-1/2})$, and $\|\hat{\sigma}^{(j)} - \sigma^{(j)}\|_{L_2(P_X | T = 1)} = o_P(n^{-1/4})$ for some optimizer $(\alpha^*(x), \eta^*(x))$ of (9) satisfying Assumption 3.

In Assumption 4, the rate of $\hat{r}^{(j)}$ depends on the estimation of $e(x) = P(T = 1 | X = x)$, a standard regression problem. The estimation of $\hat{h}^{(j)}$ is also a regression problem viewing $\hat{H}^{(j)}$ as fixed. Convergence rate guarantees for such conditional mean estimation problems are well-established in the literature (Stone, 1982; Mallat, 1999; Pagan et al., 1999; Shen and Wong, 1994; Wasserman, 2006; Simonoff, 2012). The estimation of $(\hat{\alpha}^{(j)}(\cdot), \hat{\eta}^{(j)}(\cdot))$ has been discussed in Section 3.3.

Under the above two assumptions, we show that our estimator is asymptotically normal, where the estimation error of nuisance component is negligible. The proof of Theorem 3 is deferred to Appendix C.4.

**Theorem 3.** Suppose Assumptions 3 and 4 hold. Then $\sqrt{n}(\hat{\mu}_{1,0} - \mu_{1,0}) \rightsquigarrow N(0, \text{Var}(\phi_{1,0}(X, Y, T)))$, where $
abla_{2_1} \phi_{1,0}(X_i, Y_i, T_i) = \frac{T_i}{p_1} r_{1,0}(X_i) [H(X_i, Y_i(1)) - h(X_i)] + \frac{1 - T_i}{p_0} h(X_i)$.

Here $p_1 = P(T = 1) = 1 - p_0$, and we define $H(x, y) = \ell(\theta^*, x, y), h(x) = \mathbb{E}[H(X, Y(1)) | X = x, T = 1]$. All the expectations (variances) are induced by the observed distribution. Furthermore, define

$$\hat{\sigma}^2 = \frac{1}{p_1} \left( \frac{1}{n_1} \sum_{i \in S_1} d_{2, i}^2 - \left( \frac{1}{n_1} \sum_{i \in S_1} d_{1, i} \right)^2 \right) + \frac{1}{p_0} \left( \frac{1}{n_0} \sum_{i \in S_0} d_{0, i}^2 - \left( \frac{1}{n_0} \sum_{i \in S_0} d_{0, i} \right)^2 \right)$$
where \( \hat{p}_i = |Z_i|/n \), \( \hat{p}_0 = |Z_0|/n \), \( d_{1,i} = \hat{r}^{(j[i])}(X_i)(\hat{H}^{(j[i])}(X_i,Y_i) - \hat{h}^{(j[i])}(X_i)) \), \( d_{0,i} = \hat{h}^{(j[i])}(X_i) \), and \( j[i] \in \{1,2,3\} \) is the fold that sample \( i \) lies in. Then \( \sqrt{n}(\hat{\nu}_{1,0} - \nu_{1,0})/\hat{\sigma} \rightsquigarrow N(0,1) \).

Similar results can be obtained for \( \mu_{1,0}^+ \), if we simply flip the sign of \( Y(1) \) and flip back after running the same procedure. The above procedure can also be generalized to the inference of \( \mu_{1,1}^+ \); the simplest way might be just switching the two groups. We summarize these results in Appendix B.3 for completeness.

4.3 Robustness to misspecification of ERM

Our inferential guarantee in Theorem 3 relies on consistency of both the nonparametric regression and the ERM steps. While these conditions are relatively mild, in this part, we take a step further and note that our estimator is in particular robust to the ERM step.

The following theorem shows that even though our empirical risk minimizers \( \hat{\alpha}^{(j)} \) and \( \hat{\eta}^{(j)} \) converge to something else, our procedure still provide valid, albeit more conservative, inference on the lower bound of \( \mathbb{E}[Y(1)|T=0] \). The proof of Theorem 4 is in Appendix C.5.

**Theorem 4.** Suppose Assumptions 3 and 4 with \( (\alpha^*,\eta^*) \) replaced by some fixed \( \theta^o := (\alpha^o,\eta^o) \), and the first condition of Assumption 3 is replaced by the local one: \( \mathbb{E}[\ell(\alpha^o(X),\eta^o(X),X,Y(1))|\alpha^o(X) - \alpha(X),\eta^o(X) - \eta(X)) = 0 \) for any \( (\alpha,\eta) \in \Theta_n \) in a small \( \| \cdot \|_\infty \)-neighborhood of \( \theta^o \). We additionally assume \( \| \hat{\theta} - \theta^o \|_\infty = o_P(1) \). Then \( \sqrt{n}(\hat{\mu}_{1,0} - \mu_{1,0}^O) \rightsquigarrow N(0,\text{Var}(\phi_{1,-}(X,Y,T))) \), where \( \mu_{1,0}^O \leq \mu_{1,0}^\beta \), and

\[
\phi_{1,-}^O(X_i,Y_i,T_i) = \frac{T_i}{p_1} r_{1,0}(X_i) [H^O(X_i,Y_i(1)) - h^O(X_i)] + \frac{1 - T_i}{p_0} h^O(X_i) .
\]

Here we define \( H^O(x,y) = \ell(\theta^O,x,y) \) and \( h^O(x) = \mathbb{E}[H^O(X,Y(1))|X=x,T=1] \). Furthermore, we have \( \sqrt{n}(\hat{\mu}_{1,0} - \mu_{1,0}^O)/\hat{\sigma} \rightsquigarrow N(0,1) \) for the variance estimator \( \hat{\sigma}^2 \) defined in Theorem 3.

In theorem 4, we only require the convergence of \( (\hat{\alpha}^{(j)},\hat{\eta}^{(j)}) \) in \( L_2(\mathbb{P}_{\theta=1}) \)-norm any pair of fixed functions. This might happen, for example, if the function class we employ does not approximate \( (\alpha^*,\eta^*) \) very well, but our estimators still converge to a fixed in-class risk minimizer. In this case, our estimator converges to a conservative lower bound of the counterfactual mean and still yields valid inference.

In parallel to the mean-zero gradient property of \( \theta^* \), we assume a local first-order condition for \( \theta^O \) restricted to \( \Theta_n \), which is crucial for the double robustness to the estimation error. This condition is satisfied as long as \( \theta^O \) is the population risk minimizer (with weight \( r(X) \)) among \( \Theta_n \). To obtain an estimator that converges to \( \theta^O \), we might slightly change the procedure: fit \( \hat{r}^{(j)}(x) \) on one fold and run the ERM with the fitted \( \hat{r}^{(j)}(x) \) on a new fold. The convergence of the empirical risk minimizer can be satisfied if \( \Theta_n \) is not too complex and \( \hat{r}^{(j)}(x) \) is consistent with a slow rate.

We also note that, as implied by Theorem 4, plugging in any fixed function into our procedure without ERM (or equivalently, setting \( \Theta_n = \{\theta\} \) for some fixed \( \theta \) that satisfy the regularity conditions) also yields a valid lower bound. However, this is uninteresting as it may be way too conservative.

4.4 Inference for treatment effects

With the above estimator for counterfactual means in place, we briefly discuss the construction of confidence intervals for treatment effects. Let us first start with ATT/ATC. Following the preceding example of \( \mu_{1,0}^O \), in view of (8), we can construct an estimator for the lower bound of ATC, defined as

\[
\tilde{\tau}_{ATC} := \hat{\mu}_{1,0}^O - \frac{1}{n_0} \sum_{i \in I_0} Y_i ,
\]

where \( \hat{\mu}_{1,0}^O \) is constructed as in Section 3.3. Theorem 3 directly implies the following result of double robustness and asymptotic normality for \( \tilde{\tau}_{ATC} \), and the proof is omitted for brevity.

**Corollary 1.** Under the same conditions of Theorem 3, \( \sqrt{n}(\tilde{\tau}_{ATC} - \tau_{ATC}) \rightsquigarrow N(0,\text{Var}(\phi_{ATC}(X_i,Y_i,T_i))) \), where \( \tau_{ATC} = \mu_{1,0} - \mathbb{E}[Y(0)|T=0] \) is a lower bound for ATC under the \((f,\rho)\)-selection condition, and

\[
\phi_{ATC}(X_i,Y_i,T_i) = \frac{T_i}{p_1} r_{1,0}(X_i) [H(X_i,Y_i(1)) - h(X_i)] + \frac{1 - T_i}{p_0} (h(X_i) + Y_i(0)) .
\]
Similar to Theorem 3, a consistent estimator \( \hat{\sigma}_{\text{ATE}}^2 \) can also be constructed for \( \text{Var}(\phi_{\text{ATE}}(X_i, Y_i, T_i)) \), enabling Wald-type inference. Based on the results of Section 4, we can similarly construct other bounds for ATT and ATC and combine them to obtain bounds on ATE. For example, let \( \hat{\rho}_{0,1}^+ \) estimate an upper bound on \( E[Y(0) \mid T = 1] \) with influence function \( \phi_{0,1}^+ \) (see Appendix B.3 for details). We may construct

\[
\tilde{\tau}_{\text{ATT}} := \frac{1}{n_1} \sum_{i \in I_1} Y_i - \hat{\rho}_{0,1}^+, \quad \text{and} \quad \tilde{\tau}_{\text{ATE}} := \hat{\rho}_1 \cdot \tilde{\tau}_{\text{ATT}} + \hat{\rho}_0 \cdot \tau_{\text{ATT}}.
\]

Then \( \sqrt{n}(\tilde{\tau}_{\text{ATE}} - \tau_{\text{ATE}}) \sim N(0, \text{Var}(\phi_{\text{ATE}}(X_i, Y_i, T_i))) \), where \( \tau_{\text{ATE}} \) is a lower bound for ATE under the \((f, \rho)\)-selection condition, and the influence functions are \( \phi_{\text{ATE}} = p_1 \phi_{\text{ATT}} + p_0 \phi_{\text{ATC}} \), and \( \phi_{\text{ATT}} = T_i Y_i / p_1 - \phi_{0,1}^+ \).

5 Numerical Experiments

We illustrate the performance of our procedure on simulated datasets. We focus on the estimation of the counterfactual mean \( E[Y(1) \mid T = 0] \) given confounded observational data and take \( f(t) = t \log t \).

5.1 Simulation setting

We fix the sample size at \( n = 15000 \) and the covariate dimension at \( p = 4 \). To generate the confounded dataset, setting \( U = Y(1) \), we fix the observed distribution of \( \mathbb{P}(X, Y(1) \mid T = 1) \) and \( \mathbb{P}(T = 1) \), and vary the counterfactual distribution \( \mathbb{P}(X, Y(1) \mid T = 0) \), so that \( \text{OR}(x, u) := \frac{\mathbb{P}(T = 0 \mid X = x, U = u) \mathbb{P}(T = 1 \mid X = x)}{\mathbb{P}(T = 1 \mid X = x, U = u) \mathbb{P}(T = 0 \mid X = x)} \) satisfies \((f, \rho)\)-selection condition for a sequence of \( \rho > 0 \). To be specific, we generate the covariates and treatment assignments with

\[
X \sim \text{Unif}([0,1]^p), \quad T \mid X \sim \text{Bern}(e(X)),
\]

where we set the observed propensity score as \( e(x) = \logit(\gamma^T x) \) for \( \gamma = (-0.531, 0.126, -0.312, 0.018)^T \). Finally, given \( \delta \in \mathbb{R} \), we generate the potential outcomes via

\[
Y(1) = X^\top \beta_1 - \delta \cdot (1 - T) \sigma(X) + \varepsilon \cdot \sigma(X),
\]

\[
Y(0) = X^\top \beta_0 - \delta \cdot (1 - T) \sigma(X) + \varepsilon \cdot \sigma(X),
\]

where \( \varepsilon \sim N(0, 1) \), and we set \( \beta_1 = (0.531, 1.126, -0.312, 0.671)^T \), \( \beta_0 = (-0.531, -0.126, -0.312, 0.671)^T \) and \( \sigma^2(x) = 1 + 1.25x_1^2 \). Put it another way, the observations of \( Y(1) \) in the treated group follow \( Y(1) \mid X = x, T = 1 \sim N(x^\top \beta_1, \sigma^2(x)) \), while \( Y(1) \mid X = x, T = 0 \sim N(x^\top \beta_1 - \delta \cdot \sigma(x), \sigma^2(x)) \).

In this setting, the confounder is entirely driven by \( U := Y(1) \). The odds ratio is

\[
\text{OR}(x, u) = \exp \left( -\frac{\delta(u - x^\top \beta_1) + \delta^2}{2\sigma^2(x)} \right),
\]

and we obtain an upper bound for the \( f \)-divergence as \( \rho = \delta^2 / 2 \). The same bound can be obtained for the other odds ratio of the control group. The observed dataset is thus \( \{(X_i, Y_i, T_i)\}_{i=1}^n \), where \( Y_i = Y_i(T_i) \).

Intuitively, \( \delta \) drives the direction and magnitude of confounding: when \( \delta > 0 \), larger values of \( Y(1) \) has larger probability of getting treated even conditional on \( X \); as a result, the observed \( Y(1) \) in the treated group is actually shifted to larger values, leading to overestimate of treatment effects if confounding is not accounted for. The larger \( \delta \) is, the more severe the impact of confounding is. On the other hand, when \( \delta < 0 \), inference under the strong ignorability assumption tends to underestimate the treatment effects. In this setting, although we anticipate \( \text{OR}(X, U) \) to be controlled overall, it does not admit a uniform upper bound; we plot several quantiles of \( \text{OR}(X, U) \) in the treated group in Figure 2.
5.2 Sensitivity analysis with one dataset

We first illustrate the estimators and confidence intervals we obtain under a fixed confounded data generating process. To be specific, we fix $\delta = 0.5$ (hence $\rho = 0.125$) to generate the data, and apply our procedure to the fixed dataset for a series of $\rho \in \{0.05, 0.1, \ldots, 0.95, 1.0\}$. We obtain 0.975-lower confidence bound (LCB) for the lower bound of ATC and 0.975-upper confidence bounds (UCB) for the upper bound of ATC (i.e., the bounds in (7)), which together form a 0.95-CI for ATC under a hypothesized confounding level $\rho$. The results are plotted in Figure 3.

Without accounting for confounding, reweighting on the covariates tend to overestimate the ATC (indicated by the estimators for small $\rho$). The LCB crosses the ground truth at $\hat{\rho} = 0.1$; this can be viewed as a lower confidence bound for the true confounding level $\rho = 0.125$ (we elaborate on this in the discussion when the ground truth is zero). Finally, the LCB hits zero at $\hat{\rho}_0 = 0.65$; we can thus conclude with 0.95-confidence that ATC is non-negative as long as the true confounding level does not exceed $\hat{\rho}_0$.

5.3 Validity and sharpness

To show the validity and sharpness of our procedure, we first vary $\delta \in \{0.1, 0.2, \ldots, 1.5\}$ in our data-generating process, and apply our procedure with the correct level $\rho = \delta^2/2$. Feeding the data into Algorithm 1 yields the estimator $\hat{\mu}_{1,0}$ for the lower bound on $\mathbb{E}[Y(1) \mid T = 1]$; changing the observations to $Y(1) \leftarrow -Y(1)$, the negative of the output of Algorithm 1, denoted as $\hat{\mu}_{1,0}^+$, is an estimator for the upper bound $\mu_{1,0}^+$. Based on the corresponding variance estimators $\hat{\sigma}_{1,0,\pm}$, we construct the confidence interval for $\mathbb{E}[Y(1) \mid T = 0]$ as $\text{CI}_{\text{mean}} := [\hat{\mu}_{1,0} + z_{0.025} \hat{\sigma}_{1,0,0}/\sqrt{n}, \hat{\mu}_{1,0}^+ + z_{0.975} \hat{\sigma}_{1,0,0}/\sqrt{n}]$; the confidence interval for $\mu_{1,0}$ is constructed as $\text{CI}_{\text{lower}} := [\hat{\mu}_{1,0} + z_{0.025} \hat{\sigma}_{1,0,0}/\sqrt{n}, \hat{\mu}_{1,0}^+ + z_{0.975} \hat{\sigma}_{1,0,0}/\sqrt{n}]$, and similarly $\text{CI}_{\text{upper}} := [\hat{\mu}_{1,0} + z_{0.025} \hat{\sigma}_{1,0,0}/\sqrt{n}, \hat{\mu}_{1,0}^+ + z_{0.975} \hat{\sigma}_{1,0,0}/\sqrt{n}]$ for $\mu_{1,0}$.

To obtain the ground truth of $\mu_{1,0}^+$ at each $\delta$, we evaluate the bounds on $\mathbb{P}[Y(1) \mid X = x, T = 0]$ for each $x$ by optimizing with a huge amount of samples from $\mathbb{P}[Y(1) \mid X = x, T = 1]$; we then marginalize over $X \mid T = 0$.

\footnote{This is feasible because in our setting, $\mathbb{P}[Y(1) \mid X = x, T = 1]$ is normal distribution, and the target bounds are shift-invariant; we only need to evaluate the bounds for all values of $\rho$ and a fine grid of $\sigma(x)$.}
Figure 3: The 0.975-LCB for the lower bound of ATC (blue) and 0.975-UCB for the upper bound of ATC (red), obtained from one run of our procedure on one dataset. Solid lines are the original estimators from our procedure, while dashed lines are sorted to ensure they are monotone in $\rho$. The black dashed line is the actual ATC.

to obtain an estimator the ground truth of $\mu_{1,0}$. For each $\delta$, this procedure is repeated and averaged over many runs to further reduce the random error.

The estimators for bounds of counterfactuals over $N = 500$ runs for each $\rho$ are plotted in Figure 4 (they are evenly spaced on the $x$-axis). The simulation results show the sharpness and accuracy of our estimators: they are quite close to the ground truth, especially for small values of $\rho$; they get a bit conservative and have a larger variance when $\rho$ is as large as 1. Interestingly, there are also a few outliers when $\rho$ is very small, and the estimators seem to be the most stable for an medium scale of $\rho$ (around 0.18 to 0.5). The actual value of $E[Y(1) \mid T = 0]$ in our design, represented by the red triangles, are very close to the lower solid line, the ground truth of $\mu_{1,0}$; this means our simulation design is close to the worst case.

To further validate our inference procedure, we compute the empirical coverage of CI$_{\text{lower}}$ and CI$_{\text{upper}}$ for $\mu_{1,0}^\pm$ over $N = 500$ runs. We also compute the ground truth of $E[Y(1) \mid T = 0]$ under our design as a baseline, and compute the empirical coverage of CI$_{\text{mean}}$. They are plotted in Figure 5. Our empirical coverage is close to the nominal level 0.95 in almost all settings, showing the validity of our inference procedure.

Figure 5: Empirical coverage for $\mu_{1,0}^-$ (left), $\mu_{1,0}^+$ (middle), and $E[Y(1) \mid T = 0]$ (right). The short vertical segments are the C.I.s computed with $N = 500$ replicates. The red dashed line is the nominal level 0.95.

Figure 6 plots the empirical coverage of one-sided C.I.s for $\mu_{1,0}^\pm$, defined as CI$_{\text{lower one-side}} := [\hat{\mu}_{1,0}^- + z_{0.05} \hat{\sigma}_{1,0}^- / \sqrt{n}, +\infty)$ and CI$_{\text{upper one-side}} := (-\infty, \hat{\mu}_{1,0}^+ + z_{0.95} \hat{\sigma}_{1,0}^+ / \sqrt{n}]$. Our theory shows that even though the ERM is off, these C.I.s still have valid asymptotic coverage; such robustness is also supported by empirical evidence.
Figure 4: Boxplots for $\tilde{\mu}_{1,0}$ (red ones) and $\tilde{\mu}_{1,0}$ (blue ones) over $N = 500$ replicates with each value of $\rho$. The solid lines are the ground truths of $\mu_{1,0}^+$ and $\mu_{1,0}^-$. The red triangles represent the actual value of $\mathbb{E}[Y(1) | T = 0]$ in our simulation setting.

Figure 6: Empirical coverage of one-sided C.I.s for $\mu_{1,0}$ (left), and $\mu_{1,0}^+$ (right). The short vertical segments are the C.I.s computed with $N = 500$ replicates. The red dashed line is the nominal level 0.95.

6 Discussion

In this work, we propose a new sensitivity model based on the $f$-divergence that characterizes the average effect of confounders on selection bias. Under the $f$-sensitivity model, we offer a scheme for the estimation and inference on the counterfactual and the ATE. We close the paper by a discussion on possible extensions.

Tightness. As mentioned before, the optimal value of (6) is not necessarily the tightest lower bound for $\mathbb{E}[Y(1) | T = 0]$: the sharp one under $(f, \rho)$-selection condition is given by

$$\inf\left\{\mathbb{E}^{\sup}\left[Y(1) | T = 0\right] : \mathbb{P}^{\sup} \in Q_{1,0}\right\},$$
where $Q_{1,0}$ is the identification set of all distributions that agree with the observed distribution and satisfy the $(f, \rho)$-selection condition. The constraints in (6) define a superset of $Q_{1,0}$, potentially leading to conservativeness.

Using the exact characterization of $Q_{1,0}$ provided in Proposition 5, we can represent the sharp lower (resp. upper) bound of $\mathbb{E}[Y(1) \mid T = 0]$ under the $(f, \rho)$-selection condition as the optimal value of
\[
\min_{\text{L(x) measurable}} (\text{resp. max}) \mathbb{E}[Y(1)L(X) \mid T = 1]
\]
\[
\text{s.t. } \mathbb{E}[L(x) \mid X = x, T = 1] = r_{1,0}(x)
\]
\[
\mathbb{E}[f(L(x)/r_{1,0}(x)) \mid X = x, T = 1] \leq \rho, \text{ for almost all } x.
\]
\[
\mathbb{E}[r_{1,0}(x)f(r_{1,0}(x)/L(x)) \mid X = x, T = 1] \leq \rho, \text{ for almost all } x.
\]

With the same argument, we can also develop the optimization problems for sharp bounds on the ATT and the ATC. Compared to the dual problems in Proposition 3, the additional constraints in the last line above leads to a dual form that is not as clean. Developing an efficient algorithm that solves this tight bound remains an interesting avenue for future research.

**Sensitivity analysis.** In this paper, we have focused on conducting inference on the counterfactuals and treatment effects under the $(f, \rho)$-selection condition, with a prescribed confounding parameter $\rho$. Based on this, we can make robust causal conclusions and conduct sensitivity analysis by inverting the confidence intervals as follows. Suppose the goal is to detect if there is a nonzero ATE; we can consider an increasing sequence of $\rho$, and construct a level $1 - \alpha$ confidence interval $\hat{C}(\rho)$ for the ATE using the method introduced in this paper at each value of $\rho$; finally let $\hat{\rho}$ be the smallest $\rho$ such that $\hat{C}(\rho)$ contains zero. We can interpret the results as either there is a nonzero ATE, or there is a confounder as large as $\hat{\rho}$ to explain away the observed treatment effects.

More rigorously, let $\rho^*$ denote the true confounding level and suppose the constructed confidence intervals $\hat{C}(\rho)$ are nested in $\rho$: for any $\rho_1 \leq \rho_2$, $\hat{C}(\rho_1) \subset \hat{C}(\rho_2)$. We then have
\[
\limsup_{n \to \infty} \mathbb{P}(\text{ATE} = 0, \rho^* < \hat{\rho}) \leq \limsup_{n \to \infty} \mathbb{P}(\text{ATE} \notin \hat{C}(\rho^*)) \leq \alpha,
\]
if $\hat{C}(\rho^*)$ is an asymptotically valid confidence interval for the ATE. In words, when the ATE is indeed zero, $\hat{\rho}$ is an asymptotic level-$(1 - \alpha)$ confidence lower bound for $\rho^*$. Similar to the case of Jin et al. (2021), here only point-wise validity is necessary, i.e., we only need our CIs to be asymptotically valid for each fixed ground truth of $\rho$. Finally, we note that the monotonicity of the confidence intervals is satisfied with a reasonable estimation procedure; one can also force the confidence intervals to be monotone by enlarging some of them to conform to those for smaller values of $\rho$, without hurting the asymptotic validity.

**Implications for the conditional average treatment effect (CATE).** The methodology proposed in this paper also provides bounds on CATE under the $(f, \rho)$-selection condition. For example, the proof of Propositions 2 and 3 implies that a lower bound for $\mathbb{E}[Y(1) \mid X = x, T = 1]$ is given by the optimal value of
\[
\min_{\text{L(x) measurable}} \mathbb{E}[Y(1)L \mid X = x, T = 1]
\]
\[
\text{s.t. } \mathbb{E}[L(\mid T = 1, X = x],
\]
\[
\mathbb{E}\left[f\left(\frac{L}{r_{1,0}(x)}\right) \mid X = x, T = 1\right] \leq \rho.
\]

The dual form of the above optimization problem is
\[
\sup_{\alpha \geq 0, \eta \in \mathbb{R}} -r_{1,0}(x) \cdot \mathbb{E}\left[\alpha f^*\left(-\frac{Y(1) + \eta}{\alpha}\right) + \eta + \alpha\rho \mid X = x, T = 1\right].
\]
(11)

Note that the optimizer $(\alpha^*(x), \eta^*(x))$ defined in (10) is exactly the optimizer of (11). In fact, $\hat{\rho}(x) := \hat{\tau}(\cdot)h(\cdot)(x)$ where $\hat{\tau}(\cdot), h(\cdot)$ are defined in Algorithm 1 is an estimator for the optimal objective in (11).
These quantities are repeatedly estimated on distinct folds of data as intermediate steps of our procedure. While such sample splitting does not compromise the efficiency of inference due to the final averaging step, how to efficiently estimate these CATE bound functions with statistical guarantees might call for distinct considerations from ours. We leave this for future investigation.

**Marginal \((f, \rho)\) selection condition.** We might even relax the per-\(x\) uniform bound on the \(f\)-divergence in Definition 1 to a marginal fashion, so that the selection bias is controlled averaged over both \(U\) and \(X\). More formally, we might consider the constraint that

\[
\int f \left( \frac{\Pr(T = 0 \mid X = x, U)}{\Pr(T = 1 \mid X = x, U)} \right) \Pr(U, X \mid T = 1) \, dP U, X \mid T = 1 \leq \rho.
\]

In this setting, the odds ratio can be very large for a small proportion of \(X \mid T = 1\), but still controlled in the average sense. This type of marginal \((f, \rho)\)-selection model leads to a larger class of distributional shifts than the \((f, \rho)\)-selection condition here, and a different optimization problem for bounds on counterfactual means. Following similar arguments here, we see that the dual formulation, parallel to Proposition 3, can still be viewed as a risk minimization problem; however, the risk function would involve the unknown \(X\)-shift \(r_1, 0\), which might make the estimation and inference more complicated. The estimation and inference under this marginal \(f\)-sensitivity model is an ongoing work.

### 7 Acknowledgement

The authors thank Emmanuel Candès, Kevin Guo and Dominik Rothenhäusler for helpful discussions. Z. R. was partially supported by ONR grant N00014-20-1-2337, and NIH grants R56HG010812, R01MH113078 and R01MH123157.
References

Andrews, I., Gentzkow, M., and Shapiro, J. M. (2020). On the informativeness of descriptive statistics for structural estimates. *Econometrica*, 88(6):2231–2258.

Bross, I. D. (1966). Spurious effects from an extraneous variable. *Journal of chronic diseases*, 19(6):637–647.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Christensen, T. and Connault, B. (2019). Counterfactual sensitivity and robustness. *arXiv preprint arXiv:1904.00989*.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Cox, D. R. (1958). Planning of experiments.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464.

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368.

Dorn, J. and Guo, K. (2021). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *arXiv preprint arXiv:2102.04543*.

Dorn, J., Guo, K., and Kallus, N. (2021). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):907–920.

Geer, S. A., van de Geer, S., and Williams, D. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pages 401–414.

Gupta, S. and Rothenhäusler, D. (2021). The s-value: evaluating stability with respect to distributional shifts. *arXiv preprint arXiv:2105.03067*.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.
Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jin, Y., Ren, Z., and Candès, E. J. (2021). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *arXiv preprint arXiv:2111.12161*.

Jin, Y. and Rothenhäusler, D. (2021). One estimator, many estimands: fine-grained quantification of uncertainty using conditional inference. *arXiv preprint arXiv:2104.04565*.

Kallus, N., Mao, X., and Zhou, A. (2019). Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lee, K., Bargagli-Stoffi, F. J., and Dominici, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*.

Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412.

Lukenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Morimoto, T. (1963). Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331.

Nie, X., Imbens, G., and Wager, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

Pagan, A., Ullah, A., Gourieroux, C., Phillips, P. C., and Wickens, M. (1999). *Nonparametric econometrics*, volume 10. Citeseer.

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. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rahman, S. (2016). The f-sensitivity index. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):130–162.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Rosenbaum, P. (2002a). *Observational Studies*. Springer Series in Statistics. Springer.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2002b). Attributing effects to treatment in matched observational studies. *Journal of the American statistical Association*, 97(457):183–192.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rudin, W. et al. (1976). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2020). Distributional robust batch contextual bandits. *arXiv preprint arXiv:2006.05630*.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Timan, A. F. (2014). *Theory of approximation of functions of a real variable*. Elsevier.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Luxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2017). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *arXiv preprint arXiv:1711.11286*.

Zheng, W. and Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer.
A Proofs for identification sets

A.1 Proof of Proposition 1

Here we prove a stronger result that directly implies Proposition 1. The following proposition is a tight characterization of the identification set induced by the $f$-sensitivity models.

Proposition 5. Let $(X,Y(0),Y(1),U,T) \sim \mathbb{P}_{\text{sup}}$ be the true unknown super-population over all random variables of interest. Let $\mathbb{P}$ be the set of all distributions over $(X,Y(0),Y(1),U,T)$. Let $\mathbb{P}^{\text{obs}}_{X,Y,T}$ be the joint distribution of all observable random variables $(X,Y,T)$. Let $t \in \{0,1\}$. Define $Q_{t,1-t}$ as the set of all counterfactual distributions that agrees with the observables and satisfies the $(f,\rho)$ selection condition, i.e.,

$$Q_{t,1-t} = \{ \mathbb{P}_{X,Y(T)|T=1-t} : \mathbb{P} \in \mathbb{P}, \mathbb{P}_{X,Y,T} = \mathbb{P}^{\text{obs}}_{X,Y,T}, \mathbb{P} \text{ satisfies Definition 1} \}.$$

Then $\mathbb{P}^{\text{sup}}_{X,Y(T)|T=1-t} \in Q_{t,1-t}$, and

$$Q_{t,1-t} = \left\{ Q : \frac{dQ_{X}|T=t}{d\mathbb{P}^{\text{obs}}_{X|T=t}}(x) = r_{t,1-t}(x), D_f(Q_{Y|X=x} \parallel \mathbb{P}^{\text{obs}}_{Y|X=x,T=t}) \leq \rho, \text{ for } \mathbb{P}^{\text{obs}}_{X|T=t} \text{-almost all } x, \right. \left. D_f(Q_{Y|X=x,T=t} \parallel \mathbb{Q}_{Y|X=x}) \leq \rho, \text{ for } \mathbb{Q}_{X} \text{-almost all } x \right\},$$

where $r_{0,0}(x) = \frac{(1-e(x))p_0}{e(x)(1-p_0)}$, $r_{1,0}(x) = \frac{e(x)(1-p_0)}{(1-e(x))p_1}$, and $e(x) = \mathbb{P}^{\text{obs}}(T=1|X=x)$, $p_1 = \mathbb{P}^{\text{obs}}(T=1)$.

Proof of Proposition 5. Fix $t = 1$. For any $\mathbb{P}_{X,Y(1)|T=0} \in Q_{1,0}$, since $\mathbb{P}_{X,Y,T} = \mathbb{P}^{\text{obs}}_{X,Y,T}$,

$$\frac{d\mathbb{P}_{X|T=0}}{d\mathbb{P}^{\text{obs}}_{X|T=1}} = \frac{d\mathbb{P}_{X|T=0}}{d\mathbb{P}_{X|T=1}} = r_{1,0}(x).$$

By Lemma 1, $D_f(\mathbb{P}_{X,Y(1)|T=0} \parallel \mathbb{P}_{X,Y(1)|T=1}) \leq \rho$. On the other hand,

$$D_f(\mathbb{P}_{Y(1)|X,T=1} \parallel \mathbb{P}_{Y(1)|X,T=0}) \leq D_f(\mathbb{P}_{Y(1),U|X,T=1} \parallel \mathbb{P}_{Y(1),U|X,T=0})$$

$$= \mathbb{E}_{\mathbb{P}_{Y(1),U}|X,T=0} \left[ f \left( \frac{d\mathbb{P}_{Y(1),U|X,T=1}}{d\mathbb{P}_{Y(1),U|X,T=0}} \right) \right]$$

$$= \mathbb{E}_{U|X,T=0} \left[ f \left( \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \cdot \frac{\mathbb{P}(T=0|X)}{\mathbb{P}(T=1|X)} \right) \right] \leq \rho,$$

where the last inequality is due to the $(f,\rho)$-selection condition. Combining the above, we establish the “⊂” direction. It remains to prove the reverse. We show the proof for the case of $t = 1$ here, and the $t = 0$ case follows from similar arguments.

Given any $Q \in Q_{1,0}$, we aim to find a distribution $\mathbb{P}^{\text{sup}}$ over $(X,Y(0),Y(1),U,T)$ such that

- $(Y(1),Y(0)) \in \mathbb{P}^{\text{sup}}|T|X,U$;
- $\mathbb{P}^{\text{sup}}$ is compatible with $\mathbb{P}^{\text{obs}}_{X,T,Y}$;
- $\mathbb{P}^{\text{sup}}$ satisfies the $(f,\rho)$-selection condition;
- $\mathbb{P}^{\text{sup}}_{X,Y(1)|T=0}(x,y) = Q(x,y)$.

To construct $\mathbb{P}^{\text{sup}}$, we first set $\mathbb{P}^{\text{sup}}_{X,T} = \mathbb{P}^{\text{obs}}_{X,T}$. Then we specify the distribution of $Y(1)|X,T$ via

$$\mathbb{P}^{\text{sup}}_{Y(1)|X,T=1} = \mathbb{P}^{\text{obs}}_{Y|X,T=1}, \quad \mathbb{P}^{\text{sup}}_{Y(1)|X,T=0} = Q \mathbb{P}_{Y|X}.$$

So far the joint distribution of $(X,T,Y(1))$ has been determined. We let $U = Y(1)$ be the unobserved confounder. Finally, the distribution of $Y(0)|X,T,Y(1),U$ is specified via

$$\mathbb{P}^{\text{sup}}_{Y(0)|X,T,Y(1),U} = \mathbb{P}^{\text{sup}}_{Y(0)|X} = \mathbb{P}^{\text{obs}}_{Y|X,T=0}.$$
Having constructed \( \mathbb{P}_{\text{sup}} \), we proceed the check that it satisfies the conditions. Conditional on \( X \) and \( U \), \( Y(1) \) becomes deterministic and the distribution of \( Y(0) \) only depends on \( X \). Hence \( \langle Y(1), Y(0) \rangle \perp | X, U \). By construction, it is straightforward to see that \( \mathbb{P}_{\text{sup}} \) is compatible with \( \mathbb{P}_{\text{obs}}^{X,T,Y} \). For any \( x \), again by the construction of \( \mathbb{P}_{\text{sup}} \),

\[
\mathbb{E}_{\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=1)} \left[ f \left( \frac{d\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=0)}{d\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=1)} \right) \right] = \mathbb{E}_{\mathbb{P}_{\text{obs}}(Y|X=x,T=1)} \left[ f \left( \frac{d\mathbb{Q}_Y|X=x}{d\mathbb{P}_{\text{obs}}(Y|X=x,T=1)} \right) \right] = D_f(Q_Y|X=x) \leq \rho,
\]

where the last inequality is due to the definition of \( Q \). On the other hand,

\[
\frac{d\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=0)}{d\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=1)} = \frac{d\mathbb{P}_{\text{sup}}(T=1|X=x)}{d\mathbb{P}_{\text{sup}}(T=0|X=x)} \cdot \frac{d\mathbb{P}_{\sup}(T=0|X=x)}{d\mathbb{P}_{\sup}(T=0|X=x)} \cdot \frac{d\mathbb{P}_{\sup}(T=1|X=x)}{d\mathbb{P}_{\sup}(T=0|X=x)},
\]

where the last equality is because \( U = Y(1) \) under \( \mathbb{P}_{\text{sup}} \). Combing the above, we have

\[
\mathbb{E}_{\mathbb{P}_{\text{sup}}(U|X=x,T=1)} \left[ f \left( \frac{d\mathbb{P}_{\text{sup}}(T=0|U,X=x)}{d\mathbb{P}_{\text{sup}}(T=1|U,X=x)} \cdot \frac{d\mathbb{P}_{\sup}(T=1|X=x)}{d\mathbb{P}_{\sup}(T=0|X=x)} \right) \right] \leq \rho.
\]

Similarly,

\[
\mathbb{E}_{\mathbb{P}_{\text{sup}}(U|X=x,T=0)} \left[ f \left( \frac{d\mathbb{P}_{\text{sup}}(T=1|U,X=x)}{d\mathbb{P}_{\text{sup}}(T=0|U,X=x)} \cdot \frac{d\mathbb{P}_{\sup}(T=0|X=x)}{d\mathbb{P}_{\sup}(T=0|X=x)} \right) \right] \leq \mathbb{E}_{\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=0)} \left[ f \left( \frac{d\mathbb{P}_{\text{sup}}(Y|X=x,T=1)}{d\mathbb{P}_{\text{sup}}(Y(1)|X=x,T=0)} \right) \right]
\]

Therefore, the super population \( \mathbb{P}_{\text{sup}} \) satisfies the \((f, \rho)\)-selection condition. By construction, \( \mathbb{P}_{Y(1)|X,T=0}^{\text{sup}} = Q_Y|X \). It remains to show that \( \mathbb{P}_{X|T=0}^{\text{sup}} = Q_X \). For any measurable set \( A \),

\[
\mathbb{P}_{\text{sup}}(X \in A | T=0) = \mathbb{E}_{\text{sup}} \left[ \frac{d\mathbb{P}_{\text{sup}}(X|T=0)}{d\mathbb{P}_{\text{sup}}(X|T=1)} \cdot 1\{X \in A\} | T=1 \right] = \mathbb{E}_{\text{sup}} \left[ \frac{d\mathbb{Q}_X}{d\mathbb{P}_{\text{obs}}(X|T=1)} \cdot 1\{X \in A\} | T=1 \right] = Q(X \in A).
\]

Since the above holds for any measurable set \( A \), \( \mathbb{P}_{X|T=0}^{\text{sup}} = Q_X \).

Finally, switching the role of 1 and 0 completes the proof.

\[ \square \]

### B  Deferred details and discussions

#### B.1  Proof of Proposition 4

Given \( X = x \), suppose instead \( \alpha^*(x) = 0 \). We consider the following two cases:
• If \( \eta^*(x) < -y(x) \), then

\[
\lim_{\alpha \to 0} \inf \mathbb{E} \left[ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) + \eta^*(x) + \alpha \rho \mid X = x, T = 1 \right] \\
= \lim_{\alpha \to 0} \inf \mathbb{E} \left[ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) 1 \{Y(1) \leq -\eta^*(x)\} \right] \\
+ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) 1 \{Y(1) > -\eta^*(x)\} \mid X = x, T = 1 \right] + \eta^*(x) + \alpha \rho \\
\geq \lim_{\alpha \to 0} \inf \mathbb{E} \left[ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) 1 \{Y(1) \leq -\eta^*(x)\} - \alpha L \mid X = x, T = 1 \right] + \eta^*(x) \\
+ \lim_{\alpha \to 0} \inf \left[ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) 1 \{Y(1) > -\eta^*(x)\} - \alpha L \mid X = x, T = 1 \right] + \eta^*(x) \\
\geq + \infty,
\]

where step (a) uses the fact that \( \liminf_{n \to \infty} a_n + b_n \geq \liminf_{n \to \infty} a_n + \liminf_{n \to \infty} b_n \) and step (b) follows from Fatou’s lemma and the condition that \( f^*(x)/x \to \infty \) when \( x \to \infty \).

• If \( \eta^*(x) \geq -y(x) \), then

\[
\lim_{\alpha \to 0} \mathbb{E} \left[ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) + \eta^*(x) + \alpha \rho \mid X = x, T = 1 \right] \\
= \lim_{\alpha \to 0} \mathbb{E} \left[ \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) 1 \{Y(1) \geq \text{ess inf} Y(1)\} + \rho \mid X = x, T = 1 \right] + \eta^*(x) \\
\leq \mathbb{E} \left[ \lim_{\alpha \to 0} \alpha f^* \left( \frac{Y(1) + \eta^*(x)}{-\alpha} \right) 1 \{Y(1) \geq \text{ess inf} Y(1)\} + \rho \mid X = x, T = 1 \right] + \eta^*(x) \\
\leq \eta^*(x).
\]

Above, step (a) is due to the fact that \( f^*(x) \) is bounded when \( x \leq 0 \) and the dominated convergence theorem; step (b) is because \( f^*(x)/x \to 0 \) as \( x \to -\infty \).

Combining the two cases above, we conclude that \( \eta^*(x) = -y(x) \) and the optimal value of the dual problem is \( -y(x) \). By the strong duality, the optimal value of the primal objective function is \( \mathbb{E} [r_{1,0}(X) y(X) \mid T = 1] \). As an implication, there exists a feasible \( L(x, y) \) such that \( \mathbb{E} [Y(1) L(X, Y(1)) \mid T = 1] = \mathbb{E} [r_{1,0}(X) y(X) \mid T = 1] \).

Let \( \mathbb{Q}_{Y \mid X = x} \) denote the measure induced by \( L(x, y) \):

\[
\frac{d \mathbb{Q}_{Y \mid X = x}}{d \mathbb{P}_{Y(1) \mid X = x, T = 1}}(y) = \frac{L(x, y)}{r_{1,0}(x)}.
\]

This is a valid transformation of measure because \( L(x, y) \) is feasible. Then \( Y(1) = y(X) \) a.s. under \( \mathbb{Q}_{Y \mid X} \).

Consequently,

\[
1 = \mathbb{Q}(Y(1) = y(X) \mid X = x) = \mathbb{E} \left[ \frac{L(x, y(X))}{r_{1,0}(X)} \right] \mid X = x, T = 1 \right] = \frac{L(x, y(X))}{r_{1,0}(x)} \cdot \bar{p}(x),
\]

\[
0 = L(x, Y(1)) \cdot 1 \{Y(1) > y(x)\}, \text{ a.s. under } \mathbb{Q}_{Y(1) \mid X = x, T = 1}.
\]

Again since \( L \) is feasible,

\[
\rho \geq \mathbb{E} \left[ f \left( \frac{L(x, Y(1))}{r_{1,0}(x)} \right) \mid X = x, T = 1 \right] = \bar{p}(x) \cdot f \left( \frac{1}{\bar{p}(x)} \right) + (1 - \bar{p}(x)) \cdot f(0).
\]

This is a contradiction to the condition. Hence \( \alpha^*(x) > 0 \).
B.2 Discussions on Assumption 1 on sieve estimation

We provide additional discussion on Assumption 1 for sieve estimators in the context of \((X, Y(1)) \mid T = 1\). In particular, we first justify the smoothness of the optimizers when the conditional distributions are sufficiently smooth. We then verify the technical conditions for two choices of \(f\)-divergences: KL-divergence and \(\chi^2\)-divergence. Then we discuss some considerations of relaxing the conditions with implementations in practice.

**Smoothness of the optimizers.** We first provide some justifications for assuming the optimizers are continuously differentiable. By the strong convexity of \(f\), its conjugate \(f^*\) is continuous, hence without loss of generality we always assume the differentiation and expectation are exchangeable. We also assume the conjugate \(f^*\) is sufficiently smooth, which is the case for many popular choices of \(f\)-divergence. As we discussed in Proposition 4, under mild conditions, the optimizers \((\alpha^*(x), \eta^*(x))\) lies in the interior of \(\mathbb{R}^+ \times \mathbb{R}\). The optimizers are thus the solutions to

\[
(\alpha^*(x), \eta^*(x)) = \arg\min_{\alpha, \beta} \nabla_{\alpha, \beta} \left\{ a \mathbb{E} \left[ f^* \left( \frac{Y(1)+b}{a} \right) \bigg| X = x, T = 1 \right] + b + a\rho \right\},
\]

where the right-hand side takes the form

\[
F(a, b, x) := \mathbb{E} \left[ g(Y(1), a, b) \bigg| X = x, T = 1 \right] \in \mathbb{R}^2
\]

for some differentiable or smooth function \(g\) decided by \(f^*\) and its derivative \((f^*)'\). Thus \(F(a, b, x)\) is smooth in \((a, b)\) when \(f^*\) is sufficiently smooth. Now let us assume the conditional distribution \(P_{Y(1) \mid X=x, T=1}\) is smooth; for example, for some \(h \in \mathcal{X}\), \(P_{Y(1) \mid X=x+th, T=1} = P_{Y(1) \mid X=x, T=1} + t \cdot P_h\) for some measure \(P_h\) on \(\mathcal{Y}\); and similar for higher-order expansions. This is a reasonable assumption if we are willing to assume that the conditional distributions of \(Y(1)\) are close for similar covariates. Concretely, such condition holds when \(Y(1) \mid X = x, T = 1\) is a normal distribution with homoskedastic noise and a smooth mean function, or heteroskedastic noise with a smooth mean function and smooth standard deviation function, etc. When the conditional distributions are smooth in \(x\), the function \(F(a, b, x)\) is also smooth in \(x\) by the linearity of conditional expectation. Finally, if the derivatives with respect to \(a, b\) are always invertible (which is the case under mild conditions for the examples we discuss shortly) and smooth, invoking the Implicit Function Theorem (Rudin et al., 1976), the minimizer can be smooth in \(x\).

**KL-divergence.** A popular choice for the function \(f\) is \(f(x) = x \log x\), which leads to the KL-divergence (Kullback and Leibler, 1951). The dual function in this case is \(f^*(y) = e^{-y-1}\), and the loss function becomes

\[
\ell(\theta, x, y) = \alpha(x)e^{\frac{y+\eta(x)}{\alpha(x)}} - 1 + \eta(x) + \alpha(x)\rho.
\]

The conditional expectation is

\[
\mathbb{E} \left[ \ell((a, b), x, Y(1)) \bigg| X = x, T = 1 \right] = a \mathbb{E} \left[ e^{\frac{Y(1)+b}{a}} - 1 \bigg| X = x, T = 1 \right] + b + a\rho.
\]

We first look at the strong convexity assumption. The conditional expectation is twice differentiable, with

\[
\nabla^2_\alpha \mathbb{E} \left[ \ell((a, b), x, Y(1)) \bigg| X = x, T = 1 \right] = \frac{1}{a^2} \mathbb{E} \left[ (Y(1) + b)^2 e^{\frac{Y(1)+b}{a}} - 1 \bigg| X = x, T = 1 \right],
\]

\[
\nabla^2_\theta \mathbb{E} \left[ \ell((a, b), x, Y(1)) \bigg| X = x, T = 1 \right] = \frac{1}{a} \mathbb{E} \left[ e^{\frac{Y(1)+b}{a}} - 1 \bigg| X = x, T = 1 \right],
\]

\[
\nabla^2_{a, b} \mathbb{E} \left[ \ell((a, b), x, Y(1)) \bigg| X = x, T = 1 \right] = -\frac{1}{a^2} \mathbb{E} \left[ (Y(1) + b) e^{\frac{Y(1)+b}{a}} - 1 \bigg| X = x, T = 1 \right].
\]

Therefore, a simple calculation shows that as long as \(Y(1)\) is not deterministic at \((\alpha^*(x), \beta^*(x))\), the Hessian matrix is non-singular. Also, if the underlying distribution \(P_{Y(1) \mid X=x, T=1}\) is continuous in \(x\), the above derivatives, hence the eigenvalues of the Hessian matrix is continuous; since \(\mathcal{X}\) is compact, there exists a positive uniform lower bound for the smallest eigenvalue of the Hessian matrix, leading to strong convexity.
We then consider the continuity condition \(|\ell(\theta, x, y) - \ell(\theta^*, x, y)| \leq \bar{\ell}(x, y) \|\theta(x) - \theta^*(x)\|_2\) for \(\|\theta(x) - \theta^*(x)\|_2 < \epsilon\) for some sufficiently small \(\epsilon > 0\), where \(\|\theta(x) - \theta^*(x)\|_2\) is the Euclidean norm, and \(\sup_{x \in \mathcal{X}} \mathbf{E}[\bar{\ell}(x, Y(1))^2 | X = x, T = 1] < M\) for some constant \(M > 0\). By Taylor expansion, we have
\[
\ell(\theta, x, y) - \ell(\theta^*, x, y) = \nabla_\theta \ell(\bar{\theta}, x, y)(\theta^*(x) - \theta(x)),
\]
where \(\bar{\theta}(x)\) lies between \(\theta(x)\) and \(\theta^*(x)\). We note that \(\nabla_\theta\) is also a smooth function of \(\theta\), and the gradient is uniform bounded for \(\theta(x)\) within a neighborhood of \(\theta^*(x)\) in terms of Euclidean \(L_2\)-norm. In particular,
\[
\frac{\partial}{\partial a} \ell((a, b), x, y) = (1 - \frac{y + b}{a}) e^{\frac{y + b}{a} - 1} + \rho, \quad \frac{\partial}{\partial b} \ell((a, b), x, y) = 1 - e^{\frac{y + b}{a} - 1}.
\]
For any \(\|(a, b) - \theta^*(x)\|_2 \leq \epsilon\) for sufficiently small \(\epsilon\), we can take \(\bar{\ell}(x, y)\) as the uniform upper bound of the Euclidean norm of the gradient, which has finite second moment if \(Y(1)\) is not too heavy-tailed.

Finally, the last condition is that there exists a constant \(C_1\) such that \(\mathbf{E}[\bar{\ell}(\theta, X, Y(1)) - \ell(\theta^*, X, Y(1)) | T = 1] \leq C_1 \|\theta - \theta^*\|_{L_2(\mathbb{P}|_T = 1)}^2\) when \(\theta \in \mathcal{P}_c(\mathcal{X}) \times \mathcal{P}_c(\mathcal{X})\) and \(\|\theta - \theta^*\|_{L_2(\mathbb{P}|_T = 1)}\) is sufficiently small. Similar to arguments in the proof of Theorem 1, sufficiently small \(\|\theta - \theta^*\|_{L_2(\mathbb{P}|_T = 1)}\) implies sufficiently small \(\|\theta - \theta^*\|_{\infty}\) for this function class. Therefore, we can consider \(\theta \in \mathcal{P}_c(\mathcal{X}) \times \mathcal{P}_c(\mathcal{X})\) such that \(\|\theta - \theta^*\|_{\infty}\) is sufficiently small. With a Taylor expansion of the conditional expectation of the risk at \((\alpha^*(x), \eta^*(x))\), we have
\[
\mathbf{E}[\ell(\theta, x, Y(1)) | X = x, T = 1] = \mathbf{E}[\ell(\theta^*, x, Y(1)) | X = x, T = 1]
\]
\[
= 1/2 \cdot \nabla_\theta^2 \mathbf{E}[\ell(\theta, x, Y(1)) | X = x, T = 1][\theta(x) - \theta^*(x), \theta(x) - \theta^*(x)]
\]
since the gradient is zero, where \(\bar{\theta}(x)\) lies between \(\theta(x)\) and \(\theta^*(x)\). Previous derivations have shown that the Hessian is continuous; also, by the compactness of \(\mathcal{X}\) and continuity of \(\theta^*(x)\), there is a uniform lower bound \(\epsilon > 0\) for \(\alpha^*(x)\). Thus, when \(\|\theta - \theta^*\|_{\infty}\) is sufficiently small, the Hessian is also bounded. Again by the compactness of \(\mathcal{X}\), this bound can be taken to be uniform for \(x \in \mathcal{X}\), which leads to the desired condition.

\(\chi^2\)-divergence. Another popular choice is \(f(x) = (x - 1)^2\), so that \(f^*(y) = \frac{1}{4}(y + 2)^2 - 1\). The conjugate function is a quadratic function on \([-\infty, \infty]\) and zero on \((-\infty, -2]\), with continuous gradient \((f^*)'(y) = (y^2 + 1)^+\), and second-order derivative \((f^*)''(y) = \frac{1}{2}\mathbb{1}(y > -2)\); the latter is almost-everywhere (under Lebesgue measure) except \(y = 2\). We now proceed to verify the conditions. The loss function is
\[
\ell(\theta, x, y) = \frac{\alpha(x)}{4} \left[\left(\frac{y + \eta(x)}{-\alpha(x)} + 1\right)^2 - 1\right] + \eta(x) + \alpha(x)\rho.
\]
Assuming \(Y(1)\) does not have point measure, the differentiation and expectation are exchangeable, and
\[
\nabla_\theta^2 \mathbf{E}[\ell((a, b), x, Y(1)) | X = x, T = 1] = \frac{1}{2a^2} \mathbf{E}\left[(Y(1) + b)\mathbb{1}\{Y(1) + b > -2\} | X = x, T = 1\right],
\]
\[
\nabla_\theta^2 \mathbf{E}[\ell((a, b), x, Y(1)) | X = x, T = 1] = \frac{1}{2a} \mathbf{E}\left[\mathbb{1}\{Y(1) + b > -2\} | X = x, T = 1\right],
\]
\[
\nabla_{a,b}^2 \mathbf{E}[\ell((a, b), x, Y(1)) | X = x, T = 1] = -\frac{1}{2a^2} \mathbf{E}\left[(Y(1) + b)\mathbb{1}\{Y(1) + b > -2\} | X = x, T = 1\right].
\]
Also, the gradient is given by
\[
\nabla_a \mathbf{E}[\ell((a, b), x, Y(1)) | X = x, T = 1] = \mathbf{E}\left[\frac{(Y(1) + b)^2 + 1}{2a} - 1 + \frac{Y(1) + b}{a} \left(\frac{Y(1) + b}{2a} + 1\right) | X = x, T = 1\right] + \rho,
\]
\[
\nabla_b \mathbf{E}[\ell((a, b), x, Y(1)) | X = x, T = 1] = -\mathbf{E}\left[\frac{(Y(1) + b)^2 + 1}{2a} | X = x, T = 1\right] + 1,
\]
which are both zero at \((a, b) = (\alpha^*(x), \eta^*(x))\). The form of the loss function implies that \(\alpha^*(x) > 0\) for almost all \(x\); hence there is a uniform lower bound \(\epsilon > 0\) by the compactness of \(\mathcal{X}\). By Cauchy-Schwarz inequality, the Hessian at \((a, b) = (\alpha^*(x), \eta^*(x))\) is positive \(\mathbb{P}(\frac{Y(1) + \eta(x)}{-\alpha(x)} > -2 | X = x, T = 1) = 0\) or \((Y(1) + \eta(x) - c(x))\mathbb{1}\{\frac{Y(1) + \eta(x)}{-\alpha(x)} > -2\} = 0\) almost surely for some \(c(x) \in \mathbb{R}\). By the optimality condition,
the former is impossible, and the latter is also impossible if $Y(1)$ is not deterministic conditional on $X = x$. Thus, as long as $Y(1) \mid X = x$ is not deterministic for almost all $x$, the Hessian is positive definite for all $x \in \mathcal{X}$. By compactness of $\mathcal{X}$ and the continuity, we know that the minimal eigenvalue of the Hessian is uniformly lower bounded away from zero, hence the strong convexity follows.

The other two conditions are easy to verify in this case: the conjugate function $f^*$ is a truncation of a quadratic function. Since truncation is a contraction map, these results hold easily by the uniform boundedness of second-order derivatives. We’ve thus verified the conditions in Assumption 1 for $\chi^2$-divergence.

**Practical conderations.** In practice, we might search for $(\alpha^*(x), \eta^*(x))$ within the function classes with a bounded range of coefficients in the two examples we give, leading to a compact function space. This is typically assumed in the contexts of $M$-estimators and sieve estimators (Van der Vaart, 2000; Geer et al., 2000; Chen and Shen, 1998; Chen, 2007). In this case, the regularity conditions are easier to verify given the uniform boundedness. The function space still provides finer and finer approximation to the targets if the bounded range enlarges properly with $n$.

### B.3 Estimators for bounds on counterfactual means

In this section, we summarize the application of the procedure in Section 3.3 to estimate other lower and upper bounds on counterfactual means.

(a) **Upper bound of $\mathbb{E}[Y(1) \mid T = 0]$**: Let $-\hat{\mu}_{1,0}$ be the estimator obtained from the procedure in Section 3.3 with $-Y(1)$ replacing $Y(1)$. Then $\sqrt{n}(\hat{\mu}_{1,0} - \mu_{1,0}) \rightarrow N(0, \text{Var}(\phi_{1,+}(X,Y,T)))$, with influence function

$$\phi_{1,+}(X_i, Y_i, T_i) = \frac{T_i}{p_1} r_{1,0}(X_i)[H_{1,+}(X_i, -Y_i(1)) - h_{1,+}(X_i)] + \frac{1 - T_i}{p_0} h_{1,+}(X_i),$$

where $H_{1,+}(x, y) = \alpha^*_1(x)f(x)(\frac{\eta^*_1(x)}{\alpha^*_1(x)}) + \eta^*_1(x) + \alpha^*_1(x)\rho$ with $(\alpha^*_1(x), \eta^*_1(x))$ being the minimizer of $\mathbb{E}[\alpha f^*(\frac{-Y(1)+\eta}{\alpha}) + \eta + \alpha \rho \mid X = x, T = 1]$, and $h_{1,+}(x) = \mathbb{E}[H_{1,+}(X, -Y(1)) \mid X = x, T = 1]$.

(b) **Lower bound of $\mathbb{E}[Y(0) \mid T = 1]$**: Let $\hat{\mu}_{0,1}$ be the estimator obtained from the procedure in Section 3.3 switching the role of treated and control groups. Then $\sqrt{n}(\hat{\mu}_{0,1} - \mu_{0,1}) \rightarrow N(0, \text{Var}(\phi_{0,-}(X,Y,T)))$ with influence function

$$\phi_{0,-}(X_i, Y_i, T_i) = \frac{1 - T_i}{p_0} r_{0,1}(X_i)[H_{0,-}(X, Y(0)) - h_{0,-}(X_i)] + \frac{T_i}{p_1} h_{0,-}(X_i).$$

Here $H_{0,-}(x, y) = \alpha^*_{0,-}(x)f(x)(\frac{\eta^*_{0,-}(x)}{\alpha^*_{0,-}(x)}) + \eta^*_{0,-}(x) + \alpha^*_{0,-}(x)\rho$, and $(\alpha^*_{0,-}(x), \eta^*_{0,-}(x))$ is the minimizer of $\mathbb{E}[\alpha f^*(\frac{Y(0)+\eta}{\alpha}) + \eta + \alpha \rho \mid X = x, T = 0]$, and $h_{0,-}(x) = \mathbb{E}[H_{0,-}(X, Y(0)) \mid X = x, T = 0]$.

(c) **Upper bound of $\mathbb{E}[Y(0) \mid T = 1]$**: Let $-\hat{\mu}_{0,1}$ be the estimator obtained from the procedure in Section 3.3 switching the role of treated and control groups and replacing $Y(0)$ with $-Y(0)$. Then $\sqrt{n}(\hat{\mu}_{0,1} - \mu_{0,1}) \rightarrow N(0, \text{Var}(\phi_{0,+}(X,Y,T)))$ with influence function

$$\phi_{0,+}(X_i, Y_i, T_i) = \frac{1 - T_i}{p_0} r_{1,0}(X_i)[H_{0,+}(X_i, -Y_i(1)) - h_{0,+}(X_i)] + \frac{T_i}{p_1} h_{0,+}(X_i),$$

where $H_{0,+}(x, y) = \alpha^*_0(x)f(x)(\frac{\eta^*_0(x)}{\alpha^*_0(x)})) + \eta^*_0(x) + \alpha^*_0(x)\rho$ with $(\alpha^*_0(x), \eta^*_0(x))$ being the minimizer of $\mathbb{E}[\alpha f^*(\frac{-Y(0)+\eta}{\alpha}) + \eta + \alpha \rho \mid X = x, T = 0]$, and $h_{0,+}(x) = \mathbb{E}[H_{0,+}(X, -Y(0)) \mid X = x, T = 0]$.
C Technical proofs

C.1 Proof of Proposition 3

Proposition 3. We first claim that solving (6) amounts to solving the following problem for each $x$:

$$
\min_{L(x) \text{ measurable}} \mathbb{E}[Y(1)L(x) \mid X = x, T = 1] \tag{12}
$$

s.t. $\mathbb{E}[L(x) \mid X = x, T = 1] = r_{1,0}(x)$

$\mathbb{E}[f(L(x)/r_{1,0}(x)) \mid X = x, T = 1] \leq \rho.$

To be specific, denoting the optimal objective of (12) as $\mu(x)$ and that of (6) as $\mu_{1,0}$, we are to show that $\mu_{1,0} = \mathbb{E}[\mu(X) \mid T = 1]$. To see why it is the case, suppose $L^*$ is the optimizer of (6), then it is measurable with respect to $X$ and $Y(1)$ and satisfies the constraints of (6). Then $L(x)(\cdot) := L^*(x, \cdot)$ is measurable with respect to $Y(1)$, and satisfy the constraints of (12). As a result, we have $\mathbb{E}[\mu(Y(1)L(x,y)) \mid X = x, Y(1) = T = 1] \geq \mathbb{E}[\mu(X) \mid T = 1]$. On the other hand, suppose $L^*(x,\cdot)$ is measurable with respect to $Y(1)$ and is the minimizer for (12) for $\mathbb{P}_{X \mid T = 1}$-almost all $x$. We let $L(x,y) = L^*(x,y)$, so that it is measurable with respect to $(X,Y(1))$ and satisfy the constraints of (6). Thus we have $\mathbb{E}[L(X,Y(1))Y(1) \mid T = 1] = \mathbb{E}[\mu(X) \mid T = 1] \geq \mu_{1,0}$. Combining the two directions leads to the equivalence.

In the following, we solve (12) and write $\mathbb{E}_x$ in place of $\mathbb{E}[\cdot \mid X = x, T = 1]$ for simplicity. Invoking Luenberger (1997, Theorem 8.6.1) to this convex problem, we have

$$
\min_{\mathbb{E}_x[L=\mathbb{E}_x(L(x))], \mathbb{E}_x[f(L/r_{1,0}(x))] \geq 0} \mathbb{E}_x[Y(1)L(x)] = \max_{\alpha \geq 0, \eta \in \mathbb{R}} \varphi(\alpha, \eta, x),
$$

where the Slater’s condition is satisfied and strong duality holds, and

$$
\varphi(\alpha, \eta, x) = \inf_{L \geq 0 \text{ measurable}} \mathcal{L}(\alpha, \eta, L, x),
\mathcal{L}(\alpha, \eta, L, x) = \mathbb{E}_x[Y(1)L(x)] + \eta \mathbb{E}_x[L - r_{1,0}(x)] + \alpha(\mathbb{E}_x[f(L/r_{1,0}(x))] - \rho).
$$

The minimum of $\mathcal{L}(\alpha, \eta, L, x)$ is thus given by

$$
\varphi(\alpha, \eta, x) = \mathbb{E}_x \left[ \min_{z \geq 0} \left\{ Y(1)z + \eta z - \eta r_{1,0}(x) + \alpha f(z/r_{1,0}(x)) - \alpha \rho \right\} \right]
= \mathbb{E}_x \left[ -\alpha f^* \left( \frac{r_{1,0}(x)}{-\alpha} \right) (Y(1) + \eta) - \eta r_{1,0}(x) - \alpha \rho \right].
$$

Now we write $\alpha(x)$ and $\eta(x)$ to emphasize its dependency on $x$. Therefore, by the equivalence discussed in the beginning, we have

$$
\mu_{1,0} = \mathbb{E} \left[ \max_{\alpha(X) \geq 0, \eta(X) \in \mathbb{R}} \varphi(\alpha(X), \eta(X), X) \mid T = 1 \right]
= \mathbb{E} \left[ \varphi(\alpha^*(X), \eta^*(X), X) \mid T = 1 \right],
$$

where for $\mathbb{P}_{X \mid T = 1}$-almost all $x$,

$$
(\alpha^*(x), \eta^*(x)) \in \text{argmax} \mathbb{E}_{\alpha \geq 0, \eta \in \mathbb{R}} \left[ -\alpha f^* \left( \frac{r_{1,0}(x)}{-\alpha} \right) (Y(1) + \eta) - \eta r_{1,0}(x) - \alpha \rho \mid X = x, T = 1 \right].
$$

With a change-of-variable from $\alpha(x)$ to $\alpha(x)r_{1,0}(x)$, we have

$$
(\alpha^*(x)/r_{1,0}(x), \eta^*(x)) \in \text{argmax} \mathbb{E}_{\alpha \geq 0, \eta \in \mathbb{R}} \left[ -\alpha r_{1,0}(x) f^* \left( \frac{Y(1) + \eta}{-\alpha} \right) - \eta r_{1,0}(x) - \alpha r_{1,0}(x) \rho \mid X = x, T = 1 \right].
$$
The minimum of (6) can thus be written as
\[
\mu_{1,0} = -\mathbb{E} \left[ r_{1,0}(x) \left\{ \alpha^*(X) f^* \left( \frac{Y(1) + \eta^*(X)}{-\alpha^*(X)} \right) + \eta^*(X) + \alpha^*(X) \rho \right\} \right| T = 1 ,
\]
where for \( \mathbb{P}_X | T = 1 \)-almost all \( x \), it holds that
\[
\left( \alpha^*(x), \eta^*(x) \right) \in \arg \min_{\alpha \geq 0, \eta \in \mathbb{R}} \mathbb{E} \left[ \alpha f^* \left( \frac{Y(1) + \eta}{-\alpha} \right) + \eta + \alpha \rho \right| X = x, T = 1 .
\]
Therefore, we complete the proof of Proposition 3. \( \square \)

### C.2 Proof of convergence of sieve estimator

**Proof of Theorem 1.** We analyze the behavior of \( \hat{\theta}(j) \) for each fold \( j \). As \( |Z_i^{(j)}| \approx n \), we take the generic notation of \( \hat{\theta} \) and sample size \( n \), so that
\[
\hat{\mathbb{E}}_n \left[ \ell(\hat{\theta}, X, Y(1)) \right] \geq \inf_{\theta \in \Theta} \mathbb{E}_n \left[ \ell(\theta, X, Y(1)) \right] - O_P \left( \frac{\log n}{n^{2p/(2p+d)}} \right),
\]
where \( (X_i, Y_i) \sim \mathbb{P}_{X,Y(1)} | T = 1 \) are i.i.d. data. For some fixed \( b > 0 \), we denote the sequence
\[
\delta_n := \inf \left\{ \delta \in (0,1) : \frac{1}{\sqrt{n} \delta^2} \int_{|x| \leq \delta} \log N(\epsilon_{1+d/2p}, \Theta_n, \| \cdot \|_{L_2(\mathbb{P}_T=1)}) \, d\epsilon \leq 1 \right\},
\]
where \( N(\epsilon, \Theta_n, \| \cdot \|_{L_2(\mathbb{P}_T=1)}) \) is the \( \epsilon \)-covering number of \( \Theta_n \) in the \( L_2 \)-norm under \( \mathbb{P}_T=1 \). We employ the established convergence results for sieve estimators adapted from Chen (2007, Theorem 3.2) and Yadlowsky et al. (2018, Lemma B.3), stated in Lemma 2.

**Lemma 2.** Let \( \theta^* \in \Theta \) be a population risk minimizer. Suppose there exists constants \( c_1, c_2 > 0 \) such that
\[
c_1 \mathbb{E} [\ell(\theta, X, Y) - \ell(\theta^*, X, Y)] \leq d(\theta, \theta^*)^2 \leq c_2 \mathbb{E} [\ell(\theta, X, Y) - \ell(\theta^*, X, Y)] \text{ for } \theta \text{ in a neighborhood of } \theta^*.
\]
Suppose the following conditions hold:

(i) For sufficiently small \( \epsilon > 0 \),
\[
\text{Var} (\ell(\theta, X, Y) - \ell(\theta^*, X, Y)) \leq C_1 \epsilon^2 \text{ for all } \theta \in \Theta_n \text{ such that } d(\theta, \theta^*) \leq \epsilon.
\]

(ii) For any \( \delta > 0 \), there exists a constant \( s \in (0,2) \) and a measurable function \( U_n(\cdot) \) such that
\[
\sup_{\theta \in \Theta} \mathbb{E} [U_n(X,Y)^2] \leq C_3 \text{ and } \sup_{\theta \in \Theta} \mathbb{E} [ U_n(X,Y) \left| X = x \right] \leq \delta^2 U_n(X,Y) \text{ for constant } C_3 > 0.
\]

Then \( d(\hat{\theta}_n - \theta^*) = O_P (\max \{ \delta_n, \inf \theta \in \Theta_n d(\theta, \theta^*) \} ) \).

We define the distance as \( L_2 \)-norm \( d(\theta, \theta^*) = \| \theta - \theta^* \|_{L_2(\mathbb{P})} \), and verify the conditions in Lemma 2. We define \( \Theta = \Lambda^p(X) \times \Lambda^p(X) \) without truncation. The upper bound \( \mathbb{E} [\ell(\theta, X, Y) - \ell(\theta^*, X, Y) | T = 1] \) is directly implied by Assumption 1. By the \( \lambda \)-strong convexity of \( \mathbb{E} [\ell((a,b), x, y) | X = x] \) is at \((a,b) = \theta^*(x), \)
\[
\mathbb{E} [\ell(\theta(x), x, y) | X = x] - \mathbb{E} [\ell(\theta^*(x), x, y) | X = x] \geq \lambda (\theta(x) - \theta^*(x))^2.
\]
Integrating over \( X \) yields \( \mathbb{E} [\ell(\theta(X), X, Y) - \ell(\theta^*(X), X, Y)] \geq c'' d(\theta, \theta^*) \) for some constant \( c'' > 0 \).

We then check condition (i). By the positive density condition, we have \( \| \cdot \|_{L_2(\lambda)} \propto \| \cdot \|_{L_2(\mathbb{P})} \). Hence \( \| \theta - \theta^* \|_\infty = o(1) \) once \( \| \theta - \theta^* \|_{L_2(\mathbb{P})} = o(1) \). By Lemma 2 of Chen and Shen (1998), we have \( \| \theta \|_\infty \lesssim \| \theta \|_{L_2(\lambda)}^{2p/(2p+d)} \) for any \( \theta \in \Theta \), where \( \lambda \) is the Lebesgue measure. Therefore, sufficiently small \( \| \theta - \theta^* \|_{L_2(\mathbb{P})} \) implies sufficiently small \( \| \theta - \theta^* \|_\infty \). Since for \( \| \theta - \theta^* \|_\infty \) sufficiently small, \( |\ell(\theta(x), x, y) - \ell(\theta^*(x), x, y)| \leq \ell(x, y)(\theta(x) - \theta^*(x)) \) where \( \mathbb{E} [\ell(x, y)^2 | X = x] \leq M \) for all \( x \), we have
\[
\text{Var} (\ell(\theta, X, Y) - \ell(\theta^*, X, Y)) \leq \mathbb{E} [|\ell(\theta(x), x, y) - \ell(\theta^*(x), x, y)|^2] \leq M \mathbb{E} [(\theta(X) - \theta^*(X))^2] \leq M \epsilon^2.
\]
for all $\theta \in \Theta_n$ such that $d(\theta, \theta^*) \leq \epsilon$ for sufficiently small $\epsilon > 0$. Condition (ii) follows from the same argument by taking $U_n(x, y) = \ell(x, y)$. Therefore, applying Lemma 2 we have $\|\hat{\theta}_n - \theta^*\|_{L_2(p)} = O_P(\max\{\delta_n, \inf_{\theta \in \Theta_n} d(\theta^*, \theta^*)\})$.

Here according to Chen and Shen (1998) and Geer et al. (2000), we have

$$\log N(\epsilon, \Theta_n^m, \|\cdot\|_{2,p}) \leq \text{dim}(\Theta_n^m) \log \frac{1}{\epsilon},$$

where $\text{dim}(\Theta_n^m) = J^n_p$. Since truncation is a contraction map, the covering number of $\Theta_n^m$ is upper bounded by the above quantity. As a result, we have

$$\log N(\epsilon, \Theta_n, \|\cdot\|_{2,p}) \leq J^n_p \log \frac{1}{\epsilon}.$$

Similar to the results in Yadlowsky et al. (2018), we have

$$\delta_n \leq \sqrt{\frac{J^n_d}{n}}.$$

We finally bound the approximation error using $\Theta_n$. Note that we take $\Theta_n$ to be truncated at $\epsilon$. However, since the population minimizer $\theta^*$ is uniformly bounded above $\epsilon$, since truncation is a contraction map, we have $\inf_{\theta \in \Theta_n} \|\theta - \theta^*\|_{L_2(p)} \leq \text{dim}(\Theta_n) \|\theta - \theta^*\|_{L_2(p)} \leq O(J^n_p)$, where the last inequality is a well-established result, see, e.g., Timan (2014). We now set $J_n = (n/\log n)^{1/(2p+d)}$, so that $\|\hat{\theta} - \theta^*\|_{L_2(p)} = O_P((\log n/n)^{p/(2p+d)})$. This completes our proof.

**C.3 Proof of Theorem 2**

**Proof of Theorem 2.** We consider the general scenario where $(\hat{\alpha}^{(j)}, \hat{\eta}^{(j)})$ converges in sup-norm to some fixed $(\alpha^*, \eta^*)$, and show that $-\hat{\mu}^{(j)}_{1,0} \overset{P}{\to} \mathbb{E}[r(X)\ell(\theta^*(X), X, Y(1) \mid T = 1)]$ for any fixed $j$, where the risk function $\ell$ is defined in Proposition 3. In the following, we drop the dependency on $j$ for notational convenience; we are to show that with estimators $\hat{f}, \hat{H}$ and $\hat{h}$ that are independent of $\mathcal{I}_1$ and $\mathcal{I}_0$,

$$\hat{\mu} := \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \hat{r}(X_i)(\hat{H}(X_i, Y_i) - \hat{h}(X_i)) + \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \hat{h}(X_i) \overset{P}{\to} \mathbb{E}[r(X)\ell(\theta^*(X), X, Y(1) \mid T = 1)].$$

Therefore, if $\theta^* = \theta^*$, we have $\hat{\mu}_{1,0} = \mu_{1,0} + o_P(1)$ since $\mu_{1,0} = \mathbb{E}[r(X)\ell(\theta^*(X), X, Y(1) \mid T = 1)]$ by Proposition 3. Otherwise, since $\mathbb{E}[r(X)\ell(\theta^*(X), X, Y(1)) \mid T = 1] \leq \mathbb{E}[r(X)\ell(\theta^*(X), X, Y(1)) \mid T = 1]$, we have the one-sided validity that $\hat{\mu}_{1,0} \overset{P}{\to} -\mathbb{E}[r(X)\ell(\theta^*(X), X, Y(1)) \mid T = 1] \leq \mu_{1,0}$, i.e., our estimator converges to a valid lower bound.

It thus remains to show (13). We prove the results when either $\hat{r}$ or $\hat{h}$ is consistent.

**Consistent $\hat{r}$.** We first show the case where $\hat{r}$ is consistent for $r_{1,0}$, but not necessarily the regression function $\hat{h}$. Recall that $\hat{H}(x, y) = \ell(\hat{\theta}(x), x, y)$. Note that

$$\hat{\mu} = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} (\hat{r}(X_i) - r(X_i))(\hat{H}(X_i, Y_i) - \hat{h}(X_i)) + \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} r(X_i)(\hat{H}(X_i, Y_i) - \hat{h}(X_i)) + \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \hat{h}(X_i).$$

The first summation can be controlled as (where the expectation is implicitly conditional on other folds except $\mathcal{I}^{(j)}_0 \cup \mathcal{I}^{(j)}_1$)

$$\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} (\hat{r}(X_i) - r_{1,0}(X_i))(\hat{H}(X_i, Y_i) - \hat{h}(X_i))\right)^2\right] \leq \mathbb{E}[\hat{r}(X_i) - r_{1,0}(X_i)]^2 \mathbb{E}[\hat{H}(X_i, Y_i) - \hat{h}(X_i)]^2 \leq M \cdot \|\hat{r} - r_{1,0}\|_{L_2(p \mid T = 1)}^2 = o_P(1).$$
Invoking Lemma 3, we can drop the conditioning and the first summation is \( o_P(1) \). On the other hand, since the covariate shift between \( \mathbb{P}_X | T = 1 \) and \( \mathbb{P}_X | T = 0 \) is exactly \( r_{1,0} \), we know that \( \mathbb{E}[\hat{h}(X) r(X) | T = 1] = \mathbb{E}[\hat{h}(X) | T = 0] \), where we still implicitly condition on other folds. As a result,

\[
-\frac{1}{|I_1|} \sum_{i \in I_1} r_{1,0}(X_i) \hat{h}(X_i) + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{h}(X_i)
\]

\[
= -\frac{1}{|I_1|} \sum_{i \in I_1} (r_{1,0}(X_i) \hat{h}(X_i) - \mathbb{E}[\hat{h}(X) r(X) | T = 1]) + \frac{1}{|I_0|} \sum_{i \in I_0} (\hat{h}(X_i) - \mathbb{E}[r(X) | T = 0]),
\]

where both terms are unbiased. Thus by Cauchy-Schwarz inequality,

\[
\mathbb{E} \left[ \left( -\frac{1}{|I_1|} \sum_{i \in I_1} r_{1,0}(X_i) \hat{h}(X_i) + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{h}(X_i) \right)^2 \right] 
\leq \frac{2}{|I_1|} \mathbb{Var}(r_{1,0}(X) \hat{h}(X) | T = 1) + \frac{2}{|I_0|} \mathbb{Var}(\hat{h}(X) | T = 0) = o_P(1)
\]

invoking the assumption that \( \hat{h} \) as finite second moment. Drop the conditioning by Lemma 3, we know that this summation is also \( o_P(1) \), hence

\[
\hat{\mu} = \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i) \hat{H}(X_i, Y_i) + o_P(1)
\]

\[
= \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i) \ell(\theta^*(X_i), X_i, Y_i) + \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i) \{ \ell(\hat{\theta}(X_i), X_i, Y_i) - \ell(\theta^*(X_i), X_i, Y_i) \} + o_P(1)
\]

\[
= \mathbb{E}[r(X_1) \ell(\theta^*(X), X, Y(1)) | T = 1] + \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i) \{ \ell(\hat{\theta}(X_i), X_i, Y_i) - \ell(\theta^*(X_i), X_i, Y_i) \} + o_P(1).
\]

Finally, once \( \| \theta - \theta^* \|_{\infty, \mathbb{P}_X | T = 1} = o_P(1) \), by the local expansion around \( \theta^*(x) \), we have

\[
\left| \ell(\hat{\theta}(X_i), X_i, Y_i) - \ell(\theta^*(X_i), X_i, Y_i) \right| \leq M(X_i, Y_i) \| \theta(X_i) - \theta^*(X_i) \|_2,
\]

hence (implicitly conditioning on other folds) we have

\[
\mathbb{E} \left[ \left( \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i) \{ \ell(\hat{\theta}(X_i), X_i, Y_i) - \ell(\theta^*(X_i), X_i, Y_i) \} \right)^2 \right]
\leq \mathbb{E} \left[ r(X_1)^2 M(X, Y(1))^2 \| \hat{\theta}(X_i) - \theta^*(X_i) \|_2^2 \right] = o_P(1)
\]

since \( \mathbb{E}[M(X, Y(1))^2 | T = 1] \leq M \) for some constant \( M > 0 \). We’ve thus completed the proof of (13).

**Consistent \( \hat{h} \).** We then show the results when \( \hat{h} \) is consistent, but not necessarily \( \hat{r} \). In this case, \( \| \hat{h} - \hat{h} \|_{L_2(\mathbb{P}_X | T = 1)} = o_P(1) \), where \( \hat{h} = \mathbb{E}[\hat{H}(X, Y(1)) | X = x, T = 1] \) viewing \( \hat{H} \) as fixed. Note that

\[
\hat{\mu} = \frac{1}{|I_1|} \sum_{i \in I_1} \hat{r}(X_i) (\hat{H}(X_i, Y_i) - \hat{h}(X_i)) + \frac{1}{|I_1|} \sum_{i \in I_1} \hat{r}(X_i) (\hat{h}(X_i) - \hat{h}(X_i)) + \frac{1}{|I_0|} \sum_{i \in I_0} \hat{h}(X_i).
\]

The first summation is unbiased conditional on other folds, hence

\[
\mathbb{E} \left[ \left( \frac{1}{|I_1|} \sum_{i \in I_1} \hat{r}(X_i) (\hat{H}(X_i, Y_i) - \hat{h}(X_i)) \right)^2 \right] = \frac{1}{|I_1|} \mathbb{Var}(\hat{r}(X) \{ \hat{H}(X, Y(1)) - \hat{h}(X) \} | T = 1) = o_P(1)
\]

de to the finite second moments. By Cauchy-Schwarz inequality, the second summation satisfies

\[
\mathbb{E} \left[ \left( \frac{1}{|I_1|} \sum_{i \in I_1} (\hat{h}(X_i) - \hat{h}(X_i)) \right)^2 \right] \leq \| \hat{r} \cdot (\hat{h} - \hat{h}) \|_{L_2(\mathbb{P}_X | T = 1)}^2 = o_P(1)
\]
due to the boundedness of $\hat{r}$. Similarly, we know that \( \frac{1}{|I|_0} \sum_{i \in I_0} \tilde{h}(X_i) - \hat{h}(X_i) = o_P(1) \). Consequently, invoking Lemma 3 we drop the implicit conditioning and arrive at

\[
\hat{\mu} = \frac{1}{|I_1|} \sum_{i \in I_1} \tilde{h}(X_i) + o_P(1) = \mathbb{E}[\hat{h}(X) \mid T = 0] + o_P(1)
\]

further using the finite second moment of $\hat{h}$ (or $\tilde{H}$), where we implicitly condition on other folds and view $\hat{h}$ as fixed. Finally, denoting \( h^\circ(x) = \mathbb{E}[\ell(\theta^\circ(x), x, Y(1)) \mid X = x, T = 1] \), we note that by Jensen’s inequality,

\[
\left| \mathbb{E}[\hat{h}(X) \mid T = 0] - \mathbb{E}[h^\circ(X) \mid T = 0] \right|^2 \leq \| \hat{h} - h^\circ \|^2_{L_2(\mathbb{P}^X \mid T = 0)} \leq \| \tilde{H}(X, Y(1)) - \ell(\theta^\circ(X), X, Y(1)) \|^2_{L_2(\mathbb{P}^X \mid T = 0)}.
\]

By the same argument as the previous case and due to the uniform boundedness of the covariate shift $r_{1,0}(\cdot)$, the above term is $o_P(1)$. Therefore, by the change-of-measure with $r_{1,0}$, we have

\[
\hat{\mu} = \mathbb{E}[h^\circ(X) \mid T = 0] + o_P(1) = \mathbb{E}[r(X) h^\circ(X) \mid T = 1] + o_P(1) = \mathbb{E}[r(X) \ell(\theta^\circ(X), X, Y(1)) \mid T = 1] + o_P(1)
\]

by the tower property of conditional expectations. We thus complete the proof of two cases and conclude the proof of Theorem 2.

\[\square\]

C.4 Proof of Theorem 3

Proof of Theorem 3. We show that for each $j$, we have $\tilde{\mu}_{1,0}^{(j)} = \tilde{\mu}_{1,0}^{*,(j)} + o_P(1/\sqrt{n})$, where

\[
\tilde{\mu}_{1,0}^{*,(j)} = \frac{1}{|I_1^{(j)}|} \sum_{i \in I_1^{(j)}} r_{1,0}(X_i)(H(X_i, Y_i) - h(X_i)) + \frac{1}{|I_0^{(j)}|} \sum_{i \in I_0^{(j)}} h(X_i),
\]

and we define

\[
H(x, y) = \alpha^*(x) f^*(\frac{y + \eta^*(x)}{-\alpha^*(x)}) + \eta^*(x) + \alpha^*(x) \rho, \quad h(x) = \mathbb{E}[H(X, Y(1)) \mid X = x, T = 1].
\]

We show this result for any $j$; we implicitly condition on all the remaining folds other than $I_1^{(j)}$ and $I_0^{(j)}$, so that all nuisance components are viewed as fixed. To simplify notations, we write $I_1 := I_1^{(j)}$, $I_0 := I_0^{(j)}$ and $r := r_{1,0}, \tilde{r} := \tilde{r}^{(j)}, \hat{r} := \hat{r}^{(j)}, \tilde{H} := \tilde{H}^{(j)}, \hat{H} := \hat{H}^{(j)}, \tilde{h} := \tilde{h}^{(j)}$. We also represent the parameters (functionals) with

\[
\tilde{\theta}(\cdot) = (\tilde{\alpha}(\cdot), \tilde{\eta}(\cdot)) := (\tilde{\theta}^{(j)}(\cdot), \tilde{\eta}^{(j)}(\cdot)), \quad \theta^*(\cdot) := (\alpha^*(\cdot), \eta^*(\cdot)),
\]

and recall the generic function (where $\theta = (\alpha(\cdot), \eta(\cdot))$)

\[
\ell(\theta, x, y) = \alpha(x) f^*(\frac{y + \eta(x)}{-\alpha(x)}) + \eta(x) + \alpha(x) \rho,
\]

so that $H(x, y) = \ell(\theta^*, x, y)$ and $\tilde{H}(x, y) = \ell(\tilde{\theta}, x, y)$. By definition, we have the decomposition

\[
\tilde{\mu}_{1,0}^{(j)} - \tilde{\mu}_{1,0}^{*,(j)} = \frac{1}{|I_1|} \sum_{i \in I_1} \left[ \tilde{r}(X_i)(\tilde{H}(X_i, Y_i) - \tilde{h}(X_i)) - r(X_i)(H(X_i, Y_i) - h(X_i)) \right] + \frac{1}{|I_0|} \sum_{i \in I_0} (\tilde{h}(X_i) - h(X_i))
\]

\[
= \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i)(\tilde{H}(X_i, Y_i) - H(X_i, Y_i)) - \frac{1}{|I_1|} \sum_{i \in I_1} (\tilde{r}(X_i) - r(X_i))(\tilde{h}(X_i) - h(X_i))
\]

\[+ \frac{1}{|I_1|} \sum_{i \in I_1} (\tilde{r}(X_i) - r(X_i))(\tilde{H}(X_i, Y_i) - \tilde{h}(X_i)) - \frac{1}{|I_1|} \sum_{i \in I_1} r(X_i)(\tilde{h}(X_i) - h(X_i)) + \frac{1}{|I_0|} \sum_{i \in I_0} (\tilde{h}(X_i) - h(X_i)).
\]
In the following, we are to bound the several summations separately. Firstly, by Cauchy-Schwarz inequality,
\[
\left| \frac{1}{|I_1|} \sum_{i \in I_1} (\hat{r}(X_i) - r(X_i)) (\hat{h}(X_i) - h(X_i)) \right| \leq \sqrt{\frac{1}{|I_1|} \sum_{i \in I_1} (\hat{r}(X_i) - r(X_i))^2} \cdot \sqrt{\frac{1}{|I_1|} \sum_{i \in I_1} (\hat{h}(X_i) - h(X_i))^2}
\]
\[
= O_P(\|\hat{r} - r\|_{L_2(P_X|T=1)} \cdot \|\hat{h} - h\|_{L_2(P_X|T=1)}) = o_P(1/\sqrt{n})
\]
under the given convergence rate of the product. Since \(\hat{h}(x) = \mathbb{E}[\hat{H}(X,Y) | X = x, T = 1]\) for the fixed function \(\hat{H}\), the term \((\hat{r}(X_i) - r(X_i))(\hat{H}(X_i,Y_i) - h(X_i))\) has mean zero, hence by Markov’s inequality,
\[
\frac{1}{|I_1|} \sum_{i \in I_1} (\hat{r}(X_i) - r(X_i)) (\hat{h}(X_i) - h(X_i)) = O_P(\sqrt{\text{Var}(\hat{r}(X_i) - r(X_i))(\hat{H}(X_i,Y_i) - h(X_i))/\sqrt{n}}),
\]
where by the consistency of \(\hat{r}\), this term is \(o_P(1/\sqrt{n})\). Furthermore, note that
\[
\frac{1}{|I_1|} \sum_{i \in I_1} r(X_i)(\hat{h}(X_i) - h(X_i)) - \frac{1}{|I_0|} \sum_{i \in I_0} (\hat{h}(X_i) - h(X_i))
\]
\[
= \frac{1}{|I_1|} \sum_{i \in I_1} \left( r(X_i)(\hat{h}(X_i) - h(X_i)) - \mathbb{E}[r(X)(\hat{h}(X) - h(X)) | T = 1] \right)
\]
\[
- \frac{1}{|I_0|} \sum_{i \in I_0} (\hat{h}(X_i) - h(X_i) - \mathbb{E}[\hat{h}(X_i) - h(X_i) | T = 0]),
\]
where we use the equivalence of the two expectations: this is because there is a covariate shift \(\bar{X} \neq X\) when \(n\) is sufficiently small, by the \(L^2\)-consistency of \(\hat{h}\) to \(h\) and the fact that \(\|\hat{h} - h\|_{L_2(P_X|T=1)}\) by the stability of the conditional expectations induced by the stability of \(g\) in Assumption 3. Finally, we turn to
\[
\frac{1}{|I_1|} \sum_{i \in I_1} r(X_i)(\hat{H}(X_i,Y_i) - H(X_i,Y_i)).
\]
Since \(\ell(\theta,x,y)\) is a convex function in \(\theta\) for any \((x,y)\), for any \((\mathbb{P}_X|T=1)-\text{almost all}\) \(x,\)
\[
\mathbb{E}[\ell(\theta,x,y) | X = x, T = 1] = \alpha(x) \mathbb{E} \left[ f^*(Y(1) + \eta(x)) \mid X = x, T = 1 \right] + \eta(x) + \alpha(x) \rho
\]
is also convex and differentiable by the given regularity condition. In particular, by the optimality of \((\alpha^*(x),\eta^*(x))\) for the per-\(x\) minimization problem and the exchangeability of differentiation and expectation,
\[
\nabla_{\theta} \mathbb{E}[\ell(\theta,x,Y(1)) | X = x, T = 1]_{\theta=(\alpha^*(x),\eta^*(x))} = \mathbb{E}[\nabla_{\theta} \ell(\theta^*(x),x,Y(1)) | X = x, T = 1] = 0.
\]
Multiplying \(r(X)\) and integrating over \(X | T = 1\), we know that
\[
\mathbb{E}[r(X)\nabla_{\theta} \ell(\theta^*(X),X,Y(1))]_{\theta=(\alpha^*(x),\eta^*(x))} = \mathbb{E}[\nabla_{\theta} \ell(\theta^*(x),x,Y(1)) | X = x, T = 1] = 0.
\]
By Lemma 2 of Chen and Shen (1998), when both \(\hat{\theta}\) and \(\theta^*\) is smooth enough, sufficiently small \(\|\hat{\theta} - \theta^*\|_{L_2(\mathbb{P})}\) implies sufficiently small \(\|\hat{\theta} - \theta^*\|_\infty\). As a result, when \(\|\hat{\theta}(x) - \theta^*(x)\|_{L_2(\mathbb{P}|T=1)}\) is sufficiently small, by the
condition that $|\ell(\hat{\theta}, x, y) - \ell(\theta^*, x, y) - \nabla \ell(\theta^*(x), x, y)(\theta^*(x) - \hat{\theta}(x))| \leq M(x, y)\|\hat{\theta}(x) - \theta^*(x)\|_2^2$ as well as Jensen’s inequality, we have

$$
\mathbb{E}\left[|\mathcal{H}(X, Y(1)) - H(X, Y(1))| \mid T = 1\right] = \mathbb{E}\left[|\mathcal{r}(X)\mathcal{H}(X, Y(1)) - H(X, Y(1))| \mid T = 1\right] \\
\leq \mathbb{E}\left[|\mathcal{r}(X)|\mathcal{H}(X, Y(1)) - \mathcal{H}(X) - H(X, Y(1))| \mid T = 1\right] \\
\leq \mathbb{E}\left[|\mathcal{r}(X)M(X, Y(1))\|\hat{\theta}(X) - \theta^*(X)\|_2^2 \mid T = 1\right] = M \cdot \|\hat{\theta} - \theta\|_2^2|_{L_2(P_X \mid T = 0)}.
$$

Returning to our problem, we note that due to unbiasedness,

$$
\frac{1}{|T_1|} \sum_{i \in T_1} r(X_i)\left(\mathcal{H}(X_i, Y_i) - H(X_i, Y_i)\right) = O_P\left(\|\mathcal{r}X(1)\|_{L_2(P_X \mid T = 1)}^2 / \sqrt{n}\right),
$$

where by the given conditions, we have

$$
\|\mathcal{r}(X)\mathcal{H}(X, Y(1)) - H(X, Y(1))\|_{L_2(P_X \mid T = 1)} = O\left(\|\hat{\theta} - \theta\|_{L_2(P_X \mid T = 1)}\right) = o_P(1).
$$

As a result, we have

$$
\frac{1}{|T_1|} \sum_{i \in T_1} r(X_i)\left(\mathcal{H}(X_i, Y_i) - H(X_i, Y_i)\right) \leq o_P(1/\sqrt{n}) + M\|\hat{\theta} - \theta\|_2^2|_{L_2(P_X \mid T = 0)} = o_P(1/\sqrt{n}).
$$

Putting these pieces together, we conclude the proof of $\hat{\mu}_{1,0} = \hat{\mu}_{1,0}^* + o_P(1/\sqrt{n})$ for each $j$. Therefore, averaging over the three folds, we have

$$
\sqrt{n}(\hat{\mu}_{1,0} - \mu_{1,0}) = \frac{1}{n_1} \sum_{i \in T_1} \left(r_{1,0}(X_i)\left(\mathcal{H}(X_i, Y_i) - h(X_i)\right) - h(X_i)\right) + \frac{1}{n_0} \sum_{i \in T_0} h(X_i) + o_P(1/\sqrt{n}),
$$

which, by CLT and Slutsky’s theorem, converges in distribution to $N(0, \sigma^2)$. Here $n_1$ is the total number of treated samples, and $n_0$ is the number of control samples. The asymptotic variance is

$$
\sigma^2 = \frac{1}{p_1} \text{Var} \left(r_{1,0}(X)\left(\mathcal{H}(X, Y(1)) - h(X)\right) \mid T = 1\right) + \frac{1}{p_0} \text{Var} \left(h(X) \mid T = 0\right).
$$

where $p_1 = \mathbb{P}(T = 1)$, $p_0 = \mathbb{P}(T = 0)$ and all the expectations (variances) are induced by the observed distribution.

It now remains to show that $\hat{\sigma}^2 \rightarrow \sigma^2$ in the definition of Theorem 3. As $\hat{p}_1 \rightarrow p_1$, $\hat{p}_0 \rightarrow p_0$, by the law of large numbers, it suffices to show that $\frac{1}{n_1} \sum_{i \in T_1} (d_{1,i}^d - (d_{1,i}^d)^2) = o_P(1)$ and $\frac{1}{n_0} \sum_{i \in T_0} (d_{1,i}^d - (d_{1,i}^d)^2) = o_P(1)$ and similar for $(d_{0,i}, (d_{0,i}^*)^2)$, where we define the oracle counterparts $d_{1,i}^* = r_{1,0}(X_i)\left(\mathcal{H}(X_i, Y_i) - h(X_i)\right)$, $d_{0,i}^* = h(X_i)$.

By Cauchy-Schwarz inequality,

$$
\frac{1}{n_1} \sum_{i \in T_1} (d_{1,i}^d - (d_{1,i}^d)^2) = \frac{1}{n_1} \sum_{i \in T_1} (d_{1,i} - (d_{1,i}^d)^2) + \frac{1}{n_1} \sum_{i \in T_1} 2(d_{1,i} - d_{1,i}^d) d_{1,i}^* \\
\leq \frac{1}{n_1} \sum_{i \in T_1} (d_{1,i} - (d_{1,i}^d)^2) + \sqrt{\frac{1}{n_1} \sum_{i \in T_1} (d_{1,i} - (d_{1,i}^d)^2)^2} \cdot \sqrt{\frac{1}{n_1} \sum_{i \in T_1} (d_{1,i}^d)^2}.
$$

Focusing on the summation within $\mathcal{T}_1$, we have

$$
\frac{1}{|\mathcal{T}_1|} \sum_{i \in \mathcal{T}_1} (d_{1,i} - (d_{1,i}^d)^2) = O_P\left(\left\|\mathcal{r}^d(\mathcal{H}(j) - \mathcal{H}(j)) - r_{1,0}(H - h)\right\|_{L_2(P_X \mid T = 1)}^2\right),
$$

where the right-hand side is $o_P(1)$ under the conditions of Theorem 3. Other folds and other summation terms follow similar arguments hence $\hat{\sigma}^2 \rightarrow \sigma^2$. By Slutsky’s lemma, we conclude the proof of Theorem 3. $\square$
C.5 Proof of Theorem 4

Proof of Theorem 4. The proof follows exactly the same arguments as the proof of Theorem 3 with \( \theta^0 \) in place of \( \theta^* \); where all the errors are controlled in the same way; the only difference is to show that

\[
\mathbb{E}[r(X) \nabla_\theta \ell(\theta^0(X), X, Y(1)) | \theta = \theta^0(X)] | T = 1
\]

in parallel with (15) in the proof of Theorem 3. We note that this is directly implied by our local condition. Therefore, under the conditions of Theorem 4,

\[
\hat{\mu} - \frac{1}{n_1} \sum_{i \in I_1} r_{1,0}(X_i) \left[ H^0(X_i, Y_i(1)) - h^0(X_i) \right] - \frac{1}{n_0} \sum_{i \in I_0} h^0(X_i).
\]

The terms in the summation has expectation

\[
\mu^0_{1,0} := -\mathbb{E}[r_{1,0}(X_i) H^0(X_i, Y_i(1)) | T = 1].
\]

Since \( \alpha^*(x), \eta^*(x) \) is the per-x minimizer of \( \mathbb{E}[\ell(\theta, X, Y(1)) | X = x, T = 1] \), we have

\[
\mathbb{E}[H^0(X_i, Y_i(1)) | X = x, T = 1] \geq \mathbb{E}[H(X_i, Y_i(1)) | X = x, T = 1]
\]

for \( \mathbb{P}_X | T = 1 \)-almost all \( x \), hence by tower property, we have \( \mu^0_{1,0} \leq \mu^-_{1,0} \). On the other hand, the consistency of \( \hat{\sigma}^2 \) to \( \text{Var}(\phi^0_{1,0}(X, Y, T)) \) also follows the same arguments as the proof of Theorem 3 with \( \theta^0 \), which concludes our proof of Theorem 4.

\[\square\]

D Technical lemmas

Lemma 3. Let \( F_n \) be a sequence of \( \sigma \)-algebras, and let \( A_n \geq 0 \) be a sequence of nonnegative random variables. If \( \mathbb{E}[A_n | F_n] = o_P(1) \), then \( A_n = o_P(1) \).

Proof of Lemma 3. By Markov’s inequality, for any \( \epsilon > 0 \), we have

\[
B_n := \mathbb{P}(A_n > \epsilon | F_n) \leq \frac{\mathbb{E}[A_n | F_n]}{\epsilon} = o_P(1),
\]

and \( B_n \in [0, 1] \) are bounded random variables. For any subsequence \( \{n_k\}_{k \geq 1} \) of \( \mathbb{N} \), since \( B_{n_k} \xrightarrow{P} 0 \), there exists a subsequence \( \{n_{k_i}\}_{i \geq 1} \subset \{n_k\}_{k \geq 1} \) such that \( B_{n_{k_i}} \xrightarrow{a.s.} 0 \) as \( i \to \infty \). By the dominated convergence theorem, we have \( \mathbb{E}[B_{n_{k_i}}] \to 0 \), or equivalently, \( \mathbb{P}(A_{n_{k_i}} > \epsilon) \to 0 \). Therefore, for any subsequence \( \{n_k\}_{k \geq 1} \) of \( \mathbb{N} \), there exists a subsequence \( \{n_{k_i}\}_{i \geq 1} \subset \{n_k\}_{k \geq 1} \) such that \( A_{n_{k_i}} \xrightarrow{P} 0 \) as \( i \to \infty \). By the arbitrariness of \( \{n_k\}_{k \geq 1} \), we know \( A_n \xrightarrow{P} 0 \) as \( n \to \infty \), which completes the proof. \[\square\]