Supplementary Material of "Sensitivity Analysis under the f-Sensitivity Models: A Distributional Robustness Perspective"

Ying Jin*[1], Zhimei Ren*[1], and Zhengyuan Zhou[3]

[1]Department of Statistics and Data Science, the Wharton School, University of Pennsylvania
[3]Stern School of Business, New York University

A Deferred details and discussions

A.1 Discussion on related and future directions

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.

Sensitivity analysis. In this paper, we have focused on conducting inference on the counterfactuals and treatment effects under the (f, ρ)-selection condition, with a prescribed confounding parameter ρ. 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 ρ, and construct a level $1-\alpha$ confidence interval $\widehat{C}(\rho)$ for the ATE using the method introduced in this paper at each value of ρ; then let $\widehat{\rho}$ be the smallest ρ such that $C(\rho)$ contains zero. We can interpret the results as "either there is a nonzero ATE, or there is a confounder as large as $\widehat{\rho}$ to explain away the observed treatment effects".

More rigorously, let ρ^* denote the true confounding level and suppose the constructed confidence intervals $\widehat{C}(\rho)$ are nested in ρ, in the sense that $\widehat{C}(\rho_1) \subset \widehat{C}(\rho_2)$ for any $\rho_1 \le \rho_2$. We then have

$$\limsup_{n\to\infty} \mathbb{P}(\text{ATE } = 0, \rho^* < \widehat{\rho}) \le \limsup_{n\to\infty} \mathbb{P}(\text{ATE } \notin \widehat{C}(\rho^*)) \le \alpha,$$

if $\widehat{C}(\rho^*)$ is an asymptotically valid confidence interval for the ATE. In words, when the ATE is indeed zero, $\widehat{\rho}$ is an asymptotic level-$(1-\alpha)$ confidence lower bound for ρ^*. Similar to the case of Jin et al. (2023), here only point-wise validity is necessary, i.e., we only need our CIs to be asymptotically valid for each fixed ground truth of ρ. Finally, we note that the monotonicity of the confidence intervals is satisfied with a reasonable estimation procedure; in addition, given a sequence of preliminary confidence intervals $\widehat{C}(\rho)$ that are not necessarily nested, one can enforce monotonicity while maintaining validity via

$$\widehat{C}^{\text{nest}}(\rho) := \cup_{\rho' \le \rho} \widehat{C}(\rho').$$

It is straightforward to see that $\widehat{C}^{\text{nest}}(\rho)$'s are nested, and they are still asymptotically valid since $\widehat{C}(\rho) \subseteq \widehat{C}^{\text{nest}}(\rho)$ for every $\rho \ge 0$.

*Author names listed alphabetically.

Calibrating the sensitivity parameter. It is impossible to estimate the confounding strength ρ without additional information. However, in practice, a reasonable approach for guessing the magnitude of confounding is to pretend an observed covariate is a missing confounder, and use the leave-one-out approach to estimate the resulting deviation as ρ. This can be done for both conditional and marginal f-sensitivity models. To be specific, let U be the one observed variable that is left out, and X be the remaining covariates, and suppose data are $\{X_i, Y_i, T_i\}_{i=1}^n$. One can fit propensity score models $\widehat{e}(x,u)$ for $\mathbb{P}(T=1 \,|\, X=x, U=u)$ and $\widehat{e}_0(x)$ for $\mathbb{P}(T=1 \,|\, X=x)$, and then set

$$\widehat{\mathrm{OR}}(x,u) = \frac{\widehat{e}_0(x)/(1-\widehat{e}_0(x))}{\widehat{e}(x,u)/(1-\widehat{e}(x,u))}.$$

Then, for the conditional f-sensitivity model (consider $\mathcal{M}_1^{f,\rho}$), we can set

$$\widehat{\rho} = \sup_{T_i=1} \widehat{f}(X_i),$$

where $\widehat{f}(x)$ fits $\mathbb{E}[f(\widehat{\mathrm{OR}}(x,U)) \,|\, X=x, T=1]$. For the marginal f-sensitivity model (consider $\mathcal{M}_{1,\mathrm{mgn}}^{f,\rho}$), we can set

$$\widehat{\rho} = \frac{1}{n_1} \sum_{T_i=1} f(\widehat{\mathrm{OR}}(X_i, U_i)).$$

Note that the accuracy of these estimators depends on the estimation quality of $\widehat{e}_0$, $\widehat{e}$, and $\widehat{f}$. Analyzing the properties of these estimators are beyond the scope of this paper, and we suggest using these estimators as an exploratory analysis. In addition, since $\widehat{\rho}$ for the marginal f-sensitivity model requires less functional estimation, it might be easier to calibrate than the conditional f-sensitivity model.

Conditional average treatment effect (CATE). The methodology proposed in this paper also provides bounds on CATE under the conditional (f,ρ)-selection condition (however, the situation with the marginal (f,ρ)-selection condition is more complicated). Concretely, following the proof of Propositions 2 and 3, a lower bound for $\mathbb{E}[Y(1) \,|\, X=x, T=1]$ is given by the optimal value of

$$\begin{aligned} \min_{L(x,y)\geq 0 \text{ measurable}} \quad & \mathbb{E}\big[Y(1)L(x,Y(1)) \,|\, X=x, T=1\big] \\ s.t. \quad & \mathbb{E}\big[L(x,Y(1)) \,|\, T=1, X=x\big], \\ & \mathbb{E}\bigg[f\Big(\frac{L(x,Y(1))}{r_{1,0}(x)}\Big) \,|\, X=x, T=1\bigg] \leq \rho. \end{aligned}$$

The dual form of the above optimization problem is

$$\sup_{\alpha\geq 0, \eta\in\mathbb{R}} \; -r_{1,0}(x) \cdot \mathbb{E}\bigg[\alpha f^*\Big(-\frac{Y(1)+\eta}{\alpha}\Big) + \eta + \alpha\rho \,\Big|\, X=x, T=1\bigg]. \tag{1}$$

Note that the optimizer $(\alpha^*(x), \eta^*(x))$ defined in (3) is exactly the optimizer of (1). In fact, $\widehat{\mu}(x) := \widehat{r}^{(j)}(x)\widehat{h}^{(j)}(x)$ where $\widehat{r}^{(j)}, \widehat{h}^{(j)}$ are defined in Algorithm 1 is an estimator for the optimal objective in (1). 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.

A.2 Discussions on Assumption 2 on sieve estimation

We provide additional discussion on Assumption 2 for sieve estimators in the context of $(X, Y(1)) \,|\, T=1$. 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 χ^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

$$\big(\alpha^*(x), \eta^*(x)\big) = \operatorname{argzero} \nabla_{a,b}\bigg\{a\mathbb{E}\Big[f^*\big(\tfrac{Y(1)+b}{-a}\big)\,\Big|\, X=x, T=1\Big] + b + a\rho\bigg\},$$

where the right-hand side takes the form

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

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 $\mathbb{P}_{Y(1)\,|\,X=x,T=1}$ is smooth; for example, for some $h \in \mathcal{X}$, $\mathbb{P}_{Y(1)\,|\,X=x+th,T=1} = \mathbb{P}_{Y(1)\,|\,X=x,T=1} + t\cdot\mathbb{P}_h$ for some measure $\mathbb{P}_h$ 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 $Y(1)\,|\,X=x,T=1$ 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 is 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

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

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

$$\begin{aligned}
\nabla_a^2\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big] &= \frac{1}{a^3}\mathbb{E}\Big[\big(Y(1)+b\big)^2 e^{\frac{Y(1)+b}{-a}-1}\,\Big|\,X=x,T=1\Big],\\
\nabla_b^2\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big] &= \frac{1}{a}\mathbb{E}\Big[e^{\frac{Y(1)+b}{-a}-1}\,\Big|\,X=x,T=1\Big],\\
\nabla_{a,b}^2\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big] &= -\frac{1}{a^2}\mathbb{E}\Big[\big(Y(1)+b\big)e^{\frac{Y(1)+b}{-a}-1}\,\Big|\,X=x,T=1\Big].
\end{aligned}$$

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 $\mathbb{P}_{Y(1)\,|\,X=x,T=1}$ 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 $|\ell(\theta,x,y) - \ell(\theta^*,x,y)| \le \bar{\ell}(x,y)\|\theta(x)-\theta^*(x)\|_2$ for $\|\theta(x)-\theta^*(x)\|_2 < \epsilon$ for some sufficiently small $\epsilon > 0$, where $\|\theta(x)-\theta^*(x)\|_2$ is the Euclidean norm, and $\sup_{x\in\mathcal{X}}\mathbb{E}[\bar{\ell}(x,Y(1))^2\,|\,X=x,T=1] < M$ for some constant $M>0$. By Taylor expansion, we have

$$\ell(\theta,x,y) - \ell(\theta^*,x,y) = \nabla_\theta\ell(\widetilde{\theta},x,y)\big(\theta^*(x)-\theta(x)\big),$$

where $\widetilde{\theta}(x)$ lies between $\theta(x)$ and $\theta^*(x)$. We note that ∇_θ is also a smooth function of θ, and the gradient is uniform bounded for $\theta(x)$ within a neighborhood of $\theta^*(x)$ in terms of Euclidean L_2-norm. In particular,

$$\frac{\partial}{\partial a}\ell((a,b),x,y) = \big(1+\tfrac{y+b}{a}\big)e^{\frac{y+b}{-a}-1} + \rho, \quad \frac{\partial}{\partial b}\ell((a,b),x,y) = 1 - e^{\frac{y+b}{-a}-1}.$$

For any $\|(a,b)-\theta^*(x)\|_2 \le \epsilon$ for sufficiently small ϵ, 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 $\mathbb{E}[\ell(\theta,X,Y(1))-\ell(\theta^*,X,Y(1))\,|\,T=1] \le C_1\|\theta-\theta^*\|^2_{L_2(\mathbb{P}_{\cdot\,|\,T=1})}$ when $\theta \in \Lambda^p_c(\mathcal{X})\times\Lambda^p_c(\mathcal{X})$ and $\|\theta-\theta^*\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})}$ is sufficiently small. Similar to arguments in the proof of Theorem 1, sufficiently small $\|\theta-\theta^*\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})}$ implies sufficiently small $\|\theta-\theta^*\|_\infty$ for this function class. Therefore, we can consider $\theta \in \Lambda^p_c(\mathcal{X})\times\Lambda^p_c(\mathcal{X})$ 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

$$\begin{aligned}&\mathbb{E}\big[\ell(\theta,x,Y(1))\,\big|\,X=x,T=1\big]-\mathbb{E}\big[\ell(\theta^*,x,Y(1))\,\big|\,X=x,T=1\big]\\&=1/2\cdot\nabla^2_\theta\mathbb{E}\big[\ell(\widetilde{\theta},x,Y(1))\,\big|\,X=x,T=1\big][\theta(x)-\theta*(x),\theta(x)-\theta^*(x)]\end{aligned}$$

since the gradient is zero, where $\widetilde{\theta}(x)$ 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.

χ^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 $[-2,\infty)$ and zero on $(-\infty,-2]$, with continuous gradient $(f^*)'(y)=(\frac{s}{2}+1)_+$, 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

$$\ell(\theta,x,y)=\frac{\alpha(x)}{4}\Big[\Big(\frac{y+\eta(x)}{-\alpha(x)}+1\Big)_+^2-1\Big]+\eta(x)+\alpha(x)\rho.$$

Assuming $Y(1)$ does not have point measure, the differentiation and expectation are exchangeable, and

$$\begin{aligned}\nabla^2_a\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big]&=\frac{1}{2a^3}\mathbb{E}\Big[\big(Y(1)+b\big)^2\mathbb{1}\{\tfrac{Y(1)+b}{-a}>-2\}\,\Big|\,X=x,T=1\Big],\\\nabla^2_b\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big]&=\frac{1}{2a}\mathbb{E}\Big[\mathbb{1}\{\tfrac{Y(1)+b}{-a}>-2\}\,\Big|\,X=x,T=1\Big],\\\nabla^2_{a,b}\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big]&=-\frac{1}{2a^2}\mathbb{E}\Big[\big(Y(1)+b\big)\mathbb{1}\{\tfrac{Y(1)+b}{-a}>-2\}\,\Big|\,X=x,T=1\Big].\end{aligned}$$

Also, the gradient is given by

$$\begin{aligned}\nabla_a\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big]&=\mathbb{E}\Big[\big(\tfrac{Y(1)+b}{-2a}+1\big)_+^2-1+\tfrac{Y(1)+b}{a}\big(\tfrac{Y(1)+b}{-2a}+1\big)_+\,\Big|\,X=x,T=1\Big]+\rho,\\\nabla_b\mathbb{E}\big[\ell((a,b),x,Y(1))\,\big|\,X=x,T=1\big]&=-\mathbb{E}\Big[\big(\tfrac{Y(1)+b}{-2a}+1\big)_+\,\Big|\,X=x,T=1\Big]+1,\end{aligned}$$

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 $(Y(1)+\eta^*(x)-c(x))\mathbb{1}\{\frac{Y(1)+\eta^*(x)}{-a}>-2\}=0$ 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 $Y(1)\,|\,X=x$ 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 minimial 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 2 for χ^2-divergence.

Practical considerations. 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.

A.3 Intuitions for the debiasing technique

In this part, we explain some high-level intuitions for the debiasing technique introduced in Section 3.3 which enables fast-rate inference in Section 4.2 with slow-rate estimators.

In view of the duality form in Proposition 3, with an estimator $(\widehat{\alpha}(\cdot), \widehat{\eta}(\cdot))$ for the optimizer (α^*, η^*), a natural plug-in estimator for the optimal objective is

$$\widehat{\mu}_{1,0}^{+,\text{naive}} = \frac{1}{n_1} \sum_{T_i=1} \widehat{r}(X_i) \widehat{H}(X_i, Y_i(1)), \quad \widehat{H}(x,y) = \widehat{\alpha}(x) f^*\Big(\frac{y+\widehat{\eta}(x)}{-\widehat{\alpha}(x)}\Big) + \widehat{\eta}(x) + \widehat{\alpha}(x)\rho.$$

However, the convergence rate for nonparametric estimation of $\widehat{\alpha}(\cdot)$, $\widehat{\eta}(\cdot)$, and $\widehat{r}(\cdot)$ is typically slower than $n^{-1/2}$. That is, we usually have $\|\widehat{H} - H\|_{L_2} = O_P(n^{-\beta})$ for $\beta < 1/2$, where $H(x,y)$ is the function with true (α^*, η^*). Such a bias will lead to a bias in $\widehat{\mu}_{1,0}^{+,\text{naive}}$ which is larger than the typicall $O(n^{-1/2})$ standard deviation, and further violate the Wald-type inference based on $\widehat{\mu}_{1,0}^{+,\text{naive}}$ since it is no longer asymptotically root-n normal.

We will utilize covariates X_i for $T_i = 0$ to remove the first-order bias in $\widehat{\mu}_{1,0}^{+,\text{naive}}$. Assume also that $\widehat{h}$ is an external estimator of $\bar{h}(x) := \mathbb{E}[\widehat{H}(x, Y(1)) \,|\, X = x, T = 1]$. For simplicity, assume $\widehat{r}$, $\widehat{\alpha}$ and $\widehat{\eta}$ are given and independent of all the data, and so is $\widehat{H}$. The bias in $\widehat{\mu}_{1,0}^{+,\text{naive}}$ is

$$\widehat{\mu}_{1,0}^{+,\text{naive}} - \widehat{\mathbb{E}}_1[rH] = \widehat{\mathbb{E}}_1[\widehat{r} \cdot (\widehat{H} - \bar{h})] + \widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] - \widehat{\mathbb{E}}_1[r_{1,0}H],$$

where $\widehat{\mathbb{E}}_1[\cdot]$ is the empirical mean in the treated group. The first term $\widehat{\mathbb{E}}_1[r(\widehat{H} - \bar{h})]$ is mean zero (unbiased) since $\bar{h}$ is the true conditional mean function of $\widehat{H}(X, Y(1))$, and thus negligible. Due to the optimality of α^* and η^*, as well as the convexity of H, the first-order Taylor expansion of H implies $\mathbb{E}_1[r(\widehat{H} - H)]$ is of a higher order than the estimation bias, hence also negligible. The remaining bias is thus (we use $\approx$ to denote equivalence up to higher-order error)

$$\widehat{\mu}_{1,0}^{+,\text{naive}} - \widehat{\mathbb{E}}_1[r_{1,0}H] \;\approx\; \widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] - \widehat{\mathbb{E}}_1[r_{1,0}\widehat{H}].$$

Next, the bias $\widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] - \widehat{\mathbb{E}}_1[r_{1,0}\widehat{H}]$ is removed by subtracting $\widehat{\mathbb{E}}_1[\widehat{r}\widehat{h}] - \widehat{\mathbb{E}}_0[\widehat{h}]$ from $\widehat{\mu}_{1,0}^{+,\text{naive}}$, leading to our estimator (ignoring cross-fitting). The idea is that

$$\begin{aligned}
\widehat{\mathbb{E}}_1[\widehat{r}\widehat{h}] - \widehat{\mathbb{E}}_0[\widehat{h}] \;&\approx\; \widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] + \widehat{\mathbb{E}}_1[\widehat{r}(\widehat{h} - \bar{h})] - \widehat{\mathbb{E}}_0[\widehat{h}] \\
&\approx\; \widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] + \widehat{\mathbb{E}}_1[r_{1,0}(\widehat{h} - \bar{h})] - \widehat{\mathbb{E}}_1[r_{1,0}\widehat{h}] \\
&\approx\; \widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] - \widehat{\mathbb{E}}_1[r_{1,0}\bar{h}] \;\approx\; \widehat{\mathbb{E}}_1[\widehat{r}\bar{h}] - \widehat{\mathbb{E}}_1[r_{1,0}\widehat{H}].
\end{aligned}$$

In the second line, we use the fact that $\widehat{\mathbb{E}}_1[(\widehat{r} - r_{1,0})(\widehat{h} - \bar{h})]$ is of a higher order. Also, $r_{1,0}(x)$ is the covariate shift between treated and control groups, so that for any fixed function $f(x)$, it holds that $\mathbb{E}[r_{1,0}(X)f(X) \,|\, T = 1] = \mathbb{E}[f(X) \,|\, T = 0]$.

Finally, as all nuisance component estimators have to be computed from data, we use the cross-fitting technique to decouple the estimators and the data they are applied to.

A.4 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 $\mathbb{E}[Y(1) \,|\, T = 0]$:** Let $-\widehat{\mu}_{1,0}^{+}$ be the estimator obtained from the procedure in Section 3.3 with $-Y(1)$ replacing $Y(1)$. Then $\sqrt{n}(\widehat{\mu}_{1,0}^{+} - \mu_{1,0}^{+}) \rightsquigarrow N(0, \text{Var}(\phi_{1,+}(X,Y,T)))$, with influence function

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

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

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

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

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

(c) **Upper bound of $\mathbb{E}[Y(0)\,|\,T=1]$:** Let $-\widehat{\mu}_{0,1}^{+}$ 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 $\sqrt{n}(\widehat{\mu}_{0,1}^{+}-\mu_{0,1}^{+}) \rightsquigarrow N(0,\mathrm{Var}(\phi_{0,+}(X,Y,T)))$ with influence function

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

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

B Sensitivity analysis under the marginal f-sensitivity model

This section collects the identification, estimation and inference for treatment effects under the marginal f-sensitivity model introduced in Section 2.4.

B.1 Identification results

Parallel to Proposition 1, below we show that the marginal (f,ρ)-selection condition implies a bounded average f-divergence between $\mathbb{P}_{X,Y(1)\,|\,T=0}$ and $\mathbb{P}_{X,Y(1)\,|\,T=1}$, averaged over $\mathbb{P}_{X\,|\,T=1}$. Switching the role of treated and control groups leads to a similar result for distribution shifts under $\mathcal{M}_{0,\mathrm{mgn}}^{f,\rho}$. Its proof is also similar to that of Proposition 1, and we defer it to Appendix D.1.

Proposition B.1. *Let $\mathbb{P}$ be the true unknown joint distribution over $(X,U,T,Y(0),Y(1))$, which induces $\mathbb{P}_{X,Y,T}$, the joint distribution of (X,Y,T) where $Y=Y(T)=TY(1)+(1-T)Y(0)$. Let $\mathcal{P}$ be the set of all distributions over $(X,Y(0),Y(1),U,T)$ and $\mathcal{P}_{X,Y(1)}$ be the set of all distributions over $(X,Y(1))$. Define $\mathcal{Q}_{1,0}^m$ as the ambiguity set of all counterfactual distributions that agree with the observables and satisfy marginal (f,ρ)-selection for the treated:*

$$\mathcal{Q}_{1,0}^m=\big\{QX,Y(1)\,|\,T=0\colon Q\in\mathcal{P},\ Q_{X,Y,T}=\mathbb{P}_{X,Y,T},\ Q\in\mathcal{M}_{1,\mathrm{mgn}}^{f,\rho}\big\}$$

Then $\mathbb{P}_{X,Y(1)\,|\,T=0}\in\mathcal{Q}_{1,0}^m$, and a sharp characterization of $\mathcal{Q}_{1,0}^m$ is given by

$$\mathcal{Q}_{1,0}^m=\left\{Q\colon \frac{\mathrm{d}Q_X}{\mathrm{d}\mathbb{P}_{X\,|\,T=1}^{\mathrm{obs}}}(x)=r_{1,0}(x),\ \int_x D_f\big(Q_{Y\,|\,X=x}\big\|\mathbb{P}_{Y\,|\,X=x,T=1}^{\mathrm{obs}}\big)\,\mathrm{d}\mathbb{P}^{\mathrm{obs}}(x\,|\,T=1)\le\rho\right\}.$$

B.2 Convex optimization characterization

Parallel to Proposition 2, below we characterize the partial identification bounds for the counterfactual mean $\mathbb{E}[Y(1)\,|\,T=0]$ under $\mathcal{M}_{1,\mathrm{mgn}}^{f,\rho}$ via an optimization program.

Proposition B.2. *Let $\mu_{1,0}^{m+}$ (resp. $\mu_{0,1}^{m-}$) be the optimal value of the convex optimization problem:*

$$\begin{aligned}\min(\textit{resp. }\max)_{L(x,y)\text{ measurable}}\ &\mathbb{E}\big[Y(1)L(X,Y(1))\,\big|\,T=1\big]\\ \text{s.t. }&\mathbb{E}[L(X,Y(t))\,|\,X=x,T=1]=r_{1,0}(x),\quad\forall x\\ &\mathbb{E}\big[f\big(L(X,Y(1))/r_{1,0}(X)\big)\,\big|\,T=1\big]\le\rho,\end{aligned}\tag{2}$$

Then $\mu_{1,0}^{m-} \le \mathbb{E}[Y(1) \,|\, T=0] \le \mu_{1,0}^{m+}$ *if* $\mathbb{P} \in \mathcal{M}_{1,\mathrm{mgn}}^{f,\rho}$.

The dual formulation is provided in the following proposition as preparation for our estimation procedure, whose proof is in Appendix D.2.

Proposition B.3. *The solution to* (2) *is given by*

$$\mu_{1,0}^{m,-} = -\inf_{\alpha \ge 0, \eta(\cdot)} \mathbb{E}\left[\alpha f^*\left(\frac{r_{1,0}(X)}{-\alpha}\big(Y(1)+\eta(X)\big)\right) + \eta(X) r_{1,0}(X) + \alpha\rho \,\middle|\, T=1 \right], \tag{3}$$

where $f^*(s) = \sup_{t\ge 0}\{st - f(t)\}$ *is the conjugate function of* f.

B.3 Sensitivity analysis with stable risk minimizer

We then discuss estimation and inference strategies under the marginal f-sensitivty model. At a high level, we will still follow a plug-in-then-debias strategy. However, the specific debiasing technique and the assumptions required for consistent estimation and valid inference will differ from previous ones. We first introduce the procedure.

B.3.1 The procedure

We start by dividing both the treated group and the control group into four equal-sized folds and use the notation $\mathcal{I}_1^{(j)}$ and $\mathcal{I}_0^{(j)}$ to denote the resulting folds, $j = 1, \ldots, 4$. In view of (3), for any $a \in \mathbb{R}^+$, $b, r \in \mathbb{R}$, we denote the generic risk function as

$$m(x, y, a, b, r) = a f^*\Big(\frac{r}{-a}\big(y+\eta\big)\Big) + br + a\rho.$$

For each fold $j = 1, \ldots, 4$, we use $\mathcal{I}_1^{(j+1)}$ and $\mathcal{I}_0^{(j+1)}$ to obtain an estimator $\widehat{r}^{(j)}$ for $r_{1,0}$; as $r_{1,0}$ is decided by the propensity score $e(x) = \mathbb{P}(T=1 \,|\, X=x)$, there are many methods in the literature that can be applied. We then use $\mathcal{I}_1^{(j+2)}$ to obtain the empirical risk minimizer

$$\big(\widehat{\alpha}^{(j)}, \widehat{\eta}^{(j)}\big) = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}, \eta\in\Theta_n} \frac{1}{|\mathcal{I}_1^{(j+2)}|} \sum_{i\in\mathcal{I}_1^{(j+2)}} m\big(X_i, Y_i, \alpha, \eta(X_i), \widehat{r}^{(j)}(X_i)\big),$$

where Θ_n is some function class that might grow with n. This estimation procedure is the same for the preceding naive estimator, whose convergence has been discussed in Section B.3.2.

We assume throughout that f^* is differentiable, and define

$$H^{(j)}(x, y) = \frac{\partial f^*}{\partial r}\big(x, y, \widehat{\alpha}^{(j)}, \widehat{\eta}^{(j)}(x), \widehat{r}^{(j)}(x)\big).$$

We then use the fold $\mathcal{I}_1^{(j+3)}$ to obtain an estimator $\widehat{h}^{(j)}(\cdot)$ for the conditional mean function

$$\bar{h}^{(j)}(x) = \mathbb{E}\big[H^{(j)}(x, Y(1)) \,\big|\, X=x, T=1\big],$$

viewing $H^{(j)}$ as fixed. Finally, we define the adjusted estimator on folds $\mathcal{I}_1^{(j)}$ and $\mathcal{I}_0^{(j)}$ as

$$\widehat{\mu}_{1,0}^{m,(j)} = \frac{1}{|\mathcal{I}_1^{(j)}|} \sum_{i\in\mathcal{I}_1^{(j)}} \Big\{ m\big(X_i, Y_i, \widehat{\alpha}^{(j)}, \widehat{\eta}^{(j)}(X_i), \widehat{r}^{(j)}(X_i)\big) - \widehat{\alpha}^{(j)} \widehat{r}^{(j)}(X_i) \widehat{h}^{(j)}(X_i) \Big\} + \frac{\widehat{\alpha}^{(j)}}{|\mathcal{I}_0^{(j)}|} \sum_{i\in\mathcal{I}_0^{(j)}} \widehat{h}^{(j)}(X_i).$$

The final estimator is the average of the estimator obtained from each fold:

$$\widehat{\mu}_{1,0}^{m,-} = -\frac{1}{4}\sum_{j=1}^{4} \widehat{\mu}_{1,0}^{m,(j)}. \tag{4}$$

B.3.2 Estimation of $\widehat{\alpha}$ and $\widehat{\eta}$

The estimation of $\widehat{\alpha} \in \mathbb{R}^+$ and $\widehat{\eta}\colon \mathcal{X} \to \mathbb{R}$ can be done in a similar fashion as Section 3.4. To be specific, we still employ the sieve estimator for (2), and let $\Theta_n = \mathbb{R}^+ \times \Theta_n^\eta$ where Θ_n^η is the function class in Examples 2 or 3. We define the population risk minimizer given the estimator $\widehat{r}$ as

$$(\bar{\alpha}, \bar{\eta}) = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}, \eta(\cdot)} \mathbb{E}\Big[m\big(X, Y(1), \alpha, \eta(X), \widehat{r}(X)\big) \,\Big|\, T = 1, \mathcal{I}_1^{(1)} \cup \mathcal{I}_0^{(1)}\Big].$$

We impose the following assumption on regularity conditions of the risk function under $\widehat{r}$. Here we view $\bar{m}(x, y, a, b) = m(x, y, a, b, \widehat{r}(x))$ as a fixed function; for simplicity, we write $\bar{m}(x, y, \theta) = \bar{m}(x, y, \alpha, \eta(x))$ for $\theta = (\alpha, \eta(\cdot))$ and denote $\bar{\theta} = (\bar{\alpha}, \bar{\eta}(\cdot))$, $\widehat{\theta} = (\widehat{\alpha}, \widehat{\eta}(\cdot))$, and $\theta(x) = (\alpha, \eta(x))$. We define $\bar{I}(x, a, b) = \mathbb{E}[\bar{m}(x, Y(1), a, b) \,|\, X = x, T = 1]$, which is convex in $(a, b) \in \mathbb{R}^+ \times \mathbb{R}$.

Assumption B.1. *Suppose $\mathcal{X}$ is a compact set, and $\mathbb{P}_{X \,|\, T=1}$ has positive density on $\mathcal{X}$. Suppose $\bar{\eta} \in \Theta = \mathbb{R}^+ \times \Lambda_c^p(\mathcal{X})$ for some $c > 0$. Assume $|\bar{m}(x, y, a, b) - \bar{m}(x, y, \bar{\alpha}, \bar{\eta}(x)) - \nabla_{a,b}\bar{m}(x, y, \bar{\alpha}, \bar{\eta}(x))[a - \bar{\alpha}, b - \bar{\eta}(x)]| \le g(x, y)\|(a - \bar{\alpha}, b - \bar{\eta}(x))\|_2^2$ for $\|(a - \bar{\alpha}, b - \bar{\eta}(x))\|_2 < \epsilon$ for sufficiently small $\epsilon > 0$, where $\|\cdot\|_2$ is the Euclidean norm, and $\mathbb{E}[g(x, Y(1)) \,|\, X = x, T = 1] < M$ for some constant $M > 0$. We assume that $|\bar{I}(x, a, b) - \bar{I}(x, \bar{\alpha}^{(j)}, \bar{\eta}(x)) - \nabla_{a,b}\bar{I}(x, \bar{\alpha}, \bar{\eta}(x))[a - \bar{\alpha}, b - \bar{\eta}(x)]| \ge \lambda\|(a - \bar{\alpha}, b - \bar{\eta}(x))\|_2^2$ for $\|(a - \bar{\alpha}, b - \bar{\eta}(x))\|_2 < \epsilon$ for sufficiently small $\epsilon > 0$, where $\lambda > 0$ is a constant. Also, assume $|\bar{m}(\theta, x, y) - \bar{m}(\bar{\theta}^{(j)}, x, y)| \le \bar{g}(x, y)\|\theta(x) - \bar{\theta}(x)\|_2$ for $\|\theta(x) - \bar{\theta}(x)\|_2 < \epsilon$ for sufficiently small $\epsilon > 0$, and $\sup_{x \in \mathcal{X}} \mathbb{E}[\bar{g}(x, Y(1))^2 \,|\, X = x, T = 1] < M$ for some constant $M > 0$. Furthermore, we assume the exchangeability of differentiation and expectation for $\mathbb{E}[\nabla_{a,b}\bar{m}(x, Y(1), \bar{\alpha}, \bar{\eta}(x)) \,|\, X = x, T = 1]$ and $\mathbb{E}[\nabla_{a,b}\bar{m}(X, Y(1), \bar{\alpha}, \bar{\eta}(X)) \,|\, T = 1]$.*

In Assumption B.1, we assume the smoothness of the underlying optimizer, which is reasonable if $\widehat{r}$ is smooth and the distribution of $Y(1) \,|\, X = x$ is smooth in x (c.f. the discussion in Appendix A.2 for the conditional f-sensitivity model). The stable expansion condition is also mild, which is satisfied if the risk function is second-order smooth or have a Lipschitz first-order derivative. The condition on the convex function $\bar{I}$ can be satisfied if its Hessian is positive definite (or strongly convex) at $(\bar{\alpha}, \bar{\eta}(x))$ for all x in the compact set $\mathcal{X}$. The first-order expansions of $\bar{m}$ are similar to what we require in Assumption 2; concrete examples are similar to those discussed in Appendix A.2.

The following proposition provides the convergence rate of $(\widehat{\alpha}, \widehat{\eta}(\cdot))$ to $(\bar{\alpha}, \bar{\eta}(\cdot))$ with sieve estimation under the above regularity conditions. Its proof is in Appendix D.3.

Proposition B.4. *Suppose Assumption B.1 holds. We set $J_n = (\frac{\log n}{n})^{1/(2p+d)}$ for the sieve estimators in Examples 2 or 3. Denote the empirical risk for sieve estimation with parameters θ as $\widehat{\mathbb{E}}_n\big[\bar{m}(\theta, X, Y(1))\big]$, and suppose $\widehat{\theta}$ satisfies $\widehat{\mathbb{E}}_n\big[\bar{m}(\widehat{\theta}, X, Y(1))\big] \le \inf_{\theta \in \Theta_n} \widehat{\mathbb{E}}_n\big[\bar{m}(\theta, X, Y(1))\big] - O_P((\frac{\log n}{n})^{\frac{2p}{2p+d}})$. Then $\|\widehat{\theta} - \bar{\theta}\|_{L_2(\mathbb{P}_{\cdot \,|\, T=1})} = O_P\big((\frac{\log n}{n})^{\frac{p}{2p+d}}\big)$, $\|\widehat{\theta} - \bar{\theta}\|_\infty = O_P\big((\frac{\log n}{n})^{\frac{2p^2}{(2p+d)^2}}\big)$.*

B.3.3 Theoretical guarantees

Our key additional assumption for Wald-type inference is the stability of population risk minimizers. It is needed for our analysis because estimating $\widehat{\alpha}$ and $\widehat{\eta}$ is based on an estimated $\widehat{r}$, thereby leading to optimization error compared with estimating them based on the true $r_{1,0}(x)$. In a nutshell, it ensures that the estimation error of $\widehat{\alpha}$ and $\widehat{\eta}$ is of the same order as that of $\widehat{r}$.

Assumption B.2. *For each j, it holds that $\big\|(\bar{\alpha}^{(j)}, \bar{\eta}^{(j)}(\cdot)) - (\alpha^*, \eta^*(\cdot))\big\|_{L_2(\mathbb{P}_{\cdot \,|\, T=1})} = O(\|\widehat{r}^{(j)} - r_{1,0}\|_{L_2(\mathbb{P}_{\cdot \,|\, T=1})})$, where the L_2-norm is defined viewing α and $\bar{\alpha}^{(j)}$ as constant functions.*

Under the above assumption, the correction term using $\widehat{h}^{(j)}$ helps to remove the bias incurred by $\widehat{r}^{(j)}$ in each fold, so that the errors caused by nuisance estimation is negligible compared to $n^{-1/2}$. Before presenting our inferential guarantee, we state estimation rates and some mild regularity conditions. The functions $\widehat{r}^{(j)}$, $\bar{\alpha}^{(j)}$ and $\bar{\eta}^{(j)}$ are all viewed as fixed functions when taking expectations. We also denote $\widehat{\theta}^{(j)} = (\widehat{\alpha}^{(j)}, \widehat{\eta}^{(j)})$ and $\bar{\theta}^{(j)} = (\bar{\alpha}^{(j)}, \bar{\eta}^{(j)})$, on which $L_2(\mathbb{P}_{X \,|\, T=1})$ is defined by viewing the first argument as a constant function.

Assumption B.3. *Assume* $\|\widehat{r}^{(j)} - r_{1,0}\|_{L_2(\mathbb{P}_{\cdot \mid T=1})} = o_P(n^{-1/4})$, $\|\widehat{\theta}^{(j)} - \bar{\theta}^{(j)}\|_{L_2(\mathbb{P}_{\cdot \mid T=1})} = o_P(n^{-1/4})$ *and* $\|\widehat{h}^{(j)} - \bar{h}^{(j)}\|_{L_2(\mathbb{P}_{\cdot \mid T=1})} = o_P(n^{-1/4})$ *for all* j.

For each j, *assume that* $|m(x,y,a,b,\widehat{r}^{(j)}(x)) - m(x,y,\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(x),\widehat{r}^{(j)}(x)) - \nabla_{a,b} m(x,y,\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(x),\widehat{r}^{(j)}(x))[a-\bar{\alpha}^{(j)}, b-\bar{\eta}^{(j)}(x)]| \le g(x,y)\|(a,b)-(\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(x))\|_2^2$ *for all* (a,b) *in a neighborhood of* $(\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(x)$ *when* $\|\widehat{r}^{(j)} - r_{1,0}\|_{L_2(\mathbb{P}_{X \mid T=1})}$ *is sufficiently small, where* $\mathbb{E}[g(x,Y(1))^2 \mid X=x, T=1] \le M$ *for some constant* $M > 0$. *We assume* $\mathbb{E}[(\nabla_b m(x,Y(1),\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(x),\widehat{r}^{(j)}(x))^2 \mid X=x, T=1] \le M$, $\mathbb{E}[(\nabla_a m(X,Y(1),\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(X),\widehat{r}^{(j)}(X))^2 \mid T=1] \le M$ *for some constant* $M > 0$ *and* $\mathbb{P}_{X \mid T=1}$*-almost all* x, *and the differentiation and expectation is exchangeable in* $\mathbb{E}[\nabla_b m(x,Y(1),\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(x),\widehat{r}^{(j)}(x) \mid X=x, T=1]$ *and* $\mathbb{E}[\nabla_a m(X,Y(1),\bar{\alpha}^{(j)},\bar{\eta}^{(j)}(X),\widehat{r}^{(j)}(X) \mid T=1]$. *Also assume that* $\big|f^*\big(\frac{r}{-\bar{\alpha}}\big(y+\bar{\eta}(x)\big)\big) - f^*\big(\frac{r_{1,0}(x)}{-\bar{\alpha}}\big(y+\bar{\eta}(x)\big)\big) - \frac{\partial f^*}{\partial r}\big(\frac{r_{1,0}(x)}{-\bar{\alpha}}\big(y+\bar{\eta}(x)\big)\big)(r-r_{1,0}(x))\big| \le \bar{g}(x,y)(r-r_{1,0}(x))^2$ *for* $|r - r_{1,0}(x)| < \epsilon$ *for* $\mathbb{P}_{X \mid T=1}$*-almost all* x, *where* $\epsilon > 0$ *is a sufficiently small constant, and* $\mathbb{E}[\bar{g}(x,Y(1))^2 \mid X=x, T=1] < M$ *for some constant* $M > 0$.

In Assumption B.3, we require a slow rate $o(n^{-1/4})$ of convergence for the regression problem of $\widehat{r}^{(j)}$ and $\widehat{h}^{(j)}$, which are achievable by many methods under smoothness of the underlying regression functions. The estimation of $\widehat{\theta}^{(j)}$ is the same as we discussed in Section B.3.2.

The regularity conditions we require on the smoothness of m are mild and similar to preceding ones. For example, the second order expansion can be satisfied if the first-order derivatives are locally Lipschitz or smooth, for which the uniform upper bound of $\mathbb{E}[g(x,Y(1))^2 \mid X=x, T=1]$ exists as long as the per-x expectations are finite within the compact set $\mathcal{X}$. Other stability conditions can be similarly satisfied.

We then have the following inferential guarantee for $\widehat{\mu}_{1,0}^{m,-}$. The proof of Theorem B.1 is in Appendix D.4.

Theorem B.1. *Suppose Assumptions B.2 and B.3 hold. Then the estimator* (4) *satisfies* $\sqrt{n}(\widehat{\mu}_{1,0}^{m,-} - \mu_{1,0}^{m,-}) \rightsquigarrow N(0, \mathrm{Var}(\phi_{1,-}^m(X,Y,T)))$, *where* $\mu_{1,0}^{m,-}$ *is the lower bound of* $\mathbb{E}[Y(1) \mid T=0]$ *under the marginal* (f,ρ)*-selection condition in Proposition 2, and*

$$\phi_{1,-}^m(X,Y,T) = \frac{T}{p_1}\Big\{m\big(X,Y,\alpha^*,\eta^*(X),r_{0,1}(X)\big) - \alpha^* r_{0,1}(X)h(X)\Big\} + \frac{1-T}{p_0}\alpha^* h(X),$$

with $p_1 = \mathbb{P}(T=1) = 1 - p_0$. *Here we define* $h(x) = \mathbb{E}[H(x,Y(1)) \mid X=x, T=1]$, *where* $H(x,y) = H(x,y) = \frac{\partial f^*}{\partial r}\big(x,y,\alpha^*,\eta^*(x),r_{0,1}(x)\big)$, *and* $\alpha^*, \eta^*(\cdot)$ *is the population optimizer in* (3). *Define*

$$\widehat{\sigma}_{1,0,m,-}^2 := \frac{1}{\widehat{p}_1}\left(\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} d_{1,i}^2 - \Big(\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} d_{1,i}\Big)^2\right) + \frac{1}{\widehat{p}_0}\left(\frac{1}{n_0}\sum_{i\in\mathcal{I}_0} d_{0,i}^2 - \Big(\frac{1}{n_0}\sum_{i\in\mathcal{I}_0} d_{0,i}\Big)^2\right)$$

where $\widehat{p}_1 = |\mathcal{I}_1|/n$, $\widehat{p}_0 = |\mathcal{I}_0|/n$, $d_{1,i} = m(X_i,Y_i,\widehat{\alpha}^{j[i]},\widehat{\eta}^{j[i]}(X_i),\widehat{r}^{j[i]}(X_i)) - \widehat{\alpha}^{j[i]}\widehat{r}^{j[i]}(X_i)\widehat{h}^{j[i]}(X_i)$, $d_{0,i} = \widehat{\alpha}^{(j[i])}\widehat{h}^{(j[i])}(X_i)$, *and* $j[i]$ *is the fold that sample* i *lies in. Then* $\sqrt{n}(\widehat{\mu}_{1,0}^{m,-} - \mu_{1,0}^{m,-})/\widehat{\sigma}_{1,0,m,-} \rightsquigarrow N(0,1)$.

Finally, similar to Section 4.4, the consistent naive estimators and inference procedures we introduce in this section can be applied to other quantities and combined to form bounds on average treatment effects.

C Technical proofs for conditional f-sensitivity models

C.1 Proof of Lemma 1

Proof. Proof of Lemma 1 Suppose a distribution $\mathbb{P}$ over $(X,U,T,Y(0),Y(1))$ satisfies the (f,ρ)-selection condition. We note that the likelihood ratio can be decomposed as

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}_{Y(1),U \mid X=x,T=0}}{\mathrm{d}\mathbb{P}_{Y(1),U \mid X=x,T=1}} &= \frac{\mathrm{d}\mathbb{P}_{Y(1) \mid U,X=x,T=0}}{\mathrm{d}\mathbb{P}_{Y(1) \mid U,X=x,T=1}} \cdot \frac{\mathrm{d}\mathbb{P}_{U \mid X=x,T=0}}{\mathrm{d}\mathbb{P}_{U \mid X=x,T=1}} \\ &\overset{\text{(a)}}{=} \frac{\mathrm{d}\mathbb{P}_{U \mid X=x,T=0}}{\mathrm{d}\mathbb{P}_{U \mid X=x,T=1}} = \frac{\mathbb{P}(T=0 \mid X=x,U)}{\mathbb{P}(T=1 \mid X=x,U)} \cdot \frac{\mathbb{P}(T=1 \mid X=x)}{\mathbb{P}(T=0 \mid X=x)},\end{aligned}$$

where step (a) is due to condition (1). Marginalizing over the conditional distribution of $Y(1), U$ given $X=x, T=1$ yields the first result in Lemma 1.

In addition, by condition (1) and the data-processing inequality,

$$\begin{aligned} D_f\big(\mathbb{P}_{Y(1)\,|\,X=x,T=0} \,\|\, \mathbb{P}_{Y(1)\,|\,X=x,T=1}\big) \le & D_f\big(\mathbb{P}_{Y(1),U\,|\,X=x,T=0} \,\|\, \mathbb{P}_{Y(1),U\,|\,X=x,T=1}\big) \\ = & \mathbb{E}_{Y(1),U\,|\,X=x,T=1}\left[f\Big(\frac{\mathrm{d}\mathbb{P}_{Y(1),U\,|\,X=x,T=0}}{\mathrm{d}\mathbb{P}_{Y(1),U\,|\,X=x,T=1}}\Big)\right]. \end{aligned}$$

Combining the above two facts yields

$$\begin{aligned} D_f\big(\mathbb{P}_{Y(1)\,|\,X,T=0} \,\|\, \mathbb{P}_{Y(1)\,|\,X,T=1}\big) \le & \mathbb{E}_{Y(1),U\,|\,X,T=1}\left[f\Big(\frac{\mathbb{P}(T=0\,|\,X,U)}{\mathbb{P}(T=1\,|\,X,U)}\cdot\frac{\mathbb{P}(T=1\,|\,X)}{\mathbb{P}(T=0\,|\,X)}\Big)\right] \\ = & \mathbb{E}_{U\,|\,X,T=1}\left[f\Big(\frac{\mathbb{P}(T=0\,|\,X,U)}{\mathbb{P}(T=1\,|\,X,U)}\cdot\frac{\mathbb{P}(T=1\,|\,X)}{\mathbb{P}(T=0\,|\,X)}\Big)\right] \le \rho \end{aligned}$$

almost surely, where the last inequality is due to the (f,ρ)-selection condition. □

C.2 Proof of Proposition 1

Proof. Proof of Proposition 1 We prove the claim by showing that $\mathcal{Q}_{1,0}$ and the set on the right-handed side (RHS) of Proposition 1 contain each other.

Step 1: $\mathcal{Q}_{1,0} \subseteq$ RHS.

By definition, any member of $\mathcal{Q}_{1,0}$ is the induced distribution over $(X, Y(1))$ given $T=0$ of some super-population $Q \in \mathcal{M}_1^{f,\rho}$ with $\mathbb{Q}_{X,Y,T} = \mathbb{P}_{X,Y,T}$. Thus, we write it as $\mathbb{Q}_{X,Y(1)\,|\,T=0} \in \mathcal{Q}_{1,0}$. Since $\mathbb{Q}_{X,Y,T} = \mathbb{P}_{X,Y,T}$, we know that

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

By Lemma 1, $D_f\big(\mathbb{Q}_{X,Y(1)\,|\,T=0} \,\big\|\, \mathbb{P}_{X,Y(1)\,|\,T=1}\big) \le \rho$. Also, we note that $\mathbb{Q}_{Y(1)\,|\,X,T=1} = \mathbb{Q}_{Y\,|\,X,T=1} = \mathbb{P}_{Y\,|\,X,T=1}$ since $Y = TY(1) + (1-T)Y(0)$ and $\mathbb{Q}_{X,Y,T} = \mathbb{P}_{X,Y,T}$. Combining the above, we establish the "$\subset$" direction. It remains to prove the reverse.

Step 2: RHS $\subseteq \mathcal{Q}_{1,0}$.

Given any distribution Q over $(X, Y(1))$ in RHS, we aim to find a distribution (super-population) $\mathbb{Q}$ over $(X, Y(0), Y(1), U, T)$ such that

- $(Y(1), Y(0)) \perp\!\!\!\perp T \,|\, X, U$ under $\mathbb{Q}$;
- $\mathbb{Q}_{X,T,Y} = \mathbb{P}_{X,T,Y}$;
- $\mathbb{Q} \in \mathcal{M}_1^{f,\rho}$;
- $\mathbb{Q}_{X,Y(1)\,|\,T=0}(x,y) = Q(x,y)$.

These conditions imply that $Q \in \mathcal{Q}_{1,0}$, and thus RHS $\subseteq \mathcal{Q}_{1,0}$. To construct $\mathbb{Q}$, we first set $\mathbb{Q}_{X,T} = \mathbb{P}_{X,T}$. Then we specify the distribution of $Y(1)\,|\,X,T$ via

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

So far, $\mathbb{Q}_{X,T,Y(1)}$ 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

$$\mathbb{Q}_{Y(0)\,|\,X,T,Y(1),U} = \mathbb{Q}_{Y(0)\,|\,X} = \mathbb{P}_{Y\,|\,X,T=0}.$$

Having constructed $\mathbb{Q}$, we proceed to check that it satisfies the conditions listed at the beginning of Step 2. Since $Y(1) = U$ and $\mathbb{Q}_{Y(0)\,|\,X,T,Y(1),U} = \mathbb{Q}_{Y(0)\,|\,X}$ by construction, it is easy to check that $(Y(1), Y(0)) \perp\!\!\!\perp T \,|\, X, U$. By construction, it also clearly holds that $\mathbb{Q}_{X,T,Y} = \mathbb{P}_{X,T,Y}$.

We now check that $\mathbb{Q} \in \mathcal{M}_1^{f,\rho}$. For any x, again by the construction of $\mathbb{Q}$,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}_{Y(1)\,|\,X=x,T=1}}\left[f\Big(\frac{\mathrm{d}\mathbb{Q}_{Y(1)\,|\,X=x,T=0}}{\mathrm{d}\mathbb{Q}_{Y(1)\,|\,X=x,T=1}}\Big)\right] &= \mathbb{E}_{\mathbb{P}_{Y\,|\,X=x,T=1}}\left[f\Big(\frac{\mathrm{d}Q_{Y\,|\,X=x}}{\mathrm{d}\mathbb{P}_{Y\,|\,X=x,T=1}}\Big)\right] \\ &= D_f\big(Q_{Y\,|\,X=x} \,\|\, \mathbb{P}_{Y\,|\,X=x,T=1}\big) \le \rho,\end{aligned}$$

where the last inequality is due to the fact that Q is in the set on the RHS. Therefore, we verify that $\mathbb{Q} \in \mathcal{M}_1^{f,\rho}$. By construction, $\mathbb{Q}_{Y(1)\,|\,X,T=0} = Q_{Y\,|\,X}$. It remains to show that $\mathbb{Q}_{X\,|\,T=0} = Q_X$. For any measurable set A,

$$\begin{aligned}\mathbb{Q}(X \in A\,|\,T=0) =& \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{Q}_{X\,|\,T=0}}{\mathrm{d}\mathbb{Q}_{X\,|\,T=1}} \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\right] = \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{Q}_{X\,|\,T=0}}{\mathrm{d}\mathbb{Q}_{X\,|\,T=1}} \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\right] \\ =& \mathbb{E}_{\mathbb{Q}}\Big[r_{1,0}(X) \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\Big] = \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}Q_X}{\mathrm{d}\mathbb{Q}_{X\,|\,T=1}} \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\right] = Q(X \in A).\end{aligned}$$

Since the above holds for any measurable set A, we know that $\mathbb{Q}_{X\,|\,T=0} = Q_X$. Finally, switching the role of the treated and control groups yields similar results for the counterfactual distribution of $(X, Y(0))$ given $T=1$ under $\mathcal{M}_0^{f,\rho}$. This completes the proof of Proposition 1. □

C.3 Proof of Proposition 3

Proof. Proof of Proposition 3 We first claim that solving (5) amounts to solving the following problem for each x:

$$\begin{aligned}\min_{L(x)\ \text{measurable}}\ & \mathbb{E}[Y(1)L(x)\,|\,X=x,T=1] \\ \text{s.t. }\ & \mathbb{E}[L(x)\,|\,X=x,T=1] = r_{1,0}(x) \\ & \mathbb{E}[f(L(x)/r_{1,0}(x))\,|\,X=x,T=1] \le \rho.\end{aligned} \tag{5}$$

To be specific, denoting the optimal objective of (5) as $\mu(x)$ and that of (5) as $\mu_{1,0}^-$, we are to show that $\mu_{1,0}^- = \mathbb{E}[\mu(X)\,|\,T=1]$. To see why it is the case, suppose L^* is the optimizer of (5), then it is measurable with respect to X and $Y(1)$ and satisfies the constraints of (5). Then $L(x)(\cdot) := L^*(x,\cdot)$ is measurable with respect to $Y(1)$, and satisfy the constraints of (5). As a result, we have $\mathbb{E}[Y(1)L^*(x,Y(1))\,|\,X=x,T=1] \ge \mu(x)$. Marginalizing over X yields $\mu_{1,0}^- = \mathbb{E}[Y(1)L(X,Y(1))\,|\,T=1] \ge \mathbb{E}[\mu(X)\,|\,T=1]$. On the other hand, suppose $L^*(x)(\cdot)$ is measurable with respect to $Y(1)$ and is the minimizer for (5) for $\mathbb{P}_{X\,|\,T=1}$-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 (5). Thus we have $\mathbb{E}[L(X,Y(1))Y(1)\,|\,T=1] = \mathbb{E}[\mu(X)\,|\,T=1] \ge \mu_{1,0}^-$. Combining the two directions leads to the equivalence.

In the following, we solve (5) and write $\mathbb{E}_x$ in place of $\mathbb{E}[\cdot\,|\,X=x,T=1]$ for simplicity. Invoking Luenberger (1997, Theorem 8.6.1) to this convex problem, we have

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

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

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

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

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

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

$$\begin{aligned}\mu_{1,0}^{-} &= \mathbb{E}\Big[\max_{\alpha(X)\geq 0,\eta(X)\in\mathbb{R}} \varphi\big(\alpha(X),\eta(X),X\big)\,\Big|\,T=1\Big]\\ &= \mathbb{E}\Big[\varphi\big(\alpha^*(X),\eta^*(X),X\big)\,\Big|\,T=1\Big],\end{aligned}$$

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

$$\big(\alpha^*(x),\eta^*(x)\big)\in\operatorname*{arg\,max}_{\alpha\geq 0,\eta\in\mathbb{R}}\ \mathbb{E}\Big[-\alpha f^*\Big(\frac{r_{1,0}(x)}{-\alpha}\big(Y(1)+\eta\big)\Big)-\eta r_{1,0}(x)-\alpha\rho\,\Big|\,X=x,T=1\Big].$$

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

$$\big(\alpha^*(x)/r_{1,0}(x),\eta^*(x)\big)\in\operatorname*{arg\,max}_{\alpha\geq 0,\eta\in\mathbb{R}}\ \mathbb{E}\Big[-\alpha r_{1,0}(x) f^*\Big(\frac{Y(1)+\eta}{-\alpha}\Big)-\eta r_{1,0}(x)-\alpha r_{1,0}(x)\rho\,\Big|\,X=x,T=1\Big].$$

The minimum of (5) can thus be written as

$$\mu_{1,0}^{-}=-\mathbb{E}\Big[r_{1,0}(x)\Big\{\alpha^*(X)f^*\Big(\frac{Y(1)+\eta^*(X)}{-\alpha^*(X)}\Big)+\eta^*(X)+\alpha^*(X)\rho\Big\}\,\Big|\,T=1\Big],$$

where for $\mathbb{P}_{X\,|\,T=1}$-almost all x, it holds that

$$\big(\alpha^*(x),\eta^*(x)\big)\in\operatorname*{arg\,min}_{\alpha\geq 0,\eta\in\mathbb{R}}\ \mathbb{E}\Big[\alpha f^*\Big(\frac{Y(1)+\eta}{-\alpha}\Big)+\eta+\alpha\rho\,\Big|\,X=x,T=1\Big].$$

Therefore, we complete the proof of Proposition 3. □

C.4 Proof of convergence of sieve estimator

Proof. Proof of Theorem 1 We analyze the behavior of $\widehat{\theta}^{(j)}$ for each fold j. As $|\mathcal{I}_1^{(j)}|\asymp n$, we take the generic notation of $\widehat{\theta}$ and sample size n, so that

$$\widehat{\mathbb{E}}_n\big[\ell(\widehat{\theta},X,Y(1))\big]\geq\inf_{\theta\in\Theta_n}\ \widehat{\mathbb{E}}_n\big[\ell(\theta,X,Y(1))\big]-O_P((\frac{\log n}{n})^{2p/(2p+d)}),$$

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

$$\delta_n:=\inf\Big\{\delta\in(0,1)\colon\frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta}\sqrt{\log N(\epsilon^{1+d/2p},\Theta_n,\|\cdot\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})})}\,\mathrm{d}\epsilon\leq 1\Big\},$$

where $N(\epsilon,\Theta_n,\|\cdot\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})})$ is the ϵ-covering number of Θ_n in the L_2-norm under $\mathbb{P}_{\cdot\,|\,T=1}$. 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 C.1.

Lemma C.1. *Let $\theta^*\in\Theta$ be a population risk minimizer. Suppose there exists constants $c_1,c_2>0$ such that $c_1\mathbb{E}[\ell(\theta,X,Y)-\ell(\theta^*,X,Y)]\leq d(\theta,\theta^*)^2\leq c_2\mathbb{E}[\ell(\theta,X,Y)-\ell(\theta^*,X,Y)]$ for θ in a neighborhood of θ^*. Suppose the following conditions hold:*

(i) For sufficiently small $\epsilon>0$, $\mathrm{Var}(\ell(\theta,X,Y)-\ell(\theta^*,X,Y))\leq C_1\epsilon^2$ for all $\theta\in\Theta_n$ 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 $\sup_n\mathbb{E}[U_n(X,Y)^2]\leq C_3$ and $\sup_{\theta\in\Theta_n:\,d(\theta,\theta^*)\leq\delta}|\ell(\theta,X,Y)-\ell(\theta^*,X,Y)|\leq\delta^s U_n(X,Y)$ for constant $C_3>0$.

Then $d(\widehat{\theta}_n-\theta^*)=O_P(\max\{\delta_n,\inf_{\theta'\in\Theta_n}d(\theta',\theta^*)\})$.

We define the distance as L_2-norm $d(\theta,\theta') = \|\theta-\theta'\|_{L_2(\mathbb{P})}$, and verify the conditions in Lemma C.1. We define $\Theta = \Lambda_c^p(\mathcal{X}) \times \Lambda_c^p(\mathcal{X})$ without truncation. The upper bound $\mathbb{E}[\ell(\theta,X,Y) - \ell(\theta^*,X,Y) \,|\, T=1]$ is directly implied by Assumption 2. By the λ-strong convexity of $\mathbb{E}\big[\ell((a,b),x,Y) \,\big|\, X=x\big]$ is at $(a,b)=\theta^*(x)$,

$$\mathbb{E}\big[\ell(\theta(x),x,Y) \,\big|\, X=x\big] - \mathbb{E}\big[\ell(\theta^*(x),x,Y) \,\big|\, X=x\big] \geq \lambda\big(\theta(x)-\theta^*(x)\big)^2.$$

Integrating over X yields $\mathbb{E}[\ell(\theta(X),X,Y)] - \big[\ell(\theta^*(X),X,Y)\big] \geq c''d(\theta,\theta^*)$ for some constant $c''>0$.

We then check condition (i). By the positive density condition, we have $\|\cdot\|_{L_2(\lambda)} \asymp \|\cdot\|_{L_2(\mathbb{P})}$. 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 λ 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, $|\ell(\theta,x,y)-\ell(\theta^*,x,y)| \leq \bar{\ell}(x,y)(\theta(x)-\theta^*(x))$ where $\mathbb{E}[\bar{\ell}(x,Y)^2 \,|\, X=x] \leq M$ for all x, we have

$$\mathrm{Var}\big(\ell(\theta,X,Y)-\ell(\theta^*,X,Y)\big) \leq \mathbb{E}\big[|\ell(\theta,x,y)-\ell(\theta^*,x,y)|^2\big] \leq M\mathbb{E}\big[(\theta(X)-\theta^*(X))^2\big] \leq M\epsilon^2$$

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 $U_n(x,y)=\bar{\ell}(x,y)$. Therefore, applying Lemma C.1 we have $\|\widehat{\theta}_n-\theta^*\|_{L_2(\mathbb{P})} = O_P(\max\{\delta_n, \inf_{\theta'\in\Theta_n} d(\theta',\theta^*)\})$. Here according to Chen and Shen (1998) and Geer et al. (2000), we have

$$\log N\big(\epsilon,\Theta_n^\eta,\|\cdot\|_{2,\mathbb{P}}\big) \lesssim \dim(\Theta_n^\eta)\log\frac{1}{\epsilon},$$

where $\dim(\Theta_n^\eta) = J_n^p$. Since truncation is a contraction map, the covering number of Θ_n^α is upper bounded by the above quantity. As a result, we have

$$\log N\big(\epsilon,\Theta_n,\|\cdot\|_{2,\mathbb{P}}\big) \lesssim J_n^p\log\frac{1}{\epsilon}.$$

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

$$\delta_n \asymp \sqrt{\frac{J_n^d\log n}{n}}.$$

We finally bound the approximation error using Θ_n. Note that we take Θ_n to be truncated at ϵ. However, since the population minimizer θ^* is uniformly bounded above ϵ, since truncation is a contraction map, we have $\inf_{\theta\in\Theta_n}\|\theta-\theta^*\|_{L_2(\mathbb{P})} \leq \inf_{\theta\in\Theta_n^\eta\times\Theta_n^\eta}\|\theta-\theta^*\|_{L_2(\mathbb{P})} \leq O(J_n^p)$, 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 $\|\widehat{\theta}-\theta^*\|_{L_2(\mathbb{P})} = O_P((\log n/n)^{p/(2p+d)})$. This completes our proof. □

C.5 Proof of Proposition 4

Given $X=x$, suppose instead $\alpha^*(x)=0$. We consider the following two cases:

- If $\eta^*(x) < -\underline{y}(x)$, then

$$\begin{aligned}
&\liminf_{\alpha\to0}\mathbb{E}\left[\alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big)+\eta^*(x)+\alpha\rho \,\Big|\, X=x,T=1\right]\\
=&\liminf_{\alpha\to0}\mathbb{E}\left[\alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big)\mathbb{1}\big\{Y(1)\leq-\eta^*(x)\big\}\right.\\
&\qquad\left.+\alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big)\mathbb{1}\big\{Y(1)>-\eta^*(x)\big\} \,\big|\, X=x,T=1\right]+\eta^*(x)+\alpha\rho\\
\overset{(a)}{\geq}&\liminf_{\alpha\to0}\mathbb{E}\left[\alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big)\mathbb{1}\big\{Y(1)\leq-\eta^*(x)\big\}-\alpha L \,\Big|\, X=x,T=1\right]\\
&\qquad+\liminf_{\alpha\to0}\left[\alpha f^*\Big(\frac{Y(1)+\eta}{-\alpha}\Big)\mathbb{1}\big\{Y(1)>-\eta^*(x)\big\}-\alpha L \,\Big|\, X=x,T=1\right]+\eta^*(x)\\
\overset{(b)}{\geq}&+\infty,
\end{aligned}$$

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 $\eta^*(x) \geq -\underline{y}(x)$, then

$$\begin{aligned}
&\lim_{\alpha\to 0} \mathbb{E}\Big[\alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big) + \eta^*(x) + \alpha\rho \,\Big|\, X = x, T = 1\Big] \\
&= \lim_{\alpha\to 0} \mathbb{E}\Big[\alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big)\mathbb{1}\big\{Y(1) \geq \operatorname{essinf} Y(1)\big\} + \alpha\rho \,\Big|\, X = x, T = 1\Big] + \eta^*(x) \\
&\overset{\text{(a)}}{=} \mathbb{E}\Big[\lim_{\alpha\to 0} \alpha f^*\Big(\frac{Y(1)+\eta^*(x)}{-\alpha}\Big)\mathbb{1}\big\{Y(1) \geq \operatorname{essinf} Y(1)\big\} + \alpha\rho \,\Big|\, X = x, T = 1\Big] + \eta^*(x) \\
&\overset{\text{(b)}}{=} \eta^*(x).
\end{aligned}$$

  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 $\eta^*(x) = -\underline{y}(x)$ and the optimal value of the dual problem is $-\underline{y}(x)$. By the strong duality, the optimal value of the primal objective function is $\mathbb{E}\big[r_{1,0}(X)\underline{y}(X) \,|\, T = 1\big]$. As an implication, there exists a feasible $L(x, y)$ such that $\mathbb{E}[Y(1)L(X, Y(1)) \,|\, T = 1] = \mathbb{E}\big[r_{1,0}(X)\underline{y}(X) \,|\, T = 1\big]$. Let $\mathbb{Q}_{Y\,|\,X=x}$ denote the measure induced by $L(x, y)$:

$$\frac{\mathrm{d}\mathbb{Q}_{Y\,|\,X=x}}{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X=x,T=1}}(y) = \frac{L(x, y)}{r_{1,0}(x)}.$$

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

$$\begin{aligned}
&1 = \mathbb{Q}\big(Y(1) = \underline{y}(X) \,|\, X = x\big) = \mathbb{E}\Big[\frac{L\big(x, \underline{y}(x)\big)}{r_{1,0}(x)} \cdot \mathbb{1}\big\{Y(1) = \underline{y}(x)\big\} \,\Big|\, X = x, T = 1\Big] = \frac{L(x, \underline{y}(x))}{r_{1,0}(x)} \cdot \bar{p}(x), \\
&0 = L(x, Y(1)) \cdot \mathbb{1}\{Y(1) > \underline{y}(x)\}, \text{ a.s. under } \mathbb{P}_{Y(1)\,|\,X=x,T=1}.
\end{aligned}$$

Again since L is feasible,

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

This is a contradiction to the condition. Therefore, we have $\alpha^*(x) > 0$.

C.6 Proof of Theorem 2

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

$$\begin{aligned}
\widehat{\mu} &:= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} \widehat{r}(X_i)\big(\widehat{H}(X_i, Y_i) - \widehat{h}(X_i)\big) + \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0} \widehat{h}(X_i) \\
&\xrightarrow{P} \mathbb{E}\big[r(X)\ell(\theta^\diamond(X), X, Y(1)) \,\big|\, T = 1\big]. \qquad (6)
\end{aligned}$$

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

It thus remains to show (6). We prove the results when either $\widehat{r}$ or $\widehat{h}$ is consistent.

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

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

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

$$\begin{aligned}
&\mathbb{E}\bigg[\Big(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\big(\widehat{H}(X_i,Y_i) - \widehat{h}(X_i)\big)\Big)^2\bigg] \\
&\le \mathbb{E}\Big[\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)^2\mathbb{E}\big[\big(\widehat{H}(X_i,Y_i) - \widehat{h}(X_i)\big)^2 \,\big|\, X_i\big]\Big] \le M\cdot\|\widehat{r} - r_{1,0}\|^2_{L_2(\mathbb{P}_{X\,|\,T=1})} = o_P(1).
\end{aligned}$$

Invoking Lemma E.1, 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 $\mathbb{E}[\widehat{h}(X)r(X)\,|\,T=1] = \mathbb{E}[\widehat{h}(X)\,|\,T=0]$, where we still implicitly condition on other folds. As a result,

$$\begin{aligned}
&-\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r_{1,0}(X_i)\widehat{h}(X_i) + \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\widehat{h}(X_i) \\
&= -\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(r_{1,0}(X_i)\widehat{h}(X_i) - \mathbb{E}\big[\widehat{h}(X)r(X)\,\big|\,T=1\big]\big) + \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\big(\widehat{h}(X_i) - \mathbb{E}\big[r(X)\,\big|\,T=0\big]\big),
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\bigg[\Big(-\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r_{1,0}(X_i)\widehat{h}(X_i) + \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\widehat{h}(X_i)\Big)^2\bigg] \\
&\le \frac{2}{|\mathcal{I}_1|}\operatorname{Var}\big(r_{1,0}(X)\widehat{h}(X)\,\big|\,T=1\big) + \frac{2}{|\mathcal{I}_0|}\operatorname{Var}\big(\widehat{h}(X)\,\big|\,T=0\big) = o_P(1)
\end{aligned}$$

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

$$\begin{aligned}
\widehat{\mu} &= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\widehat{H}(X_i,Y_i) + o_P(1) \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\ell\big(\theta^\diamond(X_i),X_i,Y_i\big) + \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big\{\ell\big(\widehat{\theta}(X_i),X_i,Y_i\big) - \ell\big(\theta^\diamond(X_i),X_i,Y_i\big)\big\} + o_P(1) \\
&= \mathbb{E}\big[r(X_i)\ell(\theta^\diamond(X),X,Y(1))\,\big|\,T=1\big] + \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big\{\ell\big(\widehat{\theta}(X_i),X_i,Y_i\big) - \ell\big(\theta^\diamond(X_i),X_i,Y_i\big)\big\} + o_P(1).
\end{aligned}$$

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

$$\Big|\ell\big(\widehat{\theta}(X_i),X_i,Y_i\big) - \ell\big(\theta^\diamond(X_i),X_i,Y_i\big)\Big| \le M(X_i,Y_i)\big\|\widehat{\theta}(X_i) - \theta^\diamond(X_i)\big\|_2,$$

hence (implicitly conditioning on other folds) we have

$$\begin{aligned}
&\mathbb{E}\bigg[\Big(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big\{\ell\big(\widehat{\theta}(X_i),X_i,Y_i\big) - \ell\big(\theta^\diamond(X_i),X_i,Y_i\big)\big\}\Big)^2\bigg] \\
&\le \mathbb{E}\Big[r(X)^2M(X,Y(1))^2\big\|\widehat{\theta}(X_i) - \theta^\diamond(X_i)\big\|_2^2\Big] = o_P(1)
\end{aligned}$$

since $\mathbb{E}[M(X,Y(1))^2\,|\,T=1] \le M$ for some constant $M > 0$. We've thus completed the proof of (6).

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

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

The first summation is unbiased conditional on other folds, hence

$$\mathbb{E}\bigg[\Big(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\widehat{r}(X_i)\big(\widehat{H}(X_i, Y_i) - \bar{h}(X_i)\big)\Big)^2\bigg] = \frac{1}{|\mathcal{I}_1|}\operatorname{Var}\big(\widehat{r}(X)\{\widehat{H}(X, Y(1)) - \bar{h}(X)\}\,\big|\,T = 1\big) = o_P(1)$$

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

$$\mathbb{E}\bigg[\Big(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\bar{h}(X_i) - \widehat{h}(X_i)\big)\Big)^2\bigg] \le \big\|\widehat{r}\cdot(\widehat{h} - \bar{h})\big\|^2_{L_2(\mathbb{P}_{X\,|\,T=1})} = o_P(1)$$

due to the boundedness of $\widehat{r}$. Similarly, we know that $\frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\bar{h}(X_i) - \widehat{h}(X_i) = o_P(1)$. Consequently, invoking Lemma E.1 we drop the implicit conditioning and arrive at

$$\widehat{\mu} = \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\bar{h}(X_i) + o_P(1) = \mathbb{E}\big[\bar{h}(X)\,|\,T = 0\big] + o_P(1)$$

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

$$\begin{aligned}\Big|\mathbb{E}\big[\bar{h}(X)\,|\,T = 0\big] - \mathbb{E}\big[h^\diamond(X)\,\big|\,T = 0\big]\Big|^2 &\le \|\bar{h} - h^\diamond\|^2_{L_2(\mathbb{P}_{X\,|\,T=0})}\\ &\le \big\|\widehat{H}(X, Y(1)) - \ell(\theta^\diamond(X), X, Y(1))\big\|^2_{L_2(\mathbb{P}_{\cdot\,|\,T=0})}.\end{aligned}$$

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

$$\widehat{\mu} = \mathbb{E}\big[h^\diamond(X)\,|\,T = 0\big] + o_P(1) = \mathbb{E}\big[r(X)h^\diamond(X)\,|\,T = 1\big] + o_P(1) = \mathbb{E}\big[r(X)\ell(\theta^\diamond(X), X, Y(1))\,\big|\,T = 1\big] + o_P(1)$$

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

C.7 Proof of Theorem 3

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

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

and we define

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

We show this result for any j; we implicitly condition on all the remaining folds other than $\mathcal{I}^{(j)}_1$ and $\mathcal{I}^{(j)}_0$, so that all nuisance components are viewed as fixed. To simplify notations, we write $\mathcal{I}_1 := \mathcal{I}^{(j)}_1$, $\mathcal{I}_0 := \mathcal{I}^{(j)}_0$ and $r := r_{1,0}$, $\widehat{r} := \widehat{r}^{(j)}$, $\widehat{h} := \widehat{h}^{(j)}$, $\widehat{H} := \widehat{H}^{(j)}$, $\bar{h} := \bar{h}^{(j)}$. We also represent the parameters (functionals) with

$$\widehat{\theta}(\cdot) = \big(\widehat{\alpha}(\cdot), \widehat{\eta}(\cdot)\big) := \big(\widehat{\alpha}^{(j)}(\cdot), \widehat{\eta}^{(j)}(\cdot)\big), \quad \theta^*(\cdot) := \big(\alpha^*(\cdot), \eta^*(\cdot)\big),$$

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

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

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

$$\begin{aligned}
\widehat{\mu}_{1,0}^{(j)} - \widehat{\mu}_{1,0}^{*(j)} &= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\Big[\widehat{r}(X_i)\big(\widehat{H}(X_i, Y_i) - \widehat{h}(X_i)\big) - r(X_i)\big(H(X_i, Y_i) - h(X_i)\big)\Big] + \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\big(\widehat{h}(X_i) - h(X_i)\big) \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big(\widehat{H}(X_i, Y_i) - H(X_i, Y_i)\big) - \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r(X_i)\big)\big(\widehat{h}(X_i) - \bar{h}(X_i)\big) \\
&\quad + \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r(X_i)\big)\big(\widehat{H}(X_i, Y_i) - \bar{h}(X_i)\big) \\
&\quad - \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big(\widehat{h}(X_i) - h(X_i)\big) + \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\big(\widehat{h}(X_i) - h(X_i)\big).
\end{aligned}$$

In the following, we are to bound the several summations separately. Firstly, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\Big|\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r(X_i)\big)\big(\widehat{h}(X_i) - \bar{h}(X_i)\big)\Big| &\le \sqrt{\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r(X_i)\big)^2}\sqrt{\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{h}(X_i) - \bar{h}(X_i)\big)^2} \\
&= O_P\big(\|\widehat{r} - r\|_{L_2(\mathbb{P}_{X\,|\,T=1})} \cdot \|\widehat{h} - h\|_{L_2(\mathbb{P}_{X\,|\,T=1})}\big) = o_P(1/\sqrt{n})
\end{aligned}$$

under the given convergence rate of the product. Since $\bar{h}(x) = \mathbb{E}\big[\widehat{H}(X, Y(1)) \,|\, X = x, T = 1\big]$ for the fixed function $\widehat{H}$, the term $\big(\widehat{r}(X_i) - r(X_i)\big)\big(\widehat{H}(X_i, Y_i) - \bar{h}(X_i)\big)$ has mean zero, hence by Markov's inequality,

$$\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r(X_i)\big)\big(\widehat{h}(X_i) - \bar{h}(X_i)\big) = O_P\Big(\sqrt{\mathrm{Var}(\widehat{r}(X_i) - r(X_i))(\widehat{H}(X_i, Y_i) - \bar{h}(X_i)))}/\sqrt{n}\Big),$$

where, by the consistency of $\widehat{r}$, this term is $o_P(1/\sqrt{n})$. Furthermore, note that

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big(\widehat{h}(X_i) - h(X_i)\big) - \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\big(\widehat{h}(X_i) - h(X_i)\big) \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\Big(r(X_i)\big(\widehat{h}(X_i) - h(X_i)\big) - \mathbb{E}\big[r(X)(\widehat{h}(X) - h(X)) \,\big|\, T = 1\big]\Big) \\
&\quad - \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\big(\widehat{h}(X_i) - h(X_i) - \mathbb{E}\big[\widehat{h}(X_i) - h(X_i) \,|\, T = 0\big]\big),
\end{aligned} \tag{7}$$

where we use the equivalence of the two expectations: this is because there is a covariate shift $r(X)$ from $\mathbb{P}_{X\,|\,T=1}$ to $\mathbb{P}_{X\,|\,T=0}$, hence $\mathbb{E}[\phi(X)r(X) \,|\, T = 1] = \mathbb{E}[\phi(X) \,|\, X = 0]$ for any integrable function $\phi\colon \mathcal{X} \to \mathbb{R}$. The two summations in (7) is thus both unbiased, indicating

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big(\widehat{h}(X_i) - h(X_i)\big) - \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\big(\widehat{h}(X_i) - h(X_i)\big) \\
&= O_P\Big(\sqrt{\mathrm{Var}(r(X)(\widehat{h}(X) - h(X)) \,|\, T = 1)/n} + \sqrt{\mathrm{Var}(r(X)(\widehat{h}(X) - h(X)) \,|\, T = 0)/n}\Big) \\
&= O_P\big(\|r(X)(\widehat{h}(X) - h(X))\|_{L_2(\mathbb{P}_{X\,|\,T=1})}/\sqrt{n} + \|\widehat{h} - h\|_{L_2(\mathbb{P}_{X\,|\,T=0})}/\sqrt{n}\big) = o_P(1/\sqrt{n}),
\end{aligned}$$

where the last equality follows from the L_2-consistency of $\widehat{h}$ to $\bar{h}$ and the fact that $\|\bar{h} - h\|_{L_2(\mathbb{P}_{X\,|\,T=1})}$ by the stability of the conditional expectations induced by the stability of g in Assumption 4. Finally, we turn to

$$\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} r(X_i)\big(\widehat{H}(X_i, Y_i) - H(X_i, Y_i)\big).$$

Since $\ell(\theta, x, y)$ is a convex function in θ for any (x, y), for any ($\mathbb{P}_{X \mid T=1}$-almost all) x,

$$\mathbb{E}\big[\ell(\theta, x, y) \,\big|\, X = x, T = 1\big] = \alpha(x)\mathbb{E}\left[f^*\Big(\frac{Y(1) + \eta(x)}{-\alpha(x)}\Big) \,\middle|\, X = x, T = 1\right] + \eta(x) + \alpha(x)\rho$$

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,

$$\nabla_\theta \mathbb{E}\big[\ell(\theta, x, Y(1)) \,\big|\, X = x, T = 1\big]\big|_{\theta=(\alpha^*(x),\eta^*(x))} = \mathbb{E}\big[\nabla_\theta \ell(\theta^*(x), x, Y(1)) \,\big|\, X = x, T = 1\big] = 0.$$

Multiplying $r(X)$ and integrating over $X \mid T = 1$, we know that

$$\mathbb{E}\big[r(X)\nabla_\theta \ell(\theta^*(X), X, Y(1))[\widehat{\theta}(X) - \theta^*(X)] \,\big|\, T = 1\big] = 0. \tag{8}$$

By Lemma 2 of Chen and Shen (1998), when both $\widehat{\theta}$ and θ^* is smooth enough, sufficiently small $\|\widehat{\theta} - \theta^*\|_{L_2(\mathbb{P})}$ implies sufficiently small $\|\widehat{\theta} - \theta^*\|_\infty$. As a result, when $\|\widehat{\theta}(x) - \theta^*(x)\|_{L_2(\mathbb{P}_{\cdot \mid T=1})}$ is sufficiently small, by the condition that $\big|\ell(\widehat{\theta}, x, y) - \ell(\theta^*, x, y) - \nabla_\theta \ell(\theta^*(x), x, y)[\theta^*(x) - \widehat{\theta}(x)]\big| \le M(x, y)\|\widehat{\theta}(x) - \theta^*(x)\|_2^2$ as well as Jensen's inequality, we have

$$\begin{aligned}
&\Big|\mathbb{E}\big[\widehat{H}(X, Y(1)) - H(X, Y(1)) \,\big|\, T = 1\big]\Big| \\
&= \left|\mathbb{E}\Big[r(X)\Big\{\ell\big(\widehat{\theta}, X, Y(1)\big) - \ell\big(\theta^*, X, Y(1)\big) - \nabla_\theta \ell\big(\theta^*(X), X, Y(1)\big)\big[\widehat{\theta}(X) - \theta^*(X)\big]\Big\} \,\Big|\, T = 1\Big]\right| \\
&\le \mathbb{E}\Big[r(X)\big|\ell\big(\widehat{\theta}, X, Y(1)\big) - \ell\big(\theta^*, X, Y(1)\big) - \nabla_\theta \ell\big(\theta^*(X), X, Y(1)\big)\big[\widehat{\theta}(X) - \theta^*(X)\big]\big| \,\Big|\, T = 1\Big] \\
&\le \mathbb{E}\big[r(X)M(X, Y(1))\|\widehat{\theta}(X) - \theta^*(X)\|_2^2 \,\big|\, T = 1\big] = M \cdot \|r(\widehat{\theta} - \theta)\|^2_{L_2(\mathbb{P}_{X \mid T=0})}.
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_1|}\sum_{i \in \mathcal{I}_1}\Big[r(X_i)\big(\widehat{H}(X_i, Y_i) - H(X_i, Y_i)\big) - \mathbb{E}\big[\widehat{H}(X, Y(1)) - H(X, Y(1)) \,\big|\, T = 1\big]\Big] \\
&= O_P\big(\|r(X)\widehat{H}(X, Y(1)) - H(X, Y(1))\|_{L_2(\mathbb{P}_{X \mid T=1})}/\sqrt{n}\big),
\end{aligned}$$

where by the given conditions, we have

$$\|r(X)\widehat{H}(X, Y(1)) - H(X, Y(1))\|_{L_2(\mathbb{P}_{X \mid T=1})} = O\big(\|\widehat{\theta} - \theta^*\|_{L_2(\mathbb{P}_{X \mid T=1})}\big) = o_P(1).$$

As a result, we have

$$\frac{1}{|\mathcal{I}_1|}\sum_{i \in \mathcal{I}_1} r(X_i)\big(\widehat{H}(X_i, Y_i) - H(X_i, Y_i)\big) \le o_P(1/\sqrt{n}) + M\|\widehat{\theta} - \theta\|^2_{L_2(\mathbb{P}_{X \mid T=0})} = o_P(1/\sqrt{n}).$$

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

$$\sqrt{n}\big(\widehat{\mu}_{1,0}^{-} - \mu_{1,0}^{-}\big) = \frac{\sqrt{n}}{n_1}\sum_{T_i=1}\Big(r_{1,0}(X_i)\big(H(X_i, Y_i) - h(X_i)\big) - \mu_{1,0}^{-}\Big) + \frac{\sqrt{n}}{n_0}\sum_{T_i=0} h(X_i) + o_P(1/\sqrt{n}),$$

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

$$\sigma^2 = \frac{1}{p_1}\operatorname{Var}\Big(r_{1,0}(X)\big(H(X, Y(1)) - h(X)\big) \,\Big|\, T = 1\Big) + \frac{1}{p_0}\operatorname{Var}\big(h(X) \,\big|\, T = 0\big).$$

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 $\widehat{\sigma}^2 \to \sigma^2$ in the definition of Theorem 3. As $\widehat{p}_1 \xrightarrow{p} p_1$, $\widehat{p}_0 \xrightarrow{p} p_0$, by the law of large numbers, it suffices to show that $\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i}^2 - (d_{1,i}^*)^2\big) = o_P(1)$ and $\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i} - d_{1,i}^*\big) = o_P(1)$ and similar for $(d_{0,i}, d_{0,i}^*)$, where we define the oracle counterparts

$$d_{1,i}^* = r_{1,0}(X_i)\big(H(X_i,Y_i) - h(X_i)\big), \quad d_{0,i}^* = h(X_i).$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i}^2 - (d_{1,i}^*)^2\big) &= \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i} - d_{1,i}^*\big)^2 + \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} 2\big(d_{1,i} - d_{1,i}^*\big)d_{1,i}^* \\ &\le \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i} - d_{1,i}^*\big)^2 + \sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i} - d_{1,i}^*\big)^2}\cdot\sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(d_{1,i}^*\big)^2}.\end{aligned}$$

Focusing on the summation within $\mathcal{I}_1^{(j)}$, we have

$$\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\big(d_{1,i} - d_{1,i}^*\big)^2 = O_P\Big(\big\|\widehat{r}^{(j)}(\widehat{H}^{(j)} - \widehat{h}^{(j)}) - r_{1,0}(H-h)\big\|^2_{L_2(\mathbb{P}_{\cdot\,|\,T=1})}\Big),$$

where the right-handed side is $o_P(1)$ under the conditions of Theorem 3. Other folds and other summation terms follow similar arguments hence $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$. By Slutsky's lemma, we conclude the proof of Theorem 3. □

C.8 Proof of Theorem 4

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

$$\mathbb{E}\big[r(X)\nabla_\theta\ell\big(\theta^\diamond(X), X, Y(1)\big)[\widehat{\theta}(X) - \theta^\diamond(X)]\,\big|\,T=1\big]$$

in parallel with (8) in the proof of Theorem 3. We note that this is directly implied by our local condition. Therefore, under the conditions of Theorem 4,

$$\widehat{\mu}_{1,0}^- = -\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} r_{1,0}(X_i)\big[H^\diamond(X_i,Y_i(1)) - h^\diamond(X_i)\big] - \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} h^\diamond(X_i).$$

The terms in the summation has expectation

$$\mu_{1,0}^\diamond := -\mathbb{E}\big[r_{1,0}(X_i)H^\diamond(X_i,Y_i(1))\,\big|\,T=1\big].$$

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

$$\mathbb{E}\big[H^\diamond(X_i,Y_i(1))\,\big|\,X=x,T=1\big] \ge \mathbb{E}\big[H(X_i,Y_i(1))\,\big|\,X=x,T=1\big]$$

for $\mathbb{P}_{X\,|\,T=1}$-almost all x, hence by tower property, we have $\mu_{1,0}^\diamond \le \mu_{1,0}^-$. On the other hand, the consistency of $\widehat{\sigma}^2$ to $\mathrm{Var}(\phi_{1,-}^\diamond(X,Y,T))$ also follows the same arguments as the proof of Theorem 3 with $\theta^\diamond$, which concludes our proof of Theorem 4. □

D Technical proofs for marginal f-sensitivity models

D.1 Proof of Proposition B.1

Proof. Proof of Proposition B.1 We prove the claim by showing that $\mathcal{Q}_{1,0}^{\mathrm{m}}$ and the set on the right-handed side (RHS) of Proposition 1 contain each other.

Step 1: $\mathcal{Q}_{1,0}^{\mathrm{m}} \subseteq$ RHS.

By definition, any member of $\mathcal{Q}^{\mathrm{m}}_{1,0}$ is the induced distribution over $(X, Y(1))$ given $T = 0$ of some super-population $\mathbb{Q} \in \mathcal{M}^{f,\rho}_{1,\mathrm{mgn}}$ with $\mathbb{Q}_{X,Y,T} = \mathbb{P}_{X,Y,T}$. Thus, we write it as $\mathbb{Q}_{X,Y(1)\,|\,T=0} \in \mathcal{Q}^{\mathrm{m}}_{1,0}$. Since $\mathbb{Q}_{X,Y,T} = \mathbb{P}_{X,Y,T}$, we know that $\frac{\mathrm{d}\mathbb{Q}_{X\,|\,T=0}}{\mathrm{d}\mathbb{P}_{X\,|\,T=1}} = \frac{\mathrm{d}\mathbb{P}_{X\,|\,T=0}}{\mathrm{d}\mathbb{P}_{X\,|\,T=1}} = r_{1,0}(x)$. Following the proof of Lemma 1,

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

as well as (by condition (1) and the data-processing inequality)

$$\begin{aligned} D_f\big(\mathbb{P}_{Y(1)\,|\,X=x,T=0} \,\|\, \mathbb{P}_{Y(1)\,|\,X=x,T=1}\big) \le & D_f\big(\mathbb{P}_{Y(1),U\,|\,X=x,T=0} \,\|\, \mathbb{P}_{Y(1),U\,|\,X=x,T=1}\big) \\ = & \mathbb{E}_{Y(1),U\,|\,X=x,T=1}\bigg[f\Big(\frac{\mathrm{d}\mathbb{P}_{Y(1),U\,|\,X=x,T=0}}{\mathrm{d}\mathbb{P}_{Y(1),U\,|\,X=x,T=1}}\Big)\bigg]. \end{aligned}$$

Combining the above two facts and using the tower property,

$$\mathbb{E}\big[D_f\big(\mathbb{P}_{Y(1)\,|\,X,T=0} \,\|\, \mathbb{P}_{Y(1)\,|\,X,T=1}\big)\,\big|\,T=1\big] \le \mathbb{E}_{Y(1),U\,|\,T=1}\Big[f\big(\mathrm{OR}(X,U)\big)\Big] \le \rho.$$

Also, note that $\mathbb{Q}_{Y(1)\,|\,X,T=1} = \mathbb{Q}_{Y\,|\,X,T=1} = \mathbb{P}_{Y\,|\,X,T=1}$ since $Y = TY(1) + (1-T)Y(0)$ and $\mathbb{Q}_{X,Y,T} = \mathbb{P}_{X,Y,T}$. These altogether establish the "$\subset$" direction. It remains to prove the reverse.

Step 2: RHS $\subseteq \mathcal{Q}^{\mathrm{m}}_{1,0}$.

Given any distribution Q over $(X, Y(1))$ in RHS, we aim to find a distribution (super-population) $\mathbb{Q}$ over $(X, Y(0), Y(1), U, T)$ such that

- $(Y(1), Y(0)) \perp\!\!\!\perp T \,|\, X, U$ under $\mathbb{Q}$;
- $\mathbb{Q}_{X,T,Y} = \mathbb{P}_{X,T,Y}$;
- $\mathbb{Q} \in \mathcal{M}^{f,\rho}_{1,\mathrm{mgn}}$;
- $\mathbb{Q}_{X,Y(1)\,|\,T=0}(x,y) = Q(x,y)$.

These conditions imply that $Q \in \mathcal{Q}^{\mathrm{m}}_{1,0}$, and thus RHS $\subseteq \mathcal{Q}^{\mathrm{m}}_{1,0}$. To construct $\mathbb{Q}$, we first set $\mathbb{Q}_{X,T} = \mathbb{P}_{X,T}$. Then we specify the distribution of $Y(1)\,|\,X,T$ via

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

So far, $\mathbb{Q}_{X,T,Y(1)}$ 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

$$\mathbb{Q}_{Y(0)\,|\,X,T,Y(1),U} = \mathbb{Q}_{Y(0)\,|\,X} = \mathbb{P}_{Y\,|\,X,T=0}.$$

Having constructed $\mathbb{Q}$, we proceed to check that it satisfies the conditions listed at the beginning of Step 2. Since $Y(1) = U$ and $\mathbb{Q}_{Y(0)\,|\,X,T,Y(1),U} = \mathbb{Q}_{Y(0)\,|\,X}$ by construction, it is easy to check that $(Y(1), Y(0)) \perp\!\!\!\perp T \,|\, X, U$. By construction, it also clearly holds that $\mathbb{Q}_{X,T,Y} = \mathbb{P}_{X,T,Y}$.

We now check that $\mathbb{Q} \in \mathcal{M}^{f,\rho}_{1,\mathrm{mgn}}$. For any x, again by the construction of $\mathbb{Q}$,

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}_{X,Y(1)\,|\,T=1}}\bigg[f\Big(\frac{\mathrm{d}\mathbb{Q}_{Y(1)\,|\,X=x,T=0}}{\mathrm{d}\mathbb{Q}_{Y(1)\,|\,X=x,T=1}}\Big)\bigg] &= \mathbb{E}_{\mathbb{P}_{X,Y\,|\,T=1}}\bigg[f\Big(\frac{\mathrm{d}Q_{Y\,|\,X=x}}{\mathrm{d}\mathbb{P}_{Y\,|\,X=x,T=1}}\Big)\bigg] \\ &= \mathbb{E}\big[D_f\big(Q_{Y\,|\,X=x} \,\|\, \mathbb{P}_{Y\,|\,X=x,T=1}\big)\,\big|\,T=1\big] \le \rho, \end{aligned}$$

where the last inequality is due to the fact that Q is in the set on the RHS. Therefore, we verify that $\mathbb{Q} \in \mathcal{M}^{f,\rho}_{1,\mathrm{mgn}}$. By construction, $\mathbb{Q}_{Y(1)\,|\,X,T=0} = Q_{Y\,|\,X}$. It remains to show that $\mathbb{Q}_{X\,|\,T=0} = Q_X$. For any measurable set A,

$$\begin{aligned} \mathbb{Q}(X \in A\,|\,T=0) = & \mathbb{E}_{\mathbb{Q}}\bigg[\frac{\mathrm{d}\mathbb{Q}_{X\,|\,T=0}}{\mathrm{d}\mathbb{Q}_{X\,|\,T=1}} \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\bigg] = \mathbb{E}_{\mathbb{Q}}\bigg[\frac{\mathrm{d}\mathbb{Q}_{X\,|\,T=0}}{\mathrm{d}\mathbb{Q}_{X\,|\,T=1}} \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\bigg] \\ = & \mathbb{E}_{\mathbb{Q}}\Big[r_{1,0}(X) \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\Big] = \mathbb{E}_{\mathbb{Q}}\bigg[\frac{\mathrm{d}Q_X}{\mathrm{d}\mathbb{Q}_{X\,|\,T=1}} \cdot \mathbb{1}\{X \in A\}\,\Big|\,T=1\bigg] = Q(X \in A). \end{aligned}$$

Since the above holds for any measuable set A, we know that $\mathbb{Q}_{X\,|\,T=0} = Q_X$. Finally, switching the role of the treated and control groups yields similar results for the counterfactual distribution of $(X, Y(0))$ given $T = 1$ under $\mathcal{M}^{f,\rho}_{0,\mathrm{mgn}}$. This completes the proof of Proposition 1. □

D.2 Proof of Proposition B.3

Proof. Proof of Proposition B.3 The problem (5) is also convex. For notational simplicity, throughout the proof we write $\mathbb{E}[\cdot]$ instead of $\mathbb{E}[\cdot \,|\, T=1]$, and $\mathbb{E}[\cdot \,|\, X=x]$ instead of $\mathbb{E}[\cdot \,|\, X=x, T=1]$. According to Luenberger (1997, Theorem 8.6.1), the optimal value of the objective function in (5) can be written as

$$\min_{\substack{\mathbb{E}[L \,|\, X=x]=r_{1,0}(x),\forall x \\ \mathbb{E}[f(L/r_{1,0}(X))]-\rho\le 0}} \mathbb{E}[Y(1)L] = \max_{\alpha\ge 0} \ \varphi(\alpha),$$

where the dual functional is given by (where L is essentially a measurable function of $(X, Y(1))$)

$$\begin{aligned} \varphi(\alpha) &= \min_{\mathbb{E}[L \,|\, X=x]=r_{1,0}(x),\forall x} \mathcal{L}(\alpha, L), \\ \mathcal{L}(\alpha, L) &= \mathbb{E}[Y(1)L] + \alpha\big(\mathbb{E}[f(L/r_{1,0}(X))] - \rho\big). \end{aligned} \tag{9}$$

Claim D.1. *We claim that* $\varphi(\alpha) = \mathbb{E}\big[\varphi(\alpha, X)\big]$, *where* $\varphi(\alpha, x)$ *is the optimal objective of*

$$\begin{aligned} \min \ \ & \mathbb{E}[Y(1)L \,|\, X=x] + \alpha\big(\mathbb{E}[f(L/r_{1,0}(x)) \,|\, X=x] - \rho\big) \\ s.t. \ \ & \mathbb{E}[L \,|\, X=x] = r_{1,0}(x). \end{aligned} \tag{10}$$

Proof. Proof of Claim D.1 The claim follows similar arguments as the Neyman-Pearson Lemma. To see why it is the case, let $L^*(x,\cdot)$, a function measurable with respect to $Y(1)$, be the optimizer of the per-x optimization problem (10). Then L^*, as a function of both X and $Y(1)$, satisfies the constraint of (9), hence $\varphi(\alpha) \le \mathcal{L}(\alpha, L^*) = \mathbb{E}[\mathbb{E}[\mathcal{L}(\alpha, L) \,|\, X]] = \mathbb{E}[\varphi(\alpha, X)]$. On the other hand, for every measureable function $L(\cdot,\cdot)$ that satisfies the constraint of (9), if we define $L_x = L(x,\cdot)$, then L_x satisfies the constraint of (10), hence

$$\mathbb{E}\big[Y(1)L_x(Y(1)) \,\big|\, X=x\big] + \alpha\big(\mathbb{E}[f(L_x/r_{1,0}(x)) \,|\, X=x] - \rho\big) \ge \varphi(\alpha, x).$$

Integrating over X, we have $\mathcal{L}(\alpha, L) \ge \mathbb{E}[\varphi(\alpha, X)]$. Minimizing over such L we get $\varphi(\alpha) \ge \mathbb{E}[\varphi(\alpha, X)]$. By the two inequalities we conclude the proof of Claim D.1. □

Again invoking Luenberger (1997, Theorem 8.6.1), we can further solve for $\varphi(\alpha, x)$ via its dual

$$\begin{aligned} \varphi(\alpha, x) &= \max_{\eta\in\mathbb{R}} \ \phi(\alpha, \eta, x) \\ &= \max_{\eta\in\mathbb{R}} \ \min_{L\ge 0} \ \Big\{\mathbb{E}[Y(1)L \,|\, X=x] + \alpha\big(\mathbb{E}[f(L/r_{1,0}(x)) \,|\, X=x] - \rho\big) + \eta\mathbb{E}\big[L - r_{1,0}(x) \,|\, X=x\big]\Big\} \\ &= \max_{\eta\in\mathbb{R}} \ \min_{L\ge 0} \ \Big\{\mathbb{E}\big[Y(1)L + \alpha f(L/r_{1,0}(x)) + \alpha\eta L \,\big|\, X=x\big] - \alpha\rho - \eta r_{1,0}(x)\Big\}. \end{aligned}$$

Similar to the idea of Claim D.1, the inner minimization problem can also be solved inside the expectation, which leads to

$$\begin{aligned} \varphi(\alpha, x) &= \max_{\eta\in\mathbb{R}} \ \Big\{\mathbb{E}\Big[\min_{L\ge 0} \ \{Y(1)L + \alpha f(L/r_{1,0}(x)) + \alpha\eta L\} \,\Big|\, X=x\Big] - \alpha\rho - \eta r_{1,0}(x)\Big\} \\ &= \max_{\eta\in\mathbb{R}} \ \Big\{ -\alpha\mathbb{E}\Big[f^*\Big(-\frac{r_{1,0}(x)}{\alpha}\big(Y(1)+\eta\big)\Big) \,\Big|\, X=x\Big] - \alpha\rho - \eta r_{1,0}(x)\Big\}, \end{aligned}$$

where $f^*(s) = \sup_{t\ge 0}\{st - f(t)\}$ is the conjugate function. The solution to (2) is thus given by (3). This completes the proof of Proposition B.3. □

D.3 Proof of sieve estimation

Proof. Proof of Proposition B.4 To simplify notations, sometimes we write $\mathbb{E}[\cdot]$ instead of $\mathbb{E}[\cdot \,|\, T=1]$, $\mathbb{E}[\cdot \,|\, X=x]$ instead of $\mathbb{E}[\cdot \,|\, X=x, T=1]$, $\mathbb{P}(\cdot)$ instead of $\mathbb{P}(\cdot \,|\, T=1)$ and $\mathbb{P}(\cdot \,|\, X=x)$ instead of $\mathbb{P}(\cdot \,|\, X=x, T=1)$. We will invoke Lemma C.1 again and verify the conditions therein.

To be in line with the notations of Lemma C.1, recall that viewing $\widehat{r}$ as fixed, the risk function is $\ell(\theta, x, y) = \bar{m}(x, y, \theta)$, and the population risk minimizer is $\theta^* = \bar{\theta} = \bar{\alpha}, \bar{\eta}(\cdot)$. We define the L_2-norm d on

$\Theta = \mathbb{R}^+ \times \Lambda_c^p(\mathcal{X})$ as $d(\theta, \theta') = \|\theta - \theta'\|_{L_2(\mathbb{P})} := ((\alpha - \alpha')^2 + \|\eta' - \eta\|^2_{L_2(\mathbb{P}_{\cdot \mid T=1})})^{1/2}$. By the positive density condition, we have $\|\cdot\|_{L_2(\lambda)} \asymp \|\cdot\|_{L_2(\mathbb{P})}$. Hence for $\theta, \bar{\theta} \in \Theta$, we have $\|\theta - \bar{\theta}\|_\infty = o(1)$ once $\|\theta - \bar{\theta}\|_{L_2(\mathbb{P})} = o(1)$. 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 λ is the Lebesgue measure. Therefore, sufficiently small $d(\theta - \bar{\theta})$ implies sufficiently small $\|\theta - \bar{\theta}\|_\infty$.

We first verify the equivalence of $d(\theta, \bar{\theta})^2$ and $\mathbb{E}[\ell(\theta, X, Y) - \ell(\theta^*, X, Y)]$. By the optimality of $\bar{\alpha}, \bar{\eta}(\cdot)$, we know that $\bar{\eta}(x)$ is also the per-x optimizer of the risk function at $\bar{\alpha}$, that is, for $\mathbb{P}_{X \mid T=1}$-almost all x,

$$\bar{\eta}(x) = \arg\min \eta \in \mathbb{R}\ \mathbb{E}\Big[\bar{m}\big(x, Y(1), \bar{\alpha}, \eta\big) \,\Big|\, X = x, T = 1\Big].$$

Suppose $d(\theta - \bar{\theta})$ is sufficiently small, which means $\|\theta - \bar{\theta}\|_\infty < \epsilon$ for sufficiently small $\epsilon > 0$. Therefore, by the second order expansion of $\bar{I}(x, a, b) = \mathbb{E}[\bar{m}(x, Y(1), a, b) \,|\, X = x, T = 1]$ in Assumption B.1, we have

$$\big|\bar{I}(x, y, \alpha, \eta(x)) - \bar{I}(x, y, \bar{\alpha}, \bar{\eta}(x)) - \nabla_{a,b}\bar{I}(x, y, \bar{\alpha}, \bar{\eta}(x))[\alpha - \bar{\alpha}, \eta(x) - \bar{\eta}(x)]\big| \geq \lambda \|(\alpha - \bar{\alpha}, \eta(x) - \bar{\eta}(x))\|_2^2.$$

Since $\mathbb{E}[\bar{I}(X, Y(1), \alpha, \eta(X)) \,|\, T = 1] = \ell(\theta, X, Y)$, invoking Jensen's inequality we have

$$\Big|\mathbb{E}\big[\ell(\theta, X, Y)\big] - \mathbb{E}\big[\ell(\bar{\theta}, X, Y)\big] - \mathbb{E}\big[\nabla_{a,b}\bar{I}(x, y, \bar{\alpha}, \bar{\eta}(X))[\alpha - \bar{\alpha}, \eta(X) - \bar{\eta}(X)]\big]\Big| \geq \lambda \cdot d(\theta, \bar{\theta})^2. \tag{11}$$

Since $\bar{m}$ is a convex function in (a, b) for all (x, y), the conditional expectation is also convex, and by the differentiability and exchangeability of differentiation and expectation, we have

$$\mathbb{E}\Big[\nabla_b \bar{m}\big(x, Y(1), \bar{\alpha}, \bar{\eta}(x)\big) \,\Big|\, X = x, T = 1\Big] = 0.$$

Similarly, $\bar{\alpha}$ is also the minimizer of the convex risk given $\bar{\eta}$, and the regularity conditions imply

$$\mathbb{E}\Big[\nabla_a \bar{m}\big(X, Y(1), \bar{\alpha}, \bar{\eta}(X)\big) \,\Big|\, T = 1\Big] = \nabla_a \mathbb{E}\Big[\bar{m}\big(X, Y(1), \bar{\alpha}, \bar{\eta}(X)\big) \,\Big|\, T = 1\Big] = 0.$$

Returning to (11), we then have $\mathbb{E}[\ell(\theta, X, Y) - \ell(\theta^*, X, Y)] \geq \lambda \cdot d(\theta, \bar{\theta})^2$ when $\theta \in \Theta$ and $d(\theta, \bar{\theta})$ is sufficiently small. On the other hand, when $d(\theta - \bar{\theta})$ is sufficiently small, the second-order expansion of $\bar{m}$ near $(\bar{\alpha}, \bar{\eta}(x))$ in Assumption B.1 implies

$$\begin{aligned}
&\Big|\mathbb{E}\big[\ell(\theta, X, Y)\big] - \mathbb{E}\big[\ell(\bar{\theta}, X, Y)\big] - \mathbb{E}\big[\nabla_{a,b}\bar{m}(X, Y, \bar{\alpha}, \bar{\eta}(X))[a - \bar{\alpha}, b - \bar{\eta}(X)]\big]\Big| \\
&\leq \mathbb{E}\big[g(X, Y)\|(a - \bar{\alpha}, b - \bar{\eta}(X))\|_2^2\big].
\end{aligned}$$

Here similar to previous arguments, $\mathbb{E}\big[\nabla_{a,b}\bar{m}(X, Y, \bar{\alpha}, \bar{\eta}(X))[a - \bar{\alpha}, b - \bar{\eta}(X)]\big] = 0$ by the exchangeability of differentiation and expectation; by the finite expectation of $g(x, y)$ in Assumption B.1, the right-hand side can be bounded as

$$\mathbb{E}\big[g(X, Y)\|(a - \bar{\alpha}, b - \bar{\eta}(X))\|_2^2\big] \leq \mathbb{E}\big[\mathbb{E}[g(X, Y) \,|\, X = x]\|(a - \bar{\alpha}, b - \bar{\eta}(X))\|_2^2\big] \leq M \cdot d(\theta, \bar{\theta})^2.$$

We have thus verified the equivalence of $d(\theta, \bar{\theta})$ and $|\mathbb{E}[\ell(\theta, X, Y)] - \mathbb{E}[\ell(\bar{\theta}, X, Y)]|$.

We then proceed to verify condition (i) of Lemma C.1. For $\theta \in \Theta$ such that $d(\theta, \bar{\theta})$ hence $\|\theta - \bar{\theta}\|_\infty$ is sufficiently small, since $\bar{m}(\theta, x, y) - \bar{m}(\bar{\theta}, x, y) \leq \bar{g}(x, y)\|\theta(x) - \bar{\theta}(x)\|_2$ where $\mathbb{E}[\bar{g}(x, Y)^2 \,|\, X = x] \leq M$ for all x, we have

$$\begin{aligned}
\mathrm{Var}\left(\ell(\theta, X, Y) - \ell(\bar{\theta}, X, Y)\right) &\leq \mathbb{E}\big[|\bar{m}(\theta, X, Y) - \bar{m}(\bar{\theta}, X, Y)|^2\big] \leq M\mathbb{E}\big[\big(\theta(X) - \theta^*(X)\big)^2\big] \\
&\leq \mathbb{E}\Big[\mathbb{E}\big[\bar{g}(X, Y)^2 \,\big|\, X\big] \cdot \big\|\theta(X) - \bar{\theta}(X)\big\|_2^2\Big] \leq M \cdot d(\theta, \bar{\theta}),
\end{aligned}$$

hence condition (i) is satisfied. For condition (ii), it suffices to let $U_n(x, y) = \bar{g}(x, y)$.

Therefore, applying Lemma C.1, we know that $d(\widehat{\theta}_n - \bar{\theta}) = O(\max\{\delta_n, \inf_{\theta' \in \Theta_n} d(\theta', \bar{\theta}))$. Using a similar calculation as in the proof of Theorem 1, we have

$$\delta_n \asymp \sqrt{\frac{J_n^d \log n}{n}},$$

and the approximation error using Θ_n is upper bounded as $\inf_{\theta' \in \Theta_n} d(\theta', \bar{\theta}) \leq O(J_n^p)$. We can thus set $J_n = (n/\log n)^{1/(2p+d)}$, so that $\|\widehat{\theta} - \bar{\theta}\|_{L_2(\mathbb{P})} = O_P((\log n/n)^{p/(2p+d)})$, which completes our proof. □

D.4 Proof of Theorem B.1

Proof. Proof of Theorem B.1 We are to show that, for each $j \in \{1,2,3,4\}$, $\sqrt{n}(\widehat{\mu}_{1,0}^{m,(j)} - \widehat{\mu}_{1,0}^{*}) = o(1)$, where we define

$$\widehat{\mu}_{1,0}^{*} = \frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\Big\{m\big(X_i,Y_i,\alpha^*,\eta^*(X_i),r_{0,1}(X_i)\big) - \alpha^* r_{0,1}(X_i)h(X_i)\Big\} + \frac{\alpha^*}{|\mathcal{I}_0^{(j)}|}\sum_{i\in\mathcal{I}_0^{(j)}} h(X_i).$$

Here we define $h(x) = \mathbb{E}[H(x,Y(1)) \,|\, X = x, T = 1]$, where

$$H(x,y) = \frac{\partial f^*}{\partial r}\big(x,y,\alpha^*,\eta^*(x),r_{0,1}(x)\big).$$

In the following, we supress the dependency on j as the result applies to all j. We also implicitly condition on $\mathcal{I}_1^{(j+1)} \cup \mathcal{I}_0^{(j+1)} \cup \mathcal{I}_1^{(j+2)} \cup \mathcal{I}_0^{(j+2)}$, so that $\widehat{\alpha}$, $\widehat{\eta}$, $\widehat{r}$ are viewed as fixed. Until further notice, we write $\mathcal{I}_1 = \mathcal{I}_1^{(j)}$ and $\mathcal{I}_0 = \mathcal{I}_0^{(j)}$ for notational simplicity, and write $n_1 = |\mathcal{I}_1|$, $n_0 = |\mathcal{I}_0|$.

With similar arguments as in the proof of Proposition B.4, when $\widehat{\eta}$ and $\bar{\eta}$ are sufficiently smooth, sufficiently small $\|\widehat{\theta} - \bar{\theta}\|_{L_2(\mathbb{P}_{X\,|\,T=1}} $ implies sufficiently small $\|\widehat{\theta} - \bar{\theta}\|_\infty$. By the second-order expansion at $\bar{\theta}(x) = (\bar{\alpha}, \bar{\eta}(x))$ in Assumption B.3, when $\|\widehat{\theta}(x) - \bar{\eta}(x)\|_2$ is sufficently small, we have

$$\begin{aligned}\Big|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i,Y_i,\widehat{\alpha},\widehat{\eta}(X_i),\widehat{r}(X_i)\big) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i)\big)\\ - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\nabla_{a,b} m\big(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i)\big)[\widehat{\alpha}-\bar{\alpha},\widehat{\eta}(X_i)-\bar{\eta}(X_i)]\Big| \le \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} g(X_i,Y_i)\big\|\widehat{\theta}(X_i)-\bar{\theta}(X_i)\big\|_2^2.\end{aligned}$$

Here by the optimiality of $\bar{\alpha}$ and $\bar{\eta}(x)$, we know that

$$\mathbb{E}\big[\nabla_a m\big(X_i,Y_i(1),\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i)\big)\,\big|\,T_i = 1\big] = 0$$

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

$$\mathbb{E}\big[\nabla_b m\big(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i)\big)\,\big|\,X_i = x, T_i = 1\big] = 0.$$

Consequently, we have

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\nabla_a m\big(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i)\big)(\widehat{\alpha}-\bar{\alpha}) = O_p\big(|\widehat{\alpha}-\bar{\alpha}|/\sqrt{n}\big) = o_P(1/\sqrt{n})$$

and

$$\begin{aligned}&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\nabla_a m\big(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i)\big)(\widehat{\eta}(X_i)-\bar{\eta}(X_i))\\ &= O_P\Big(\big\{\operatorname{Var}\big(\nabla_a m(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i))(\widehat{\eta}(X_i)-\bar{\eta}(X_i))\,\big|\,T_i=1\big)/n\big\}^{1/2}\Big) = o(1/\sqrt{n})\end{aligned}$$

since the latter, due to Assumption B.3, is further bounded as

$$\begin{aligned}&\operatorname{Var}\big(\nabla_b m(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i))(\widehat{\eta}(X_i)-\bar{\eta}(X_i))\big)\\ &\le \mathbb{E}\Big[\mathbb{E}\big[\nabla_b m(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),\widehat{r}(X_i))^2\,\big|\,X_i,T_i=1\big](\widehat{\eta}(X_i)-\bar{\eta}(X_i))^2\,\Big|\,T_i=1\Big] \le M\|\widehat{\eta}-\eta\|^2_{L_2(\mathbb{P}_{X\,|\,T=1})}.\end{aligned}$$

Furthermore, by Markov's inequality we have

$$\begin{aligned}&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} g(X_i,Y_i)\big\|\widehat{\theta}(X_i)-\bar{\theta}(X_i)\big\|_2^2\\ &= O_P\big(\mathbb{E}\big[g(X_i,Y_i)\big\|\widehat{\theta}(X_i)-\bar{\theta}(X_i)\big\|_2^2\,|\,T_i=1\big]\big) = O_P(\|\widehat{\theta}-\bar{\theta}\|_2^2) = o_P(1/\sqrt{n}),\end{aligned}$$

with the last rate implied by Assumption B.3. As a result, we have

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i, Y_i, \widehat{\alpha}, \widehat{\eta}(X_i), \widehat{r}(X_i)\big) = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big) + o_P(1/\sqrt{n}). \tag{12}$$

On the other hand, we write

$$\bar{H}(x,y) = \frac{\partial f^*}{\partial r}\big(x, y, \bar{\alpha}, \bar{\eta}(x), \widehat{r}^{(j)}(x)\big).$$

Then the expansion with respect to r in Assumption B.3 implies

$$\begin{aligned}
&\Big|m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big) - m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big) - \bar{\alpha}\bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\Big| \\
&= \Big|\bar{\alpha}\Big[f^*\Big(\frac{\widehat{r}(X_i)}{-\bar{\alpha}}\big(Y_i + \bar{\eta}(X_i)\big)\Big) - f^*\Big(\frac{r}{-\bar{\alpha}}\big(Y_i + \bar{\eta}(X_i)\big)\Big)\Big] \\
&\qquad - \bar{\alpha}\frac{\partial f^*}{\partial r}\big(x, y, \bar{\alpha}, \bar{\eta}(x), \widehat{r}^{(j)}(x)\big)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\Big| \\
&\le \bar{\alpha}\cdot \bar{g}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)^2,
\end{aligned} \tag{13}$$

where by the finite conditional expectation of $\bar{g}(X,Y)$ and the convergence rate of $\widehat{r}$, we have

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} \bar{g}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)^2 = O_P\big(\mathbb{E}[(\widehat{r}(X_i) - r_{1,0}(X_i))^2]\big) = o(1/\sqrt{n}).$$

Therefore, combining the above residual bound with (12), the difference between $\widehat{\mu}_{1,0}^{m,(j)}$ and $\widehat{\mu}_{1,0}^{*}$ can be written as

$$\begin{aligned}
\widehat{\mu}_{1,0}^{m,(j)} - \widehat{\mu}_{1,0}^{*} &= \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big) \\
&\quad + \frac{\bar{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1} \bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big) - \frac{\widehat{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\widehat{r}(X_i)\widehat{h}(X_i) + \frac{\widehat{\alpha}}{n_0}\sum_{i\in\mathcal{I}_0}\widehat{h}(X_i) \\
&\quad + \frac{\alpha^*}{n_1}\sum_{i\in\mathcal{I}_1} r_{0,1}(X_i)h(X_i) - \frac{\alpha^*}{n_0}\sum_{i\in\mathcal{I}_0} h(X_i) + o_P(1/\sqrt{n}).
\end{aligned} \tag{14}$$

We now proceed to treat the summations in (14). Firstly, we note that

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} r_{0,1}(X_i)\widehat{h}(X_i) + \frac{1}{n_0}\sum_{i\in\mathcal{I}_0}\widehat{h}(X_i) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} r_{0,1}(X_i)h(X_i) + \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} h(X_i) \\
&= \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\Big(r_{0,1}(X_i)\big(\widehat{h}(X_i) - h(X_i)\big) - \mathbb{E}\big[\widehat{h}(X) - h(X)\,|\,T=1\big]\Big) \\
&\qquad + \frac{1}{n_0}\sum_{i\in\mathcal{I}_0}\Big(\widehat{h}(X_i) - h(X_i) - \mathbb{E}\big[\widehat{h}(X) - h(X)\,|\,T=0\big]\Big) \\
&= O_P\big(\|r_{1,0}(\widehat{h} - h)\|_{L_2(\mathbb{P}_{X\,|\,T=1})}/\sqrt{n} + \|\widehat{h} - h\|_{L_2(\mathbb{P}_{X\,|\,T=0})}/\sqrt{n}\big) = o_P(1/\sqrt{n}),
\end{aligned}$$

where the first line uses the fact that $\mathbb{P}_{X\,|\,T=1}$ to $\mathbb{P}_{X\,|\,T=0}$ admits a covariate shift $r_{1,0}(\cdot)$, hence the two expectations are equivalent, and the two summations consist of i.i.d. mean zero samples. The last rate uses the fact that

$$\big\|r_{1,0}(\widehat{h} - h)\big\|_{L_2(\mathbb{P}_{X\,|\,T=1})} \le \big\|r_{1,0}(\widehat{h} - \bar{h})\big\|_{L_2(\mathbb{P}_{X\,|\,T=1})} + \big\|r_{1,0}(\bar{h} - h)\big\|_{L_2(\mathbb{P}_{X\,|\,T=1})} = o_P(1)$$

and similar for the other term. Again by the covariate shift $r_{1,0}(x)$ between two groups, we have

$$(\widehat{\alpha} - \alpha^*)\cdot\Big(\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} r_{0,1}(X_i)h(X_i) - \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} h(X_i)\Big) = O_P(|\widehat{\alpha} - \alpha^*|/\sqrt{n}) = o_P(1/\sqrt{n})$$

since $|\widehat{\alpha} - \alpha^*| \le |\widehat{\alpha} - \bar{\alpha}| + |\bar{\alpha} - \alpha^*| = o_P(1)$. Therefore, we have

$$\begin{aligned}
&\frac{\widehat{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\widehat{r}(X_i)\widehat{h}(X_i) - \frac{\widehat{\alpha}}{n_0}\sum_{i\in\mathcal{I}_0}\widehat{h}(X_i) \\
&= \frac{\widehat{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{0,1}(X_i)\big)\widehat{h}(X_i) + \frac{\alpha^*}{n_1}\sum_{i\in\mathcal{I}_1} r_{0,1}(X_i)h(X_i) - \frac{\alpha^*}{n_0}\sum_{i\in\mathcal{I}_0} h(X_i) + o_P(1/\sqrt{n}).
\end{aligned}$$

Combined with (14), the above error bounds imply

$$\begin{aligned}
\widehat{\mu}_{1,0}^{m,(j)} - \widehat{\mu}_{1,0}^{*} = &\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big) \\
&+ \frac{\bar{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big) \\
&- \frac{\widehat{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{0,1}(X_i)\big)\widehat{h}(X_i) + o_P(1/\sqrt{n}). \qquad (15)
\end{aligned}$$

We further note that since $\mathbb{E}[\bar{H}(x, Y(1)) \,|\, X = x, T = 1] \le M$ for all $x \in \mathcal{X}$, we have

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big) = O_P(\|\widehat{r} - r_{1,0}\|_{L_2(\mathbb{P}_{X\,|\,T=1})}),$$

hence by the convergence rate in Assumption B.3, we have

$$\frac{\widehat{\alpha} - \bar{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big) = O_P(|\widehat{\alpha} - \bar{\alpha}| \cdot \|\widehat{r} - r_{1,0}\|_{L_2(\mathbb{P}_{X\,|\,T=1})}) = o_P(1/\sqrt{n}).$$

As a result, we obtain

$$\begin{aligned}
&\frac{\bar{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big) - \frac{\widehat{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{0,1}(X_i)\big)\widehat{h}(X_i) \qquad (16) \\
&= \frac{\widehat{\alpha}}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\big(\bar{H}(X_i, Y_i) - \widehat{h}(X_i)\big) + o_P(1/\sqrt{n}).
\end{aligned}$$

The above term excluding the $\widehat{\alpha}$ factor can be decomposed as

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\big(\bar{H}(X_i, Y_i) - \widehat{h}(X_i)\big) \\
&= \underbrace{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\big(\bar{H}(X_i, Y_i) - \bar{h}(X_i)\big)}_{\text{(i)}} + \underbrace{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)\big(\bar{h}(X_i, Y_i) - \widehat{h}(X_i)\big)}_{\text{(ii)}}.
\end{aligned}$$

The first term is unbiased since $\mathbb{E}[\bar{H}(x, Y(1)) \,|\, X = x, T = 1] = \bar{h}(x)$ almost surely. We thus have

$$\begin{aligned}
\big|\text{(i)}\big| &= O_P\Big(\big\{\mathbb{E}\big[\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)^2\big(\bar{H}(X_i, Y_i) - \bar{h}(X_i)\big)^2 \,\big|\, T_i = 1\big]\big\}^{1/2}/\sqrt{n}\Big) \\
&= O_P(\|\widehat{r} - r_{1,0}\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})}/\sqrt{n}) = o_P(1/\sqrt{n})
\end{aligned}$$

since the conditional variance of $\bar{H}(X, Y)$ is uniformly bounded. By Cauchy-Schwarz inequality, the other term is bounded as

$$\begin{aligned}
\big|\text{(ii)}\big| &\le \sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)^2}\sqrt{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\big(\bar{h}(X_i, Y_i) - \widehat{h}(X_i)\big)^2} \\
&= O_P\big(\|\widehat{r} - r_{1,0}\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})} \cdot \|\widehat{h} - \bar{h}\|_{L_2(\mathbb{P}_{\cdot\,|\,T=1})}\big) = o_P(1/\sqrt{n})
\end{aligned}$$

by the convergence rates in Assumption B.3. Hence (16) is also $o_P(1/\sqrt{n})$. This further leads (15) to

$$\widehat{\mu}_{1,0}^{m,(j)} - \widehat{\mu}_{1,0}^{*} = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big) + o_P(1/\sqrt{n}). \tag{17}$$

Again with similar arguments as in the proof of Proposition B.4, when η^* and $\bar{\eta}$ are sufficiently smooth, sufficiently small $\|\theta^* - \bar{\theta}\|_{L_2(\mathbb{P}_{X\mid T=1}}$ implies sufficiently small $\|\theta^* - \bar{\theta}\|_\infty$. Therefore, the second-order expansion at $(\alpha^*, \eta^*(x))$ of Assumption B.3 implies

$$\begin{aligned}
&\Big|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big) - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big) \\
&\quad - \frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} \nabla_{a,b} m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)[\bar{\alpha} - \alpha^*, \bar{\eta}(X_i) - \eta^*(X_i)]\Big| \\
&\le \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} \bar{g}(X_i, Y_i)\big\|\bar{\theta}(X_i) - \theta^*(X_i)\big\|_2^2.
\end{aligned} \tag{18}$$

By the optimality of α^* and the exchangeability of differentiation and expectation, $\nabla_a m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)$ has mean zero, thus we have

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} \nabla_a m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)[\bar{\alpha} - \alpha^*] = O_P(|\bar{\alpha} - \alpha^*|/\sqrt{n}) = o_P(1/\sqrt{n}).$$

Meanwhile, since $\eta^*(x)$ is the per-x optimizer given α^*, $\nabla_b m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)$ has conditional mean zero, hence

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i\in\mathcal{I}_1^{(j)}} \nabla_b m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)[\bar{\eta}(X_i) - \eta^*(X_i)] \\
&= O_P\Big(\big\|\nabla_b m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)[\bar{\eta}(X_i) - \eta^*(X_i)]\big\|_{L_2(\mathbb{P}_{\cdot\mid T=1})}/\sqrt{n}\Big) \\
&= O_P\big(\|\bar{\eta} - \eta^*\|_{L_2(\mathbb{P}_{\cdot\mid T=1})}/\sqrt{n}\big) = o_P(1/\sqrt{n}),
\end{aligned}$$

where we use the fact that the second moment of $\nabla_b m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)$ conditional on X is uniformly upper bounded as stated in Assumption B.3. Also, the uniformly bounded conditional expectation of $\bar{g}(X, Y)$ implies

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} \bar{g}(X_i, Y_i)\big\|\bar{\theta}(X_i) - \theta^*(X_i)\big\|_2^2 = O_P(\|\bar{\theta} - \theta^*\|^2_{L_2(\mathbb{P}_{X\mid T=1})}) = o_P(1/\sqrt{n}).$$

Returning to (17), this leads to $\sqrt{n}\big(\widehat{\mu}_{1,0}^{m,(j)} - \widehat{\mu}_{1,0}^{*}\big) = o_P(1)$ for each fold.

We now return to the original notations for all folds of data, where $n_1 = |\mathcal{I}_1|$ and $n_0 = |\mathcal{I}_0$ represent all treated or control samples. Averaging over the four folds, we have $\sqrt{n}(\widehat{\mu}_{1,0}^{m,-} - \widehat{\mu}_{1,0}^{m,-,*}) = o_P(1)$, where

$$\widehat{\mu}_{1,0}^{m,-,*} = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\Big\{m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big) - \alpha^* r_{0,1}(X_i)h(X_i)\Big\} + \frac{\alpha^*}{|\mathcal{I}_0^{(j)}|}\sum_{i\in\mathcal{I}_0} h(X_i).$$

Since $\mathbb{E}[\widehat{\mu}_{1,0}^{m,0,*}] = \mu_{1,0}^{m,-}$, invoking Central Limit Theorem and Slutsky's lemma, we have $\sqrt{n}(\widehat{\mu}_{1,0}^{m,-} - \mu_{1,0}^{m,-}) \rightsquigarrow N(0, \mathrm{Var}(\phi_{1,-}^m(X, Y, T)))$, where the influence function is given by

$$\phi_{1,-}^m(X, Y, T) = \frac{T}{p_1}\Big\{m\big(X, Y, \alpha^*, \eta^*(X), r_{0,1}(X)\big) - \alpha^* r_{0,1}(X)h(X)\Big\} + \frac{1-T}{p_0}\alpha^* h(X),$$

with $p_1 = \mathbb{P}(T=1) = 1 - p_0$. It is also straightforward to see that

$$\begin{aligned}\operatorname{Var}\big(\phi^m_{1,-}(X,Y,T)\big) =& \frac{1}{p_1}\operatorname{Var}\Big(m\big(X,Y,\alpha^*,\eta^*(X),r_{0,1}(X)\big) - \alpha^* r_{0,1}(X)h(X)\,\Big|\,T=1\Big)\\ &+ \frac{1}{p_0}\operatorname{Var}\big(\alpha^* h(X)\,\big|\,T=0\big).\end{aligned}$$

Therefore, a consistent variance estimator can be constructed by plugging in their empirical counterparts. To be specific, for $i \in \mathcal{I}_1$ we let

$$\widehat{d}_{1,i} = m(X_i, Y_i, \widehat{\alpha}^{j[i]}, \widehat{\eta}^{j[i]}(X_i), \widehat{r}^{j[i]}(X_i)\big) - \widehat{\alpha}^{j[i]}\widehat{r}^{j[i]}(X_i)\widehat{h}^{j[i]}(X_i),$$

where $j[i]$ is the fold that sample i belongs to. It aims at approximating

$$d_{1,i} := m(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big) - \alpha^* r_{0,1}(X_i)h(X_i).$$

We show $\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}(\widehat{d}^2_{1,i} - d^2_{1,i}) = o_P(1)$ as an example. By Cauchy-Schwarz inequality,

$$\begin{aligned}\Big|\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\big(\widehat{d}^2_{1,i} - d^2_{1,i}\big)\Big| &= \Big|\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\big(\widehat{d}_{1,i} - d_{1,i}\big)^2 + \frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}} 2\big(\widehat{d}_{1,i} - d_{1,i}\big)d_{1,i}\Big|\\ &\le \frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\big(\widehat{d}_{1,i} - d_{1,i}\big)^2 + 2\sqrt{\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\big(\widehat{d}_{1,i} - d_{1,i}\big)^2}\cdot\sqrt{\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}} d^2_{1,i}}.\end{aligned}$$

It thus boils down to show that $\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\big(\widehat{d}_{1,i} - d_{1,i}\big)^2 = o_P(1)$. We supress the dependence on j to simplify notations. By the second-order expansion at $\bar{\theta}(x) = (\bar{\alpha}, \bar{\eta}(x))$, we have

$$\begin{aligned}&\Big|m\big(X_i, Y_i, \widehat{\alpha}, \widehat{\eta}(X_i), \widehat{r}(X_i)\big) - m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big)\Big|\\ &\le \Big|\nabla_{a,b} m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big)[\widehat{\alpha} - \bar{\alpha}, \widehat{\eta}(X_i) - \bar{\eta}(X_i)]\Big| + g(X_i, Y_i)\big\|\widehat{\theta}(X_i) - \bar{\theta}(X_i)\big\|_2^2.\end{aligned}$$

when $\|\widehat{\theta} - \bar{\theta}\|_{L_2(\mathbb{P}_{X\,|\,T=1})}$ is sufficently small. By the finite second moment condition and the consistency of $\widehat{\alpha}$ and $\widehat{\eta}$, we have

$$\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\Big|m\big(X_i, Y_i, \widehat{\alpha}, \widehat{\eta}(X_i), \widehat{r}(X_i)\big) - m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big)\Big|^2 = o_P(1).$$

Furthermore, the expansion with repsect to r in (13) implies

$$\begin{aligned}&\Big|m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big) - m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big)\Big|\\ &\le \bar{\alpha}\bar{H}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big) + \bar{\alpha}\cdot\bar{g}(X_i, Y_i)\big(\widehat{r}(X_i) - r_{1,0}(X_i)\big)^2,\end{aligned}$$

the finite moment conditions of which imply

$$\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\Big|m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), \widehat{r}(X_i)\big) - m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big)\Big|^2 = o_P(1).$$

Finally, the expansion around $(\alpha^*, \eta^*(x))$ similar to (18) leads to

$$\begin{aligned}&\Big|m\big(X_i, Y_i, \bar{\alpha}, \bar{\eta}(X_i), r_{1,0}(X_i)\big) - m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)\Big|\\ &\le \Big|\nabla_{a,b} m\big(X_i, Y_i, \alpha^*, \eta^*(X_i), r_{0,1}(X_i)\big)[\bar{\alpha} - \alpha^*, \bar{\eta}(X_i) - \eta^*(X_i)]\Big| + \bar{g}(X_i, Y_i)\big\|\bar{\theta}(X_i) - \theta^*(X_i)\big\|_2^2.\end{aligned}$$

By the consistency of $\bar{\theta}$ and finite moments, we thus have

$$\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\left|m\big(X_i,Y_i,\bar{\alpha},\bar{\eta}(X_i),r_{1,0}(X_i)\big)-m\big(X_i,Y_i,\alpha^*,\eta^*(X_i),r_{0,1}(X_i)\big)\right|^2=o_P(1).$$

Futhermore, we note that

$$\left|\widehat{\alpha}\widehat{r}(X_i)\widehat{h}(X_i)-\alpha^* r_{1,0}(X_i)h(X_i)\right|^2\le 2\widehat{\alpha}^2\left|\widehat{r}(X_i)\widehat{h}(X_i)-r_{1,0}(X_i)h(X_i)\right|^2+2|\widehat{\alpha}-\alpha^*|^2\cdot r_{1,0}(X_i)^2h(X_i)^2,$$

which is $o_P(1)$ by the consistency of estimators. Invoking Cauchy-Schwarz inequality, we thus obtain $\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\left(\widehat{d}_{1,i}-d_{1,i}\right)^2=o_P(1)$. This also implies $\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\left(\widehat{d}_{1,i}-d_{1,i}\right)=o_P(1)$, hence $\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\widehat{d}_{1,i}^2-\left(\frac{1}{|\mathcal{I}_1^{(j)}|}\sum_{i\in\mathcal{I}_1^{(j)}}\widehat{d}_{1,i}\right)^2$ is consistent for $\mathrm{Var}\left(m(X,Y,\alpha^*,\eta^*(X),r_{0,1}(X))-\alpha^* r_{0,1}(X)h(X)\,\middle|\,T=1\right)$. Similar consistency can be obtained for the other variance term, which completes the proof of the consistency of the variance estimator. Therefore, we conclude the proof of Theorem B.1. □

E Technical lemmas

Lemma E.1. *Let $\mathcal{F}_n$ be a sequence of σ-algebra, and let $A_n\ge 0$ be a sequence of nonnegative random variables. If $\mathbb{E}[A_n\,|\,\mathcal{F}_n]=o_P(1)$, then $A_n=o_P(1)$.*

Proof. Proof of Lemma E.1 By Markov's inequality, for any $\epsilon>0$, we have

$$B_n:=\mathbb{P}(A_n>\epsilon\,|\,\mathcal{F}_n)\le\frac{\mathbb{E}[A_n\,|\,\mathcal{F}_n]}{\epsilon}=o_P(1),$$

and $B_n\in[0,1]$ are bounded random variables. For any subsequence $\{n_k\}_{k\ge1}$ of $\mathbb{N}$, since $B_{n_k}\xrightarrow{P}0$, there exists a subsequence $\{n_{k_i}\}_{i\ge1}\subset\{n_k\}_{k\ge1}$ such that $B_{n_{k_i}}\xrightarrow{\text{a.s.}}0$ as $i\to\infty$. By the dominated convergence theorem, we have $\mathbb{E}[B_{n_{k_i}}]\to0$, or equivalently, $\mathbb{P}(A_{n_{k_i}}>\epsilon)\to0$. Therefore, for any subsequence $\{n_k\}_{k\ge1}$ of $\mathbb{N}$, there exists a subsequence $\{n_{k_i}\}_{i\ge1}\subset\{n_k\}_{k\ge1}$ such that $A_{n_{k_i}}\xrightarrow{P}0$ as $i\to\infty$. By the arbitrariness of $\{n_k\}_{k\ge1}$, we know $A_n\xrightarrow{P}0$ as $n\to\infty$, which completes the proof. □

References

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.

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

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120.

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

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

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

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.

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*.