Counterfactual Risk Assessments under Unmeasured Confounding*

Ashesh Rambachan† Amanda Coston‡ Edward H. Kennedy §

December 19, 2022

Abstract

Statistical risk assessments inform consequential decisions such as pretrial release in criminal justice, and loan approvals in consumer finance. Such risk assessments make *counterfactual* predictions, predicting the likelihood of an outcome under a proposed decision (e.g., what would happen *if* we approved this loan?). A central challenge, however, is that there may have been unmeasured confounders that jointly affected past decisions and outcomes in the historical data. This paper proposes a tractable mean outcome sensitivity model that bounds the extent to which unmeasured confounders could affect outcomes on average. The mean outcome sensitivity model partially identifies the conditional likelihood of the outcome under the proposed decision, popular predictive performance metrics (e.g., accuracy, calibration, TPR, FPR), and commonly-used predictive disparities. We derive their sharp identified sets, and we then solve three tasks that are essential to deploying statistical risk assessments in high-stakes settings. First, we propose a doubly-robust learning procedure for the bounds on the conditional likelihood of the outcome under the proposed decision. Second, we translate our estimated bounds on the conditional likelihood of the outcome under the proposed decision into a robust, plug-in decision-making policy. Third, we develop doubly-robust estimators of the bounds on the predictive performance of an existing risk assessment. We apply our methods to analyze a real-world credit-scoring task, illustrating how varying assumptions on unmeasured confounding leads to substantive changes in the credit score's predictions and evaluations of its predictive disparities.

*First version: September 2022. We thank Alexandra Chouldechova, Jonathan Roth, participants at the Bravo Center Workshop on the Economics of Algorithms at Brown University, INFORMS 2022 session on "Causality and Robustness", and the Yale econometrics seminar for useful comments. We especially thank our partners at Commonwealth Bank of Australia. All errors are our own.

†Microsoft Research New England: arambachan@microsoft.com

‡Carnegie Mellon University, Heinz College and Machine Learning Department: acoston@andrew.cmu.edu

§Carnegie Mellon University, Department of Statistics and Data Science: edward@stat.cmu.edu

1 Introduction

Statistical risk assessments inform high-stakes decisions by providing *counterfactual* predictions of the likelihood of an outcome under a proposed decision.[1] A central challenge, however, is that the available training and evaluation data only contain observed outcomes under historical decision-making policies. For example, pretrial risk assessments aim to predict the likelihood that a defendant would fail to appear in court *if* they were released prior to their trial; but we only observe whether a past defendant failed to appear in court if a judge decided to release them. Consumer credit scores aim to predict the likelihood an applicant would default on a loan *if* the applicant were approved; but we only observe whether a past applicant defaulted if the financial institution approved them and the applicant accepted the offered terms.

Many existing counterfactual methods for predicting individual risk (e.g., Schulam and Saria, 2017; Coston et al., 2020; Mishler, Kennedy and Chouldechova, 2021; Mishler and Kennedy, 2021) or individual causal effects (e.g., Shalit, Johansson and Sontag, 2017; Wager and Athey, 2018; Künzel et al., 2019; Nie and Wager, 2020; Kennedy, 2022*b*) tackle this challenge by making the strong assumption of *unconfoundedness*. Unconfoundedness requires that there are no unmeasured confounders that affected both historical decisions and outcomes, or equivalently that historical decisions were as-good-as randomly assigned conditional on recorded features. In many consequential decision-making settings, however, unconfoundedness is an unreasonable assumption because historical decisions may have been based on additional information that we do not have access to. Ignoring such unmeasured confounding would lead to inaccurate individual risk predictions and misleading evaluations of existing risk assessments.

This paper develops a comprehensive framework for learning and evaluating statistical risk assessments that is robust to unmeasured confounding. We propose the *mean outcome sensitivity model* (MOSM) as a nonparametric sensitivity analysis model for unmeasured confounding in settings where risk assessments are deployed. The MOSM bounds the extent to which unmeasured confounders could possibly affect the likelihood of the outcome in the population (e.g., "how much could default rates possibly vary between observably similar approved and rejected applicants?"). In this sense, the MOSM translates statistical assumptions about unmeasured confounding into interpetable units for practitioners. Over all levels of unmeasured confounding consistent with the MOSM, we robustly solve three tasks essential for deploying a statistical risk assessment in high-stakes settings: (i) estimate personalized risk predictions; (ii) translate personalized risk predictions into recommended interventions; and (iii) audit the predictive performance and disparities of a risk assessment.

1.1 Setting and background:

We consider a setting with data $O_i = (X_i, D_i, Y_i)$ for $i = 1, \ldots, n$ drawn i.i.d. from some joint distribution $\mathbb{P}(\cdot)$, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a feature vector, $D_i \in \{0, 1\}$ is a binary intervention that was determined by some historical decision-making policy, and $Y_i \in \{0, 1\}$ is the binary observed outcome. Let $Y_i(0), Y_i(1)$ denote potential outcomes under $D_i = 0, D_i = 1$ respectively, and the observed outcome satisfies $Y_i = Y_i(D_i)$. We assume $\mathbb{P}(Y_i(1) = 1) > 0$ and $\mathbb{P}(D_i = 1 \mid X_i) \geq \delta$ with probability one for

[1] These settings include, for example, child welfare screenings (Chouldechova et al., 2018; Saxena et al., 2020), consumer finance (Khandani, Kim and Lo, 2010; Einav, Jenkins and Levin, 2013; Blattner and Nelson, 2021; Fuster et al., 2022), criminal justice (Berk, 2012; Kleinberg et al., 2018), education (Smith, Lange and Huston, 2012; Sansone, 2019), and health care (Caruana et al., 2015; Choi et al., 2016; Chen et al., 2020) among many others.

some $\delta > 0$.

The goal in constructing a *counterfactual risk assessment* or *risk score* $s(\cdot)\colon \mathcal{X} \to [0,1]$ is to predict the conditional probability $Y_i(1) = 1$ given the features X_i. We therefore refer to $\mu^*(x) := \mathbb{P}(Y_i(1) = 1 \mid X_i = x)$ as the *target regression*. The goal in auditing an existing risk assessment $s(\cdot)$ is to estimate its *predictive performance*

$$\text{perf}(s;\beta) := \mathbb{E}[\beta_0(X_i;s) + \beta_1(X_i;s)Y_i(1)], \tag{1}$$

$$\text{perf}_+(s;\beta) := \mathbb{E}[\beta_0(X_i;s) \mid Y_i(1) = 1], \tag{2}$$

where $\beta_0(X_i;s), \beta_1(X_i;s)$ are user-specified functions that may depend on the features X_i and risk assessment $s(\cdot)$. We refer to (1) as the overall predictive performance of $s(X_i)$ and (2) as its positive class predictive performance. Analogously define $\text{perf}_-(s;\beta) := \mathbb{E}[\beta_0(X_i;s) \mid Y_i(1) = 0]$ to be the negative class predictive performance. As we discuss in Section 2, these predictive performance measures recover commonly used risk functions or predictive diagnostics for alternative choices of $\beta_0(X_i;s), \beta_1(X_i;s)$. They can also be used to audit the group fairness properties of a risk assessment (e.g., Mitchell et al., 2019; Barocas, Hardt and Narayanan, 2019).

Under unconfoundedness $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid X_i$, the target regression and the predictive performance measures are point identified using the observed outcome regression $\mu_1(X_i) := \mathbb{P}(Y_i = 1 \mid D_i = 1, X_i = x)$ or inverse propensity score weighting (e.g., Coston et al., 2020). Unconfoundedness, however, is particularly implausible in settings where counterfactual risk assessments are deployed. In settings like pretrial release and consumer lending, historical decisions were chosen by existing decision-makers that likely observed additional information relevant to the outcome $Y_i(1)$ but are not captured by the recorded features X_i. For example, judges interact with defendants during pretrial release hearings and may learn extra extenuating information that affected their release decision (Kleinberg et al., 2018; Arnold, Dobbie and Hull, 2020*b*; Rambachan, 2021). In consumer lending, an applicant's decision to accept an offered loan may depend on whether they secured a credit offer at a competing financial institution.

1.2 Contributions

In this paper, we propose a flexible, nonparametric mean outcome sensitivity model (MOSM) for unmeasured confounding that is natural in high-stakes settings. The MOSM bounds the extent to which the likelihood of the outcome $Y_i(1) = 1$ could be affected by unmeasured confounders conditional on the observed features X_i – formally, pointwise bounds on the difference $\mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i) - \mathbb{P}(Y_i(1) = 1 \mid D_i = 1, X_i)$. Since practitioners already model and evaluate risk predictions in these settings, the MOSM enables them to directly translate their intuitions about how much risk could plausibly vary in the population into statistical assumptions on unmeasured confounding. We offer several ways to specify such bounds under the MOSM in practice. For example, we show that the MOSM is implied by the existence of an instrumental variable for historical decisions (e.g., Manski, 1994; Balke and Pearl, 1997).

Under the MOSM, the target regression, predictive performance measures, and predictive disparity measures are partially identified. We derive their sharp identified sets under the MOSM. We provide closed-form expressions for the smallest and largest values of the target regression and overall predictive performance measures that are compatible with the MOSM, and show that the sharp bounds on the positive class (and negative class) predictive performance can be characterized by linear-fractional programs. We then solve three tasks that are essential to deploying counterfactual risk assessments in high-stakes settings.

Our first task is to estimate the sharp bounds on the target regression $\mu^*(x)$ under the MOSM. In Section 3, we develop nonparametric estimators for the sharp bounds, which we refer to as DR-Learners, that leverage sample-splitting and take the form of two-stage regression procedures (e.g., Foster and Syrgkanis, 2020; Kennedy, 2022*b*). The first stage uses one fold of the data to construct nonparametric estimates of nuisance functions. The second stage applies the estimated nuisance functions on the other fold to construct a pseudo-outcome regression estimator based on efficient influence functions. We derive the integrated mean square error convergence rate of our DR-Learners to the true bounds relative to that of an oracle non-parametric regression procedure under generic assumptions. When the oracle error is small, our DR-Learners converge to the true bounds quickly whenever the first-stage nuisance functions are estimated at sufficiently fast rates, which are achievable using classic nonparametric regression techniques or modern machine learning methods. This result is agnostic – it applies to any choice of nonparametric estimators of the nuisance functions in the first stage and for a large class of nonparametric regression procedures in the second stage. To prove this result, we build on Kennedy (2022*b*), and provide a model-free oracle inequality for the $L_2(\mathbb{P})$-error of nonparametric regression with estimated pseudo-outcomes that may be of independent interest.

Since counterfactual risk assessments are typically deployed to inform existing decision-makers about a possible intervention, our second task is to use the historical data to make robust recommended interventions. We evaluate the performance of a plug-in recommendation rule that thresholds our DR-Learners of the target regression bounds in Section 4 by analyzing its worst-case performance across all levels of unmeasured confounding consistent with the MOSM. We derive bounds on the worst-case performance of our plug-in rule relative to the optimal (infeasible) max-min recommendation rule. This bound implies that the plug-in decision rule is asymptotically max-min optimal again whenever the first-stage nuisance functions are estimated at sufficiently fast rates.

Our final task is to robustly audit or evaluate the predictive performance ($\text{perf}(s; \beta)$ or $\text{perf}_+(s; \beta)$) and predictive disparities of an existing risk assessment $s(X_i)$ under the MOSM in Section 5. Our estimators for the sharp bounds on overall predictive performance have a closed-form. We derive their rates of convergence, and provide conditions under which they are $\sqrt{n}-$consistent and asymptotically normally distributed. Our estimators for the sharp bounds on positive class predictive performance solve a sample linear-fractional program, and we derive their rates of convergence. Our estimators leverage efficient influence functions and sample-splitting to control bias from the nonparametric estimation of first-stage nuisance functions, and therefore allow the use of complex machine learning estimators (e.g., Robins et al., 2008; Zheng and van der Laan, 2011; Chernozhukov et al., 2018) in estimating the scalar, predictive performance measure of interest.

Altogether, our framework provides a full pipeline for the learning and evaluation of counterfactual risk assessments under unmeasured confounding. We illustrate our theoretical analysis of these methods in Monte Carlo simulations. Finally, we apply our framework to a real-world credit-scoring task, showing how our methods can be used to develop a confounding-robust credit risk score and robustly audit the predictive disparities of an existing credit score. We find that varying the assumptions on the strength of unmeasured confounding leads to substantive changes in the credit score's predictions, and our evaluations of its predictive disparities.

1.3 Related work

This paper relates to a vast literature on sensitivity analysis in causal inference. One popular approach assumes the existence of some unmeasured confounder U_i that satisfies $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \{X_i, U_i\}$ and bounds how much the unmeasured confounder may affect decisions. For example, Rosenbaum's Γ-sensitivity model bounds the extent to which the true odds of treatment $\mathbb{P}(D_i = 1 \mid X_i, U_i)/\mathbb{P}(D_i = 0 \mid X_i, U_i)$ may vary across values of the unmeasured confounder $U_i = u$, $U_i = u'$ (e.g., Rosenbaum, 1987, 2002). Yadlowsky et al. (2018) derives sharp bounds on the average treatment effect and conditional average treatment effect under Rosenbaum's Γ-sensitivity model, developing nonparametric estimators for the bounds. Zhang et al. (2020) robustly ranks alternative treatment assignment rules under Rosenbaum's Γ-sensitivity model. Tan (2006)'s marginal sensitivity model bounds the extent to which the true odds of treatment may differ from the observed odds $\mathbb{P}(D_i = 1 \mid X_i)/\mathbb{P}(D_i = 0 \mid X_i)$. A recently active literature studies robust estimation/inference on average treatment effects, conditional average treatment effects, and policy learning under the marginal sensitivity model – for example, see Kallus, Mao and Zhou (2018); Zhao, Small and Bhattacharya (2019); Dorn and Guo (2021); Dorn, Guo and Kallus (2021); Kallus and Zhou (2021); Jin, Ren and Candès (2021); Nie, Imbens and Wager (2021); Sahoo, Lei and Wager (2022).[2] In settings where risk assessments are deployed, historical decisions were made by prior decision makers, such as judges, doctors, or managers. It may therefore be difficult to place assumptions on how unmeasured confounders may have affected past decision-making, but easier to reason about how they may have possibly affected outcomes.

In this sense, our work sits in a line of causal inference research that proposes sensitivity analysis models directly on outcome distributions $Y_i(1) \mid \{D_i = 0, X_i\}$ vs. $Y_i(1) \mid \{D_i = 1, X_i\}$. Brumback et al. (2004) consider six parametric functional forms for specifying the exact relationship between these conditional distributions. Díaz and van der Laan (2013); Luedtke, Diaz and van der Laan (2015); Díaz, Luedtke and van der Laan (2018) assume the difference in means of the potential outcome under treatment versus control is bounded by a user-specified, scalar quantity. Robins, Rotnitzky and Scharfstein (2000*a*); Franks, D'Amour and Feller (2019); Scharfstein et al. (2021) assume that the unidentified distribution $Y_i(1) \mid \{D_i = 0, X_i\}$ is some known transformation ("tilting function") of the identified distribution $Y_i(1) \mid \{D_i = 1, X_i\}$. In practice, practitioners may lack sufficient knowledge to exactly and fully specify the relationship between these conditional distributions. Any particular choice of the tilting function may therefore itself be misspecified, and it is common for users to only report a few choices. In contrast, the MOSM considers all joint distributions that are consistent with the observable data and the user's specified bounds. Furthermore, our sensitivity analysis for statistical risk assessments, predictive performance measures and predictive disparities is novel relative to both of these literatures.

More broadly, we argue that the MOSM complements alternative sensitivity analysis models for

[2] Recently, Jin, Ren and Zhou (2022) proposed the f-sensitivity model as a refinement of the marginal sensitivity model.

violations of unconfoundedness. The effective, reliable, and safe use of statistical risk assessments in high-stakes settings requires there to be a suite of sensitivity analysis models that can applied off-the-shelf depending on what is most intuitive/applicable to the practitioner. We formally discuss the relationship between the MOSM and these existing sensitivity analysis frameworks in Section 8, showing how practitioners can map between the MOSM and these existing frameworks.

2 The mean outcome sensitivity model

We consider a setting with data $O_i = (X_i, D_i, Y_i)$ for $i = 1, \dots, n$ drawn i.i.d. from some joint distribution $\mathbb{P}(\cdot)$, where $Y_i = Y_i(D_i)$ for potential outcomes $Y_i(0), Y_i(1)$. Our tasks are to use $\mathcal{O} = \{O_i\}_{i=1}^n$ to (i) estimate a new counterfactual risk assessment; (ii) provide personalized recommendations for future interventions; or (iii) audit the predictive performance of an existing counterfactual risk assessment.[3]

Example (Consumer lending)**.** A financial institution observes historical data on past loan applicants, where X_i contains applicant information such as their reported income, D_i is whether the applicant was granted a loan, and $Y_i = Y_i(1)D_i$ is whether the applicant defaulted on the loan if they were granted ($Y_i(0) := 0$ since applicants that were not granted the loan cannot default). We use this data to either audit an existing credit score or construct a new credit score that predicts the likelihood a new applicant would default on a loan, $Y_i(1) = 1$ (e.g., Blattner and Nelson, 2021; Coston, Rambachan and Chouldechova, 2021; Fuster et al., 2022). ▲

Example (Pretrial release)**.** A pretrial release system observes historical data on past defendants, where X_i contains defendant information such as their current charge and prior conviction history, D_i is whether the defendant was released prior to their trial, and $Y_i = Y_i(1)D_i$ is whether the defendant failed to appear in court if they were released ($Y_i(0) := 0$ since detained defendants cannot fail to appear in court). We use this data to either audit an existing pretrial risk score or construct a new counterfactual pretrial risk score that predicts the likelihood a new defendant would fail to appear in court, $Y_i(1) = 1$ (e.g., Kleinberg et al., 2018; Jung et al., 2020*b*,*a*; Arnold, Dobbie and Hull, 2020*a*). ▲

Notation: We write sample averages of a random variable V_i as $\mathbb{E}_n[V_i] := n^{-1}\sum_{i=1}^n V_i$. Denote the observed propensity scores as $\pi_d(x) := \mathbb{P}(D_i = d \mid X_i = x)$ for $d \in \{0,1\}$. Let $\|\cdot\|$ denote the appropriate L_2-norm by context. That is, $\|f(\cdot)\| = \left(\int f(v)^2 dP(v)\right)^{1/2}$ for a measurable function $f(\cdot)$ taking values in $\mathbb{R}$, and $\|v\| = \left(\sum_{j=1}^k v_j^2\right)^{1/2}$ for a vector $v \in \mathbb{R}^k$.

2.1 Target regression and predictive performance measures

The goal in constructing a counterfactual risk assessment is to estimate the target regression $\mu^*(x) := \mathbb{P}(Y_i(1) = 1 \mid X_i = x)$. The goal in auditing an existing risk assessment $s(X_i)$ is to estimate various predictive performance measures $\mathrm{perf}(s;\beta) := \mathbb{E}[\beta_0(X_i;s) + \beta_1(X_i;s)Y_i(1)]$ and $\mathrm{perf}_+(s;\beta) := \mathbb{E}[\beta_0(X_i;s) \mid Y_i(1) = 1]$, where $\beta_0(X_i), \beta_1(X_i) \in \mathbb{R}$ are user-specified functions of X_i. As shorthand, write $\beta_{0,i} := \beta_0(X_i)$ and $\beta_{1,i} := \beta_1(X_i)$.

[3] Our results on auditing directly extend to the evaluation of a counterfactual decision rule $d(\cdot)\colon \mathcal{X} \to \{0,1\}$. In many cases, such a decision rule is constructed by thresholding a counterfactual risk score – that is, $d(x) = 1\{s(x) \leq \tau\}$ for some $\tau \in [0,1]$.

For alternative choices of $\beta_0(X_i; s)$ and $\beta_1(X_i; s)$, these predictive performance measures recover commonly used risk functions or predictive diagnostics.

Example 1 (MSE, accuracy, cross-entropy, calibration, and failure rate)**.**

a. For $\beta_0(X_i) = s^2(X_i)$ and $\beta_1(X_i) = 1 - 2s(X_i)$, $\text{perf}(s; \beta) = \mathbb{E}[(s(X_i) - Y_i(1))^2]$ is the *mean square error* of $s(X_i)$.

b. For $\beta_0(X_i) = 1 - s(X_i)$ and $\beta_1(X_i) := 2s(X_i) - 1$, $\text{perf}(s, \beta) = \mathbb{E}[s(X_i)Y_i(1) + (1 - s(X_i))(1 - Y_i(1))]$ is the *accuracy* of $s(X_i)$.

c. For $\beta_0(X_i) = -\log(1 - s(X_i))$ and $\beta_1(X_i) = \log(1 - s(X_i)) - \log(s(X_i))$, $\text{perf}(s; \beta) = -\mathbb{E}[Y_i(1)\log(s(X_i)) + (1 - Y_i(1))\log(1 - s(X_i))]$ is the *cross-entropy* of $s(X_i)$.

d. The *calibration* of $s(X_i)$ at prediction bin $[r_1, r_2] \subseteq [0, 1]$ is $\mathbb{E}[Y_i(1) \mid r_1 \leq s(X_i) \leq r_2] := \text{perf}(s; \beta)$ for $\beta_0(X_i) := 0$, $\beta_1(X_i) := \frac{1\{r_1 \leq s(X_i) \leq r_2\}}{\mathbb{E}[1\{r_1 \leq s(X_i) \leq r_2\}]}$ assuming $P(r_1 \leq s(X_i) \leq r_2) > 0$.

Example 2 (TPR and FPR)**.** For $\beta_0(X_i) = s(X_i)$, the *true positive rate* of $s(X_i)$ is $\mathbb{E}[s(X_i) \mid Y_i(1) = 1] = \text{perf}_+(s; \beta)$, and the *false positive rate* of $s(X_i)$ is $\mathbb{E}[s(X_i) \mid Y_i(1) = 0] = \text{perf}_-(s; \beta)$.

Example 3 (ROC curve)**.** The *true positive rate* at threshold $\tau \in [0, 1]$ is $\mathbb{E}[1\{s(X_i) \geq \tau\} \mid Y_i(1) = 1] = \text{perf}_+(s; \beta_\tau)$ for $\beta_\tau(X_i) = 1\{s(X_i) \geq \tau\}$. The *false positive rate* at threshold $\tau \in [0, 1]$ is analogously $\mathbb{E}[1\{s(X_i) \geq \tau\} \mid Y_i(1) = 0] = \text{perf}_-(s; \beta_\tau)$. The *ROC curve* of $s(X_i)$ is the set $\{(\text{perf}_-(s; \beta_\tau), \text{perf}_+(s; \beta_\tau) : \tau \in [0, 1]\}$.

These predictive performance measures are useful to evaluate the group fairness properties of a risk assessment (e.g., Mitchell et al., 2019). More concretely, suppose there is a binary sensitive attribute $G_i \in \{0, 1\}$ with $X_i = (\bar{X}_i, G_i)$ (e.g., ethnicity, gender, race, etc). Define the overall predictive performance of $s(X_i)$ on group $G_i = g$ as $\text{perf}_g(s; \beta) := \mathbb{E}[\beta_0(X_i) + \beta_1(X_i)Y_i(1) \mid G_i = g]$. The *overall predictive disparity* of the risk assessment is

$$\text{disp}(s; \beta) := \text{perf}_1(s; \beta) - \text{perf}_0(s; \beta). \tag{3}$$

The class-specific predictive performance on group $G_i = g$, $\text{perf}_{+,g}(s; \beta)$ and $\text{perf}_{-,g}(s; \beta)$, and the class-specific predictive disparities, $\text{disp}_+(s; \beta)$ and $\text{disp}_-(s; \beta)$, are defined analogously. By analyzing the difference in predictive performance measures across groups, the user can summarize average violations of widely-used predictive fairness definitions.

Example 4 (Equality of opportunity)**.** The risk assessment $s(X_i)$ satisfies *equality of opportunity* or balance for the positive class if $s(X_i) \perp\!\!\!\perp G_i \mid \{Y_i(1) = 1\}$ (e.g., Hardt, Price and Srebro, 2016; Chouldechova, 2017). The positive class predictive disparity $\text{disp}_+(s; \beta)$ for $\beta_0(X_i) = s(X_i)$ measures the difference in average risk assessments across groups given $Y_i(1) = 1$.

Example 5 (Bounded group loss)**.** For $\beta_0(X_i), \beta_1(X_i)$ as defined in Example 1, the risk assessment $s(X_i)$ violates *bounded group* accuracy, MSE, or cross-entropy for some $\epsilon > 0$ if either $\text{perf}_g(s; \beta) \geq \epsilon$ for $g \in \{0, 1\}$ (e.g., Agarwal, Dudík and Wu, 2019).

2.2 The mean outcome sensitivity model

Since $Y_i(1)$ is only observed under intervention $D_i = 1$, the target regression $\mu^*(x)$ and predictive performance measures $\mathrm{perf}(s;\beta), \mathrm{perf}_+(s;\beta)$ are not point identified without further assumptions. Rather than assuming the historical decisions were unconfounded, we propose an interpretable relaxation that we call the *mean outcome sensitivity model* (MOSM). Under the MOSM, the user bounds the extent to which the outcome $Y_i(1)$ could be affected by unmeasured confounders on average.

Let $\delta(X_i) := \mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i) - \mathbb{P}(Y_i(1) = 1 \mid D_i = 1, X_i)$ denote the difference in the probability $Y_i(1) = 1$ given $D_i = 0$ and $D_i = 1$ conditional on X_i. Since $Y_i(1)$ is unobserved if $D_i = 0$, neither $\mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i)$ nor $\delta(X_i)$ is identified. The mean outcome sensitivity model specifies pointwise bounds on the difference $\delta(X_i)$.

Assumption 2.1 (Mean outcome sensitivity model)**.** There exists bounding functions $\underline{\delta}(x), \overline{\delta}(x)\colon \mathcal{X} \to [-1, 1]$ satisfying $\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\underline{\delta}(X_i)] > 0$ and

$$\underline{\delta}(x) \leq \delta(x) \leq \overline{\delta}(x) \text{ for all } x \in \mathcal{X}. \tag{4}$$

Let Δ be the set of all functions $\delta(\cdot)$ satisfying (4), and write $\underline{\delta}_i := \underline{\delta}(X_i)$, $\overline{\delta}_i := \overline{\delta}(X_i)$.[4]

In the consumer lending setting, the MOSM bounds how much the probability of default may differ among applicants that were not granted a loan relative to observably similar applicants that were granted a loan. In the pretrial release example, the MOSM bounds how much the failure to appear rate may differ between observably similar detained defendants and released defendants. The MOSM nests the assumption of no unmeasured confounding by setting $\underline{\delta}(x) = \overline{\delta}(x) = 0$ for all x.

2.3 Choice of bounding functions

The choice of bounding functions $\underline{\delta}(\cdot), \overline{\delta}(\cdot)$ is crucial to the specification of the MOSM. We provide a few examples of how users may specify these bounds in practice.

Stratified outcome bounds: The user may specify the bounding functions by discretizing the feature space into strata, and then using domain knowledge to directly specify outcome bounds within each stratum.

Suppose that for some known stratification function $\kappa(\cdot)\colon \mathcal{X} \to \{1, \ldots, K\}$ and constants $\underline{\delta}_k, \overline{\delta}_k$ for $k = 1, \ldots, K$, the bounding functions further satisfy

$$\underline{\delta}(x) = \underline{\delta}_{\kappa(x)} \text{ and } \overline{\delta}(x) = \overline{\delta}_{\kappa(x)} \text{ for all } x \in \mathcal{X}. \tag{5}$$

Let $\Delta(\kappa)$ be the set of all functions $\delta(\cdot)$ that satisfy (5). The stratification function $\kappa(x)$ describes the user's domain-specific knowledge about which coarse strata summarize how unobserved confounders affect the outcome $Y_i(1)$ on average. In the consumer lending example, it may be that most of the variation in difference of default rates between rejected and approved applicant's is summarized by small

[4] By placing bounds on the difference in the probability $Y_i(1) = 1$ between $D_i = 0$ and $D_i = 1$ conditionally on X_i, the MOSM can be seen as the covarate-conditional generalization of the bounding approach taken in Luedtke, Diaz and van der Laan (2015) for average treatment effects. While covariates have been previously used in the sensitivity model of Brumback et al. (2004), they made parametric assumptions that our approach avoids.

set of known income or wealth brackets.[5] In the pretrial release example, much of the variation in the difference of failure to appear rates between released and detained defendants may be summarized by the arresting charge category (e.g., violent vs. non-violent charges).

Nonparametric outcome regression bounds: The user may wish to avoid having to specify a particular choice of strata, and instead place bounds directly in terms of the true, nonparametric outcome regression $\mu_1(x)$. Our framework allows the user to specify rich bounds of this form. For some choices $\underline{\Gamma}, \overline{\Gamma} > 0$, define

$$\underline{\delta}(x) = (\underline{\Gamma} - 1)\,\mu_1(x), \text{and } \overline{\delta}(x) = \left(\overline{\Gamma} - 1\right)\mu_1(x). \tag{6}$$

Let $\Delta(\Gamma)$ denote the set of bounding functions $\delta(\cdot)$ that satisfies these bounds.[6]

This choice implies that $\mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i)$ cannot be too different than the outcome regression $\mu_1(x)$, and satisfies $\underline{\Gamma}\mu_1(x) \leq \mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i) \leq \overline{\Gamma}\mu_1(x)$. In the pretrial release example, setting $\underline{\Gamma} = 2$ and $\overline{\Gamma} = 3$ implies we are willing to assume that detained defendants are no less risky than released defendants, but simultaneously they cannot be more than twice as risky as released defendants. Rambachan (2021) refers to such an assumption as "direct imputation," and it generalizes common strategies used to evaluate risk assessment tools in the criminal justice system. For example, Kleinberg et al. (2018), and Jung et al. (2020*a*) report results by assuming that the unobserved failure to appear rate among detained defendants is equal to some known function of the observed failure to appear rate among released defendants. As we show in Section 8, nonparametric outcome regression bounds are equivalent to common models for sensitivity analysis on unobserved confounding such as marginal sensitivity models.

Instrumental variable bounds: The existence of an instrumental variable that generates random variation in historical interventions implies the MOSM (Manski, 1994; Balke and Pearl, 1997). Such instrumental variables are common in settings where risk assessments are deployed. A classic example arises through the random assignment of judges to cases in the pretrial release system (e.g., Kleinberg et al., 2018; Arnold, Dobbie and Hull, 2020*b*,*a*; Rambachan, 2021), where an observed judge identifier is an instrument Z_i for the historical release decision D_i.[7,8]

Proposition 2.1. *Suppose* $O_i = (X_i, Z_i, D_i, Y_i) \sim P(\cdot)$ *i.i.d. for* $i = 1, \ldots, n$*, where* $Z_i \in \mathcal{Z}$ *has finite support and satisfies* $(Y_i(0), Y_i(1)) \perp Z_i \mid X_i$*. Define* $\underline{\delta}_z(x) = \left(\mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] - \mu_1(x)\right)/\pi_0(x)$

[5] The choice of stratification function relates to reject inference procedures used by industry practitioners in consumer finance (e.g., see Hand and Henley, 1993; Zeng and Zhao, 2014), which apply coarse adjustments to observed default rates among accepted applicants to impute the missing default rates among rejected applicants.

[6] Nonparametric outcome regression bounds can be combined with stratified outcome bounds. For example, for stratification function $\kappa(\cdot)$ and constants $\underline{\Gamma}_k, \overline{\Gamma}_k$ for $k = 1, \ldots, K$, the user may assume $\underline{\delta}(x) = \underline{\Gamma}_{\kappa(x)}\mu_1(x)$, $\overline{\Gamma}_{\kappa(x)}\mu_1(x)$ for all $x \in \mathcal{X}$.

[7] Lakkaraju et al. (2017) propose a "contraction procedure" that uses the random assignment of decision-makers to evaluate the performance of a risk assessments in the presence of unobserved confounding. Contraction only delivers point estimates of the failure rate of the risk assessment (see Example 1) at particular choices of threshold τ. In contrast, we sharply bound $\delta(x)$ using an instrument, which in turn enables the user to construct sharp bounds on the target regression, any overall predictive performance measure $\text{perf}(s;\beta)$ or any class-specific performance measure $\text{perf}_+(s;\beta)$, $\text{perf}_-(s;\beta)$.

[8] In related work, Qiu et al. (2021) and Pu and Zhang (2021) analyze optimal individual assignment rules when there exists a binary instrumental variable for past decisions. By establishing that the existence of such an instrumental variable implies the MOSM, our results enable users to bound the target regression, develop robust recommendation rules, and bound predictive performance of a given risk score.

and $\overline{\delta}_z(x) = (\pi_0(x,z) + \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] - \mu_1(x))/\pi_0(x)$ *for any* $z \in \mathcal{Z}$. *Then, for all* $x \in \mathcal{X}$,

$$\underline{\delta}_z(x) \leq \delta(x) \leq \overline{\delta}_z(x).$$

Let $\Delta(z)$ *denote the set of bounding functions* $\delta(\cdot)$ *satisfying these bounds for some* $z \in \mathcal{Z}$.

2.4 Sharp bounds on the target regression and predictive performance measures

The target regression and predictive performance measures are bounded under the MOSM, and we next derive their sharp bounds.

Observe that the target regression can be written as $\mu^*(x) = \mu_1(x) + \pi_0(x)\delta(x)$. We can therefore rewrite the predictive performance measures for a given risk assessment as

$$\text{perf}(s;\beta) = \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}\pi_0(X_i)\delta(X_i)] \tag{7}$$

$$\text{perf}_+(s;\beta) = \mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i)]^{-1}\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i)] \tag{8}$$

Define $\mathcal{H}(\mu^*(x);\Delta) = \{m \colon m(x) = \mu_1(x) + \delta(x)\pi_0(x) \text{ for } \delta \in \Delta\}$ to be the set of all target regression values consistent with the MOSM. Analogously, define $\mathcal{H}(\text{perf}(s;\beta);\Delta) = \{\text{perf}(s;\beta) \text{ satisfying (7) for } \delta \in \Delta\}$ and $\mathcal{H}(\text{perf}_+(s;\beta);\Delta) = \{\text{perf}_+(s;\beta) \text{ satisfying (8) for } \delta \in \Delta\}$. The sharp set of target regression values and predictive performance measures that are consistent with the MOSM can be characterized by closed intervals.

Lemma 2.1. *Suppose Assumption 2.1 is satisfied. Then,*

$$\begin{aligned}
&\mathcal{H}(\mu^*(x);\Delta) = \left[\underline{\mu}^*(x;\Delta), \overline{\mu}^*(x;\Delta)\right] \text{ for all } x \in \mathcal{X},\\
&\mathcal{H}(\mathit{perf}(s;\beta);\Delta) = \left[\underline{\mathit{perf}}(s;\beta,\Delta), \overline{\mathit{perf}}(s;\beta,\Delta)\right],\\
&\mathcal{H}(\mathit{perf}_+(s;\beta);\Delta) = \left[\underline{\mathit{perf}}_+(s;\beta,\Delta), \overline{\mathit{perf}}_+(s;\beta,\Delta)\right],
\end{aligned}$$

where $\overline{\mu}^*(x;\Delta) = \mu_1(x) + \pi_0(x)\overline{\delta}(x)$, $\underline{\mu}^*(x;\Delta) = \mu_1(x) + \pi_0(x)\underline{\delta}(x)$, *and*

$$\begin{aligned}
\overline{\mathit{perf}}(s;\beta,\Delta) &= \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}\pi_0(X_i)\left(1\{\beta_{1,i} > 0\}\overline{\delta}_i + 1\{\beta_{1,i} \leq 0\}\underline{\delta}_i\right)],\\
\underline{\mathit{perf}}(s;\beta,\Delta) &= \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}\pi_0(X_i)\left(1\{\beta_{1,i} \leq 0\}\overline{\delta}_i + 1\{\beta_{1,i} > 0\}\underline{\delta}_i\right)],\\
\overline{\mathit{perf}}_+(s;\beta,\Delta) &= \sup_{\delta(\cdot)\in\Delta} \mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i)]^{-1}\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i)],\\
\underline{\mathit{perf}}_+(s;\beta,\Delta) &= \inf_{\delta(\cdot)\in\Delta} \mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i)]^{-1}\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i)].
\end{aligned}$$

In Appendix C, we derive bounds on the predictive disparities of the risk assessment $s(X_i)$ under the MOSM.

3 Estimating the target regression bounds under the mean outcome sensitivity model

In this section, we propose estimators for the bounds $[\underline{\mu}^*(x;\Delta), \overline{\mu}^*(x;\Delta)]$ on the target regression under the MOSM. Following the heterogeneous treatment effects literature (e.g., Künzel et al., 2019; Nie and

Wager, 2020; Kennedy, 2022*b*), we refer to our estimators as "DR-Learners" since they incorporate a doubly-robust style bias correction in the second-stage regression and their construction is agnostic to the user's choice of nonparametric regression method through its use of sample splitting. By extending the analysis of pseudo-outcome regressions in Kennedy (2022*b*), we derive the integrated mean square error convergence rate of our DR-Learners to the true bounds.

We first develop our estimators for the case in which the bounding functions $\underline{\delta}(\cdot), \overline{\delta}(\cdot)$ are known. We then extend to the case in which the bounding functions themselves must be estimated using nonparametric outcome regression bounds and instrumental variable bounds.

3.1 DR-Learners for MOSM bounds on the target regression

To construct our proposed estimators for the bounds $[\underline{\mu}^*(x;\Delta), \overline{\mu}^*(x;\Delta)]$, we make use of sample-splitting. We illustrate our procedure by a single split procedure to simplify notation, but the analysis for multiple splits is straightforward.

We randomly split the data $\mathcal{O}$ into two disjoint subsets $\mathcal{O}_1$ and $\mathcal{O}_2$. We construct an estimator of the outcome regression $\hat{\mu}_1$ and propensity score $\hat{\pi}_1$ using only the observations $\mathcal{O}_1$. Using the observations $\mathcal{O}_2$, we construct the pseudo-outcomes

$$\phi_1(Y_i;\hat{\eta}) + \underline{\delta}(X_i)(1-D_i) \text{ and } \phi_1(Y_i;\hat{\eta}) + \overline{\delta}(X_i)(1-D_i), \tag{9}$$

where $\phi_1(Y_i;\eta) := \mu_1(X_i) + \frac{D_i}{\pi_1(X_i)}(Y_i - \mu_1(X_i))$ is the efficient uncentered influence function for $\mathbb{E}\{\mathbb{E}[Y_i \mid D_i = 1, X_i]\}$, $\eta = (\pi_1(X_i), \mu_1(X_i))$ are the relevant nuisance functions. We regress these constructed pseudo-outcomes on the features X_i using a user-specified nonparametric regression procedure in fold $\mathcal{O}_2$. This yields the DR-Learners $\widehat{\underline{\mu}}(x;\Delta)$, $\widehat{\overline{\mu}}(x;\Delta)$ of the target regression bounds under the MOSM. Algorithm 1 summarizes the construction of the DR-Learners.

Algorithm 1: Pseudo-algorithm for DR-Learners of MOSM target regression bounds.

Input: Data $\mathcal{O} = \{(O_i)\}_{i=1}^n$ where $O_i = (X_i, D_i, Y_i)$, number of folds K; $x \in \mathcal{X}$.

Split $\mathcal{O}$ into two independent folds $\mathcal{O}_1, \mathcal{O}_2$.

Estimate $\hat{\mu}_1$, $\hat{\pi}_1$ using only $\mathcal{O}_1$, and define $\hat{\eta} = (\hat{\pi}_1, \hat{\mu}_1)$.

Regress $\phi_1(Y_i;\hat{\eta}) + \underline{\delta}(X_i)(1-D_i) \sim X_i$ using $i \in \mathcal{O}_2$ to yield $\widehat{\underline{\mu}}(x;\Delta)$.

Regress $\phi_1(Y_i;\hat{\eta}) + \overline{\delta}(X_i)(1-D_i) \sim X_i$ using $i \in \mathcal{O}_2$ to yield $\widehat{\overline{\mu}}(x;\Delta)$.

Output: Estimated bounds $[\widehat{\underline{\mu}}(x;\Delta), \widehat{\overline{\mu}}(x;\Delta)]$.

3.2 Convergence rate of DR-Learners

We provide a theoretical guarantee on the integrated mean square error (MSE) convergence rate of the DR-Learners $\widehat{\underline{\mu}}(x;\Delta), \widehat{\overline{\mu}}(x;\Delta)$ to the true bounds. Our result compares the integrated MSE of the DR-Learners against that of an oracle nonparametric regression that has access to the true nuisance functions and can therefore form the true influence functions.

Consider an oracle that observes the true nuisance functions $\eta(X_i)$ and bounding functions $\underline{\delta}(X_i), \overline{\delta}(X_i)$ for each observation in the data. This infeasible oracle estimates the target regression bounds at any $x \in \mathcal{X}$ by regressing the true pseudo-outcomes $\phi_1(Y_i;\eta) + \underline{\delta}(X_i)(1-D_i)$ and $\phi_1(Y_i;\eta) + \overline{\delta}(X_i)(1-D_i)$ on the features X_i in the second fold using the same second-stage nonparametric regression procedure as the user. Let $\underline{\widehat{\mu}}_{oracle}(x;\Delta), \widehat{\overline{\mu}}_{oracle}(x;\Delta)$ denote these oracle estimators. Under an $L_2(\mathbb{P})$-stability condition on the user's second-stage nonparametric regression estimator, the integrated MSE of the DR-Learner equals the integrated MSE of the oracle regression plus a smoothed, doubly robust remainder term.

Theorem 3.1. *Let $\widehat{\mathbb{E}}_n[\cdot \mid X_i = x]$ denote the user-specified, second-stage pseudo-outcome regression estimator. Suppose that $\widehat{\mathbb{E}}_n[\cdot \mid X_i = x]$ satisfies the $L_2(\mathbb{P})$-stability condition (Assumption B.1), and $\mathbb{P}(\epsilon \leq \hat{\pi}_1(X_i) \leq 1-\epsilon) = 1$ for some $\epsilon > 0$. Define $\tilde{R}(x) = \widehat{\mathbb{E}}_n[(\pi_1(X_i) - \hat{\pi}_1(X_i))(\mu_1(X_i) - \hat{\mu}_1(X_i)) \mid X_i = x]$, and $R^2_{oracle} = \mathbb{E}[\|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\|^2]$. Then,*

$$\|\widehat{\overline{\mu}}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\| \leq \|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\| + \epsilon^{-1}\|\tilde{R}(\cdot)\| + o_{\mathbb{P}}(R_{oracle})$$

and $\widehat{\overline{\mu}}(\cdot;\Delta)$ is therefore oracle efficient in the $L_2(\mathbb{P})$-norm if further $\|\tilde{R}(\cdot)\| = o_{\mathbb{P}}(R_{oracle})$. The analogous result holds for the estimator of the lower-bound, $\underline{\widehat{\mu}}^*(x;\Delta)$.

Theorem 3.1 establishes that the integrated MSE of the DR-Learners for the target outcome regression bounds can be no larger than that of an infeasible oracle nonparametric regression (i.e., an oracle that has access to the true nuisance functions and bounding functions) plus the $L_2(\mathbb{P})$-norm of a smoothed remainder term $\tilde{R}(x)$ that depends on the product of errors in the estimation of the first-stage nuisance functions. The first-step, nonparametric estimation of the nuisance functions in the DR-Learners therefore only its error through this remainder term. Key to this bound is that the DR-Learners (i) use sample-splitting, estimating the nuisance parameters $\mu_1(X_i)$ and $\pi_1(X_i)$ on a separate fold of the data than the fold used for the pseudo-outcome regression, and (ii) construct pseudo-outcomes based on efficient influence functions. To prove this result, we prove an oracle inequality on the $L_2(\mathbb{P})$-error of regression with estimated pseudo-outcomes (Lemma B.1), extending Kennedy (2022*b*)'s analysis of the pointwise convergence of pseudo-outcome regression procedures.

The $L_2(\mathbb{P})$-stability condition (Assumption B.1) on the second-stage regression estimators is quite mild in practice, and Proposition B.1 shows that it is satisfied by a variety of generic linear smoothers such as linear regression, series regression, nearest neighbor matching, random forest models, and several others. The bounds in Theorem 3.1 are therefore agnostic to the underlying nonparametric regression method chosen by the user. As a result, the result can be easily applied in settings where the nuisance functions $\eta(\cdot)$ and bounding functions $\underline{\delta}(\cdot), \overline{\delta}(\cdot)$ themselves satisfy additional smoothness or sparsity conditions, and for particular choices of the second-stage regression estimator $\widehat{\mathbb{E}}_n[\cdot \mid X_i = x]$ and nuisance function estimators. Known results on the mean-squared error convergence rates of nonparametric regression procedures can then be applied to analyze $\|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\|$ and $\|\tilde{R}(\cdot)\|$. Applying Theorem 3.1 would then provide the convergence rate of our proposed DR-Learners as an explicit function of the sample size and dimensionality of the features.

3.3 Incorporating estimated bounding functions under MOSM

We now show how to extend our proposed DR-Learners when the user must also estimate the bounding functions $\underline{\delta}(\cdot)$, $\overline{\delta}(\cdot)$. The main conclusions of Theorem 3.1 continue to hold with additional remainder terms that arise from the nonparametric estimation of the bounding functions.

Nonparametric outcome bounds: Suppose the user specifies nonparametric outcome regression bounds under the MOSM (6) for some $\underline{\Gamma}, \overline{\Gamma} > 0$. In this case, the worst-case bounds on the target regression can be directly written as $\overline{\mu}(x; \Delta(\Gamma)) = \mu_1(x) + (\overline{\Gamma} - 1)\pi_0(x)\mu_1(x)$ and $\underline{\mu}^*(x) = \mu_1(x) + (\underline{\Gamma} - 1)\pi_0(x)\mu_1(x)$. We therefore modify our DR-Learners by simply modifying the pseudo-outcomes. Using the observations in $\mathcal{O}_2$, we now construct the two pseudo-outcomes $\phi_1(Y_i; \hat{\eta}) + (\overline{\Gamma} - 1)\phi(\pi_0(X_i)\mu_1(X_i); \hat{\eta})$ and $\phi_1(Y_i; \hat{\eta}) + (\underline{\Gamma} - 1)\phi(\pi_0(X_i)\mu_1(X_i); \hat{\eta})$, where

$$\phi(\pi_0(X_i)\mu_1(X_i); \eta) = ((1 - D_i) - \pi_0(X_i))\,\mu_1(X_i) + \frac{D_i}{\pi_1(X_i)}(Y_i - \mu_1(X_i))\pi_0(X_i) + \pi_0(X_i)\mu_1(X_i) \quad (10)$$

is the efficient influence function for $\mathbb{E}[\pi_0(X_i)\mu_1(X_i)]$ by standard influence function calculations (Kennedy, 2022*a*; Hines et al., 2022). Regressing these constructed pseudo-outcomes on the features X_i yields our DR-Learners $\widehat{\overline{\mu}}(x; \Delta(\Gamma))$, $\widehat{\underline{\mu}}(x; \Delta(\Gamma))$.

Under the same conditions as Theorem 3.1, the integrated MSE of these DR-Learners is equal to the integrated MSE of the oracle estimators plus the same second-order remainder term.

Proposition 3.1. *Under the same conditions as Theorem 3.1,*

$$\|\widehat{\overline{\mu}}(\cdot; \Delta(\Gamma)) - \overline{\mu}^*(\cdot; \Delta(\Gamma))\| \leq \|\widehat{\overline{\mu}}_{oracle}(\cdot; \Delta(\Gamma)) - \overline{\mu}^*(\cdot; \Delta(\Gamma))\| + \epsilon^{-1}\sqrt{\overline{\Gamma} - 1}\|\tilde{R}(\cdot)\| + o_{\mathbb{P}}(R_{oracle})$$

The analogous result holds for $\widehat{\underline{\mu}}(x; \Delta(\Gamma))$.

Instrumental variable bounds: Suppose the user specifies instrumental variable bounds under the MOSM. To derive the worst-case bounds on the target regression, it is convenient to rewrite Proposition 2.1 as bounds on the product $\pi_0(x)\delta(x)$

$$\underline{\delta}_z(x) \leq \pi_0(x)\delta(x) \leq \overline{\delta}_z(x)$$

for $\underline{\delta}_z(x) = \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] - \mu_1(x)$ and $\overline{\delta}_z(x) = \pi_0(x, z) + \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] - \mu_1(x)$. It then immediately follows the target regression bounds can be directly written as $\overline{\mu}^*(x; \Delta(z)) = \pi_0(x, z) + \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z]$ and $\underline{\mu}^*(x; \Delta(z)) = \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z]$. This suggests that we can modify our DR-Learners by modifying the nuisance functions that are estimated on each fold of the data and the pseudo-outcomes that are constructed.

We now construct an estimator of the regression $\widehat{\mathbb{E}}[Y_i D_i \mid X_i = x, Z_i = z]$, treatment propensity score $\hat{\pi}_0(x, z)$, and instrument propensity score $\widehat{\mathbb{P}}[Z_i = z \mid X_i = x]$ using only the observations $\mathcal{O}_1$. Using the observations $\mathcal{O}_2$, we construct the two pseudo-outcomes $\phi_z(Y_i D_i; \hat{\eta}) + \phi_z(1 - D_i; \hat{\eta})$ and $\phi_z(Y_i D_i; \hat{\eta})$, where

$$\phi_z(D_i Y_i; \eta) := \frac{1\{Z_i = z\}}{\mathbb{P}(Z_i = z \mid X_i = x)}(Y_i D_i - \mathbb{E}[D_i Y_i \mid X_i = x, Z_i = z]) + \mathbb{E}[D_i Y_i \mid X_i = x, Z_i = z] \quad (11)$$

$$\phi_z(1 - D_i; \eta) := \frac{1\{Z_i = z\}}{\mathbb{P}(Z_i = z \mid X_i = x)}\left((1 - D_i) - \pi_0(X_i, z)\right) + \pi_0(X_i, z) \quad (12)$$

are the efficient influence functions for $\mathbb{E}\{\mathbb{E}[D_i Y_i \mid X_i, Z_i = z]\}$, $\mathbb{E}\{\mathbb{E}[1 - D_i \mid X_i, Z_i = z]\}$ respectively, where $\eta = (\mathbb{P}(Z_i = z \mid X_i = x), \mathbb{E}[D_i Y_i \mid X_i = x, Z_i = z], \pi_0(x, z))$ are the relevant nuisance functions

(Kennedy, Balakrishnan and G'Sell, 2020). We then regress these constructed pseudo-outcomes on the features X_i, yielding our DR-Learners $\widehat{\underline{\mu}}(x; \Delta(z))$, $\widehat{\overline{\mu}}(x; \Delta(z))$.

Proposition 3.2. *Suppose the second-stage pseudo-outcome regression estimators $\widehat{\mathbb{E}}_n[\cdot \mid X_i = x]$ satisfy the $L_2(\mathbb{P})$-stability condition (Assumption B.1) and $\mathbb{P}(\epsilon \leq \widehat{\mathbb{P}}(Z_i = z \mid X_i = x)) = 1$ for some $\epsilon > 0$. Define $\tilde{R}_1(x) = \widehat{\mathbb{E}}_n[(\mathbb{P}(Z_i = z \mid X_i = x) - \widehat{\mathbb{P}}(Z_i = z \mid X_i = x))(\pi_0(x,z) - \hat{\pi}_0(x,z)) \mid X_i = x]$, $\tilde{R}_2(x) = \widehat{\mathbb{E}}_n[(\mathbb{P}(Z_i = z \mid X_i = x) - \widehat{\mathbb{P}}(Z_i = z \mid X_i = x))(\mathbb{E}[D_i Y_i \mid Z_i = z, X_i = x] - \widehat{\mathbb{E}}[D_i Y_i \mid Z_i = z, X_i = x]) \mid X_i = x]$, and $R^2_{oracle}(z) = \mathbb{E}[\|\widehat{\overline{\mu}}(\cdot; \Delta(z)) - \overline{\mu}^*(\cdot; \Delta(z))\|^2]$. Then,*

$$\|\widehat{\overline{\mu}}(\cdot; \Delta(z)) - \overline{\mu}(\cdot; \Delta(z))\| \leq \|\widehat{\overline{\mu}}_{oracle}(\cdot; \Delta(z)) - \overline{\mu}(\cdot; \Delta(z))\| + \epsilon^{-1}\left(\|\tilde{R}_1(\cdot)\| + \|\tilde{R}_2(\cdot)\|\right) + o_{\mathbb{P}}(R_{oracle}(z)).$$

The analogous result holds for the estimator of the lower bound $\widehat{\underline{\mu}}(x; \Delta(z))$.

4 Robust recommendations under the mean outcome sensitivity model

While in some settings a risk assessment alone is sufficient, in many others decision makers must translate the risk assessment into an intervention. We show how our proposed DR-Learners for the bounds on the target regression under the MOSM can be translated into a plug-in decision-making policy that has desirable robustness properties. We bound the worst-case performance of our plug-in decision-making policy relative to the max-min optimal decision rule. Our results for the DR-Learner (Theorem 3.1) then imply conditions under which our estimated decision-making policy is asymptotically max-min optimal.

4.1 Expected counterfactual utility and optimal max-min decision policies

We consider a setting in which a decision maker selects a deterministic personalized decision-making policy $d(\cdot)\colon \mathcal{X} \to \{0,1\}$ mapping features into recommendations for whether the intervention $D_i = 1$ should be implemented. We assume the decision maker prefers to provide the intervention $D_i = 1$ only when $Y_i(1) = 0$, and so they evaluate $d(X_i)$ by its *expected counterfactual utility*

$$U(d) := \mathbb{E}[(-u_{1,1}(X_i)Y_i(1) + u_{1,0}(X_i)(1-Y_i(1)))d(X_i) + (-u_{0,0}(X_i)(1-Y_i(1)) + u_{0,1}(X_i)Y_i(1))(1-d(X_i))],$$

where the utility functions $u_{1,1}(\cdot), u_{1,0}(\cdot), u_{0,0}(\cdot), u_{0,1}(\cdot) \geq 0$ specify the known payoff associated with each possible combination of decision D_i and counterfactual outcome $Y_i(1)$ at features X_i. We assume the utility functions satisfy the normalization $\sum_{d,y \in \{0,1\}^2} u_{d,y}(x) = 1$ with probability one.

This objective function is quite general, and arises naturally in our earlier running examples. In the consumer lending example, the profitability of approving customers that would not default $u_{1,0}(\cdot)$ may vary based on observed features such as the requested loan size. Analogously, in pretrial release, the benefits of releasing defendants that would not fail to appear in court $u_{0,1}(X_i)$ may vary based on observable features such as the defendant's age, charge severity, or prior history of pretrial misconduct.

While the expected counterfactual utility of a decision rule $d(X_i)$ is not point identified due to the missing data problem, it is nonetheless bounded under the MOSM. By iterated expectations, it equals

$$U(d) = \mathbb{E}[(-u_{1,1,i}\mu^*(X_i) + u_{1,0,i}(1-\mu^*(X_i)))d(X_i) + (-u_{0,0,i}(1-\mu^*(X_i)) + u_{0,1,i}\mu^*(X_i))(1-d(X_i))], \quad (13)$$

where $u_{d,y,i} := u_{d,y}(X_i)$ for $(d,y) \in \{0,1\}^2$. For any $d(X_i)$, we can directly characterize the sharp set of expected counterfactual utilities $\mathcal{H}(U(d);\Delta)$ consistent with the MOSM.

Lemma 4.1. *Suppose Assumption 2.1 is satisfied. Then, for any decision rule $d(\cdot)\colon \mathcal{X} \to \{0,1\}$, $\mathcal{H}(U(d);\Delta) = [\underline{U}(d;\Delta), \overline{U}(d;\Delta)]$, where*

$$\underline{U}(d;\Delta) := \mathbb{E}[(u_{1,0,i} - (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(X_i;\Delta))\, d(X_i) + \big(-u_{0,0,i} + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(X_i;\Delta)\big)\,(1 - d(X_i))]$$
$$\overline{U}(d;\Delta) := \mathbb{E}[\big(u_{1,0,i} - (u_{1,1,i} + u_{1,0,i})\underline{\mu}^*(X_i;\Delta)\big)\, d(X_i) + (-u_{0,0,i} + (u_{0,0,i} + u_{0,1,i})\overline{\mu}^*(X_i;\Delta))\,(1 - d(X_i))].$$

Since the expected counterfactual utility of any decision policy can only be sharply bounded under the MOSM, the decision maker must address this inherent ambiguity (e.g., Manski, 2007). We assume the decision maker selects a decision policy to solve

$$d^*(\cdot;\Delta) \in \arg\max_{d(\cdot)\colon \mathcal{X}\to[0,1]} \underline{U}(d;\Delta). \tag{14}$$

The decision maker therefore compares decision policies based on their worst-case performance. In defining the optimal max-min decision rule $d^*(\cdot;\Delta)$, we place no restrictions on the class of decision rules.[9,10] Given the structure of the sharp lower bound on expected counterfactual utility, the optimal, max-min decision rule thresholds a weighted average of the target regression bounds under the MOSM.

Lemma 4.2. *Define $\widetilde{\mu}^*(x;\Delta) = (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(x;\Delta) + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(x;\Delta)$ the utility weighted-average of the target regression bounds. The optimal max-min decision rule is*

$$d^*(X_i;\Delta) = 1\{\widetilde{\mu}^*(X_i;\Delta) \le u_{1,0,i} + u_{0,0,i}\}.$$

Since the target regression bounds $[\underline{\mu}^*(X_i), \overline{\mu}^*(X_i)]$ are not known exactly, the optimal max-min decision rule is infeasible. We therefore next consider the performance of a plug-in version based on our DR-Learners for the target regression bounds.

4.2 Regret bounds for the plug-in max-min decision policy

Using our DR-Learners for the target regression bounds, we consider a plug-in version of the optimal max-min decision rule under the MOSM. Define $\widehat{\widetilde{\mu}}(x;\Delta) = (u_{1,1,i}+u_{1,0,i})\widehat{\overline{\mu}}(x;\Delta)+(u_{0,0,i}+u_{0,1,i})\widehat{\underline{\mu}}(x;\Delta)$ to be the estimator of $\widetilde{\mu}^*(x;\Delta)$ at $x \in \mathcal{X}$ that plugs in our DR-Learners for the target regression bounds. The plug-in decision rule is

$$\hat{d}(x;\Delta) = 1\{\widehat{\widetilde{\mu}}(x;\Delta) \le u_{1,0,i} + u_{0,0,i}\}. \tag{15}$$

How much worse does the decision-maker do if she makes decisions under this plug-in decision rule rather than the optimal max-min decision rule? To do so, we define regret of the feasible, plug-in

[9] This contrasts with recent work on statistical treatment assignment rules such as Kitagawa and Tetenov (2018); Athey and Wager (2021); Kallus and Zhou (2021), which typically compares estimated decisions rules against the best decision rule in some restricted policy class.

[10] In related work, Ben-Michael, Imai and Jiang (2022) study optimal policy learning in a setting where utility depends on both the decision D_i and the full potential outcome vector $(Y_i(0), Y_i(1))$ (i.e., is "assymmetric") and the decision is unconfounded in the historical data. Ben-Michael et al. (2021) and Zhang, Ben-Michael and Imai (2022) also study optimal policy learning in settings where treatment effects can only be bounded since, respectively, the historical decision-making policy may be deterministic or there exists a historical regression discontinuity in decisions.

decision rule relative to the max-min optimal decision rule as

$$R(\hat{d};\Delta) = \underline{U}(d^*;\Delta) - \underline{U}(\hat{d};\Delta). \tag{16}$$

Notice that $R(\hat{d};\Delta) \geq 0$, and it measures the difference between the worst-case expected counterfactual utility of the max-min optimal decision rule against what is attained by the feasible plug-in decision rule. Our next result derives bounds on the squared regret of the plug-in decision rule.

Theorem 4.1. *Under the same conditions as Theorem 3.1, for* $\tilde{R}(x) = \hat{\mathbb{E}}_n[(\pi_1(X_i) - \hat{\pi}_1(X_i))(\mu_1(X_i) - \hat{\mu}_1(X_i)) \mid X_i = x]$,

$$R(\hat{d};\Delta)^2 \leq 2\|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\| + 2\|\widehat{\underline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\| + 4\epsilon^{-1}\|\tilde{R}(x)\| + o_{\mathbb{P}}(R_{oracle})$$

Theorem 4.1 shows that the squared regret of the plug-in decision rule is bounded above by the oracle integrated MSE for the target regression bounds plus again a smoothed, doubly robust remainder term that depends on the smoothed product of errors in the estimation of the first-stage nuisance parameters. There are a few points to emphasize about this bound. First, it compares the lower bound on expected counterfactual utility of the feasible decision rule against the unrestricted, max-min optimal decision rule. Second, this result immediately implies that the worst-case regret of the plug-in decision rule will converge to zero whenever the integrated MSE of the oracle converges to zero (and therefore so does the integrated MSE of the DR-Learners). In such a case, the plug-in decision rule would be asymptotically max-min optimal.

5 Robust audits under the mean outcome sensitivity model

Finally, in this section, we robustly audit the performance of an existing risk assessment $s(X_i)$ by constructing estimators of its worst-case predictive performance $\overline{\text{perf}}(s;\beta,\Delta)$, $\overline{\text{perf}}_+(s;\beta,\Delta)$ under the MOSM. Our proposed estimators are based on efficient influence functions and cross-fitting, which will enable us to control bias from the nonparametric estimation of nuisance functions and allow the use of complex machine learning estimator for these nuisance functions (e.g., Robins et al., 2008; Zheng and van der Laan, 2011; Chernozhukov et al., 2018). As for the DR-Learners, we first develop our estimators for the case in which the bounding functions $\underline{\delta}(\cdot)$, $\overline{\delta}(\cdot)$ are known, and then extend to the case in which the bounding functions themselves must be estimated.

5.1 Estimating bounds on overall predictive performance

We first estimate the bounds on overall predictive performance $\underline{\text{perf}}(s;\beta,\Delta)$, $\overline{\text{perf}}(s;\beta,\Delta)$ of a risk assessment $s(X_i)$ under the MOSM. As in the construction of the DR-Learners, $\phi_1(Y_i;\eta)$ denotes the efficient influence function for $\mathbb{E}\{\mathbb{E}[Y_i \mid D_i = 1, X_i]\}$ and $\eta := (\pi_1(X_i), \mu_1(X_i))$.

We randomly split the historical data into K folds, letting $\mathcal{O}_k$ denote the observations in the k-th fold and $\mathcal{O}_{-k}$ denote the observations not in the k-th fold. For each fold k, we construct estimators of the nuisance functions $\hat{\eta}_{-k}$ using only the sample of observations $\mathcal{O}_{-k}$ not in the k-th fold. For each observation in the k-th fold $\mathcal{O}_k$, we construct

$$\overline{\text{perf}}(O_i; \hat{\eta}_{-k}) := \beta_{0,i} + \beta_{1,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{1,i}(1 - D_i)(1\{\beta_{1,i} > 0\}\overline{\delta}_i + 1\{\beta_{1,i} \leq 0\}\underline{\delta}_i), \tag{17}$$

$$\underline{\text{perf}}(O_i; \hat{\eta}_{-k}) := \beta_{0,i} + \beta_{1,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{1,i}(1 - D_i)(1\{\beta_{1,i} \leq 0\}\overline{\delta}_i + 1\{\beta_{1,i} > 0\}\underline{\delta}_i). \tag{18}$$

We then estimate the upper-bound on overall predictive performance under the MOSM by taking the average across all units in the historical data $\mathcal{O}$, or equivalently the weighted average of the corresponding fold-specific estimators

$$\widehat{\overline{\text{perf}}}(s; \beta, \Delta) := \mathbb{E}_n\left[\overline{\text{perf}}(O_i; \hat{\eta}_{-K_i})\right] = \sum_{k=1}^{K}\left(n^{-1}\sum_{i=1}^{n} 1\{K_i = k\}\right)\mathbb{E}_n^k\left[\overline{\text{perf}}(O_i; \hat{\eta}_{-k})\right], \tag{19}$$

where $\mathbb{E}_n^k[\cdot]$ denotes the sample average over the k-th fold $\mathcal{O}_k$. Our estimator for the lower-bound $\widehat{\underline{\text{perf}}}(s; \beta, \Delta) := \mathbb{E}_n\left[\underline{\text{perf}}(O_i; \eta_{-K_i})\right]$ is defined analogously. Algorithm 2 summarizes our proposed estimators for the overall predictive performance bounds under the MOSM and their associated standard errors.

Algorithm 2: Pseudo-algorithm for overall predictive performance bounds estimators.

Input: Data $\mathcal{O} = \{(O_i)\}_{i=1}^n$ where $O_i = (X_i, D_i, Y_i)$, number of folds K.

for $k = 1, \ldots, K$ **do**
- Estimate $\hat{\eta}_{-k} = (\hat{\pi}_{1,-k}, \hat{\mu}_{1,-k})$.
- Set $\overline{\text{perf}}(O_i; \hat{\eta}_{-K(i)})$ and $\underline{\text{perf}}(O_i; \hat{\eta}_{-K(i)})$ for all $i \in \mathcal{O}_k$.

Set $\widehat{\overline{\text{perf}}}(s; \beta, \Delta) = \mathbb{E}_n[\overline{\text{perf}}(O_i; \hat{\eta}_{-K(i)})]$, $\widehat{\underline{\text{perf}}}(s; \beta, \Delta) = \mathbb{E}_n[\underline{\text{perf}}(O_i; \hat{\eta}_{-K(i)})]$;

Set $\hat{\sigma}_{i,11} = (\overline{\text{perf}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\overline{\text{perf}}}(s; \beta, \Delta))^2$,
$\hat{\sigma}_{i,12} = (\overline{\text{perf}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\overline{\text{perf}}}(s; \beta, \Delta))(\underline{\text{perf}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\underline{\text{perf}}}(s; \beta, \Delta))$, and
$\hat{\sigma}_{i,22} = (\underline{\text{perf}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\underline{\text{perf}}}(s; \beta, \Delta))^2$;

Output: Estimates $\widehat{\overline{\text{perf}}}(s; \beta, \Delta) = \mathbb{E}_n[\overline{\text{perf}}(O_i; \hat{\eta}_{-K(i)})]$, $\widehat{\underline{\text{perf}}}(s; \beta, \Delta) = \mathbb{E}_n[\underline{\text{perf}}(O_i; \hat{\eta}_{-K(i)})]$.

Output: Estimated covariance matrix $n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\hat{\sigma}_{i,11} & \hat{\sigma}_{i,12}\\ \hat{\sigma}_{i,12} & \hat{\sigma}_{i,22}\end{pmatrix}$

Our next theorem derives the rate of convergence of our proposed estimators of the bounds, and provides conditions under which they are jointly asymptotically normal.

Theorem 5.1. *Define the remainder* $R_{1,n}^k := \|\hat{\mu}_{1,-k}(\cdot) - \mu_1(\cdot)\|\|\hat{\pi}_{1,-k}(\cdot) - \pi_1(\cdot)\|$ *for each fold* $k = 1, \ldots, K$*. Assume (i)* $\mathbb{P}(\pi_1(X_i) \geq \delta) = 1$ *for some* $\delta > 0$*, (ii) there exists* $\epsilon > 0$ *such that* $\mathbb{P}(\hat{\pi}_{1,-k}(X_i) \geq \epsilon) = 1$ *for each fold* k*, and (iii)* $\|\hat{\mu}_{1,-k}(\cdot) - \mu_1(\cdot)\| = o_P(1)$ *and* $\|\hat{\pi}_{1,-k}(\cdot) - \pi_1(\cdot)\| = o_P(1)$ *for each fold* k*. Then,*

$$\left|\widehat{\overline{perf}}(s; \beta, \Delta) - \overline{perf}(s; \beta, \Delta)\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R_{1,n}^k\right),$$

$$\left|\widehat{\underline{perf}}(s; \beta, \Delta) - \underline{perf}(s; \beta, \Delta)\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R_{1,n}^k\right).$$

If further $R^k_{1,n} = o_{\mathbb{P}}(1/\sqrt{n})$ for each fold k, then

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{perf}}(s;\beta,\Delta)\\ \widehat{\underline{perf}}(s;\beta,\Delta)\end{pmatrix} - \begin{pmatrix}\overline{perf}(s;\beta,\Delta)\\ \underline{perf}(s;\beta,\Delta)\end{pmatrix}\right) \xrightarrow{d} N\left(0,\Sigma\right)$$

for covariance matrix Σ defined in the proof.

Theorem 5.1 establishes that the errors associated with our estimators of the bounds on overall predictive performance under the MOSM consists of fold-specific doubly robust remainders $R^k_{1,n}$ that will be small if either the propensity score or the outcome regression are estimated well. Furthermore, the rate condition required for our proposed estimators of the bounds to be asymptotically normal will be satisfied if all nonparametric estimators of the nuisance parameters converge at a rate faster than $O_{\mathbb{P}}(n^{-1/4})$, which is the familiar condition required on first-stage nuisance parameter estimators from the double/debiased machine learning (e.g., Robins et al., 2008; Zheng and van der Laan, 2011; Chernozhukov et al., 2018). The user can therefore use a wide-suite of nonparametric regression methods or modern machine learning based methods to construct the first-stage nuisance parameter estimators.

In Appendix C, we construct a consistent estimator of the asymptotic covariance matrix in Theorem 5.1. As a consequence, the user can conduct statistical inference by reporting asymptotically valid confidence intervals for either the upper bound or lower bound on overall predictive performance. The joint normality of our estimators of the bounds combined with the consistent estimator of the asymptotic covariance matrix also imply that researchers can construct confidence intervals for the sharp identified set $\mathcal{H}(\text{perf}(s;\beta,\Delta)$ using standard methods for inference under partial identification (e.g., Imbens and Manski, 2004; Stoye, 2009).

In Appendix C, we also develop estimators for the bounds on the overall predictive disparities of the risk assessment $s(X_i)$ under the MOSM. We again show that these estimators converge at a fast rate, and are asymptotically normally distributed.

5.1.1 Incorporating estimated bounding functions under MOSM

We now extend our proposed estimators of the bounds on overall predictive performance when the user must also estimate the bounding functions $\underline{\delta}(\cdot)$, $\overline{\delta}(\cdot)$. Provided that the nonparametric estimators for the appropriate nuisance functions converge at a sufficiently fast rate, the main conclusions of Theorem 5.1 continue to hold.

Nonparametric outcome bounds: Suppose the user specifies nonparametric outcome regression bounds under the MOSM (6) for some $\underline{\Gamma}, \overline{\Gamma} > 0$. In this case, the worst-case bounds on overall predictive performance can be directly written as

$$\overline{\text{perf}}(s;\beta,\Delta(\Gamma)) = \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}\left(1\{\beta_{1,i} > 0\}(\overline{\Gamma}-1) + 1\{\beta_{1,i} \leq 0\}(\underline{\Gamma}-1)\right)\pi_0(X_i)\mu_1(X_i)]$$
$$\underline{\text{perf}}(s;\beta,\Delta(\Gamma)) = \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}\left(1\{\beta_{1,i} \leq 0\}(\overline{\Gamma}-1) + 1\{\beta_{1,i} > 0\}(\underline{\Gamma}-1)\right)\pi_0(X_i)\mu_1(X_i)].$$

We therefore can directly extend our proposed estimators by simply instead constructing

$$\overline{\text{perf}}(O_i;\hat{\eta}_{-k}) := \beta_{0,i} + \beta_{1,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \beta_{1,i}\left(1\{\beta_{1,i} > 0\}(\overline{\Gamma}-1) + 1\{\beta_{1,i} \le 0\}(\underline{\Gamma}-1)\right)\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k}),$$
$$\underline{\text{perf}}(O_i;\hat{\eta}_{-k}) := \beta_{0,i} + \beta_{1,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \beta_{1,i}\left(1\{\beta_{1,i} \le 0\}(\overline{\Gamma}-1) + 1\{\beta_{1,i} > 0\}(\underline{\Gamma}-1)\right)\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k})$$

for each observation in the k-th fold $\mathcal{O}_k$, where $\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta})$ is the uncentered efficient influence function for $\mathbb{E}\{\pi_0(X_i)\mu_1(X_i)\}$ as defined in (10). Our estimators for the worst-case bounds on overall predictive performance under nonparametric outcome regression bounds are then $\widehat{\overline{\text{perf}}}(s;\beta,\Delta(\Gamma)) := \mathbb{E}_n[\overline{\text{perf}}(O_i;\hat{\eta}_{-K_i})]$ and $\widehat{\underline{\text{perf}}}(s;\beta,\Delta(\Gamma)) := \mathbb{E}_n[\underline{\text{perf}}(O_i;\hat{\eta}_{-K_i})]$. Under the same conditions as Theorem 5.1, our estimators for the worst-case bounds on overall predictive performance under nonparametric outcome regression bounds continue to converge quickly to the true bounds.

Proposition 5.1. *Suppose the user specifies outcome regression bounds for some $\underline{\Gamma}, \overline{\Gamma} > 0$. Under the same conditions as Theorem 5.1,*

$$\left|\widehat{\overline{perf}}(s;\beta,\Delta(\Gamma)) - \overline{perf}(s;\beta,\Delta(\Gamma))\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R_{1,n}^k\right),$$
$$\left|\widehat{\underline{perf}}(s;\beta,\Delta(\Gamma)) - \underline{perf}(s;\beta,\Delta(\Gamma))\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R_{1,n}^k\right).$$

If further $R_{1,n}^k = o_{\mathbb{P}}(1/\sqrt{n})$ for all folds k, then

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{perf}}(s;\beta,\Delta(\Gamma))\\ \widehat{\underline{perf}}(s;\beta,\Delta(\Gamma))\end{pmatrix} - \begin{pmatrix}\overline{perf}(s;\beta,\Delta(\Gamma))\\ \underline{perf}(s;\beta,\Delta(\Gamma))\end{pmatrix}\right) \xrightarrow{d} N\left(0,\Sigma(\Gamma)\right)$$

for covariance matrix $\Sigma(\Gamma)$ defined in the proof.

Instrumental variable bounds: Suppose the user specifies instrumental variable bounds under the MOSM. To derive the worst-case bounds on overall predictive performance, it is again convenient to first rewrite the IV bounds in Proposition 2.1 as bounds on the product $\pi_0(x)\delta(x)$ as $\underline{\delta}_z(x) \le \pi_0(x)\delta(x) \le \overline{\delta}_z(x)$ for $\underline{\delta}_z(x)$ and $\overline{\delta}_z(x)$ defined earlier in Section 3.3. The worst-bounds on overall predictive performance are therefore

$$\overline{\text{perf}}(s;\beta,\Delta(z)) = \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}(1\{\beta_{1,i} > 0\}\overline{\delta}_z(X_i) + 1\{\beta_{1,i} \le 0\}\underline{\delta}_z(X_i))],$$
$$\underline{\text{perf}}(s;\beta,\Delta(z)) = \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}(1\{\beta_{1,i} > 0\}\underline{\delta}_z(X_i) + 1\{\beta_{1,i} \le 0\}\overline{\delta}_z(X_i))].$$

Based on this expression, we extend our proposed estimators by constructing

$$\overline{\text{perf}}(O_i;\hat{\eta}_{-k}) = \beta_{0,i} + \beta_{1,i}\phi(Y_i;\hat{\eta}_{-k}) + \beta_{1,i}(1\{\beta_{1,i} > 0\}\phi(\overline{\delta}_z(X_i);\hat{\eta}_{-k}) + 1\{\beta_{1,i} \le 0\}\phi(\underline{\delta}_z(X_i);\hat{\eta}_{-k}))$$
$$\underline{\text{perf}}(O_i;\hat{\eta}_{-k}) = \beta_{0,i} + \beta_{1,i}\phi(Y_i;\hat{\eta}_{-k}) + \beta_{1,i}(1\{\beta_{1,i} \le 0\}\phi(\overline{\delta}_z(X_i);\hat{\eta}_{-k}) + 1\{\beta_{1,i} > 0\}\phi(\underline{\delta}_z(X_i);\hat{\eta}_{-k}))$$

for each observation in the k-fold $\mathcal{O}_k$, where $\phi(\overline{\delta}_z(X_i);\eta)$, $\phi(\underline{\delta}_z(X_i);\eta)$ are the efficient influence functions for $\mathbb{E}[\overline{\delta}_z(X_i)]$, $\mathbb{E}[\underline{\delta}_z(X_i)]$ respectively, and η are the relevant nuisance functions. Recall that efficient influence functions for $\mathbb{E}[\overline{\delta}_z(X_i)]$, $\mathbb{E}[\underline{\delta}_z(X_i)]$ are

$$\phi(\overline{\delta}_z(X_i);\eta) = \phi_z(1-D_i;\eta) + \phi_z(D_iY_i;\eta) - \phi_1(Y_i;\eta)$$
$$\phi(\underline{\delta}_z(X_i);\eta) = \phi_z(D_iY_i;\eta) - \phi_1(Y_i;\eta),$$

where $\phi_z(1-D_i;\eta)$ and $\phi_z(D_iY_i;\eta)$ are the efficient influence functions for $\mathbb{E}\{\mathbb{E}[1-D_i \mid X_i, Z_i = z]\}$ and $\mathbb{E}\{\mathbb{E}[D_iY_i \mid X_i, Z_i = z]\}$ respectively defined in (12) and (11). We therefore define our estimators for the worst-case bounds on overall predictive performance under the instrumental variable bounds as $\widehat{\overline{\text{perf}}}(s;\beta,\Delta(z)) := \mathbb{E}_n[\overline{\text{perf}}(O_i;\hat{\eta}_{-K_i})]$ and $\widehat{\underline{\text{perf}}}(s;\beta,\Delta(z)) := \mathbb{E}_n[\underline{\text{perf}}(O_i;\hat{\eta}_{-K_i}]$.

We extend Theorem 5.1 to derive the rate of convergence of our proposed estimators under instrumental variable bounds.

Proposition 5.2. *Define $R^k_{1,n} = \|\hat{\mu}_{1,-k}(\cdot)-\mu_1(\cdot)\|\|\hat{\pi}_{1,-k}(\cdot)-\pi_1(\cdot)\|$ as before, and let $R^k_{2,n} = \|\hat{\mathbb{E}}_{-k}[Y_iD_i \mid X_i, Z_i = z] - \mathbb{E}[Y_iD_i \mid X_i, Z_i = z]\|\|\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) - \mathbb{P}(Z_i = z \mid X_i)\|$, $R^k_{3,n} = \|\hat{\pi}_{0,-k}(\cdot,z) - \pi_0(\cdot,z)\|\|\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) - \mathbb{P}(Z_i = z \mid X_i)\|$. Assume that (i) $\mathbb{P}\{\mathbb{P}(Z_i = z \mid X_i) \geq \delta\} = 1$ and $\mathbb{P}(\pi_1(X_i) \geq \delta) = 1$ for some $\delta > 0$; (ii) there exists $\epsilon > 0$ such that $\mathbb{P}\{\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) \geq \epsilon\} = 1$ and $\mathbb{P}(\hat{\pi}_{1-k}(X_i) \geq \epsilon) = 1$ for all folds k; and (iii) $\|\hat{\mathbb{E}}_{-k}[D_iY_i \mid X_i, Z_i = z] - \mathbb{E}[D_iY_i \mid X_i, Z_i = z]\| = o_{\mathbb{P}}(1)$, $\|\hat{\pi}_{0,-k}(\cdot;z) - \pi_0(\cdot;z)\| = o_{\mathbb{P}}(1)$, $\|\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) - \mathbb{P}(Z_i = z \mid X_i)\| = o_{\mathbb{P}}(1)$, $\|\hat{\mu}_{1,-k}(\cdot) - \mu_1(\cdot)\| = o_{\mathbb{P}}(1)$, and $\|\hat{\pi}_{1,-k}(\cdot) - \pi_1(\cdot)\| = o_{\mathbb{P}}(1)$ for all folds k. Then,*

$$\left|\widehat{\overline{perf}}(s;\beta,\Delta(z)) - \overline{perf}(s;\beta,\Delta(z))\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K}(R^k_{1,n} + R^k_{2,n} + R^k_{3,n})\right)$$
$$\left|\widehat{\underline{perf}}(s;\beta,\Delta(z)) - \underline{perf}(s;\beta,\Delta(z))\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K}(R^k_{1,n} + R^k_{2,n} + R^k_{3,n})\right).$$

If further $R^k_{1,n} = o_{\mathbb{P}}(1/\sqrt{n})$, $R^k_{2,n} = o_{\mathbb{P}}(1/\sqrt{n})$, and $R^k_{3,n} = o_{\mathbb{P}}(1/\sqrt{n})$ for all folds k, then

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{perf}}(s;\beta,\Delta(z))\\ \widehat{\underline{perf}}(s;\beta,\Delta(z))\end{pmatrix} - \begin{pmatrix}\overline{perf}(s;\beta,\Delta(z))\\ \underline{perf}(s;\beta,\Delta(z))\end{pmatrix}\right) \xrightarrow{d} N\left(0,\Sigma(z)\right)$$

for covariance matrix $\Sigma(z)$ defined in the proof.

5.2 Estimating bounds on positive-class predictive performance

We next consider the problem of estimating the bounds $\underline{\text{perf}}_+(s;\beta,\Delta)$ on positive-class predictive performance of a risk assessment $s(X_i)$ under the MOSM. Our estimators directly solve the empirical analogues of the population optimization problems that characterize the sharp bounds given in Lemma 2.1. We will again make use of K-fold cross-fitting. We now explicitly assume n is divisible by K, and each fold contains n/K observations for simplicity. For each fold $k = 1,\ldots,K$, we construct estimators of the nuisance functions $\hat{\eta}_{-k}$ using only the sample of observations $\mathcal{O}_{-k}$ not in the k-th fold. We then construct a fold-specific estimate of the upper bound by solving

$$\widehat{\overline{\text{perf}}}^k_+(s;\beta,\Delta_n) := \max_{\tilde{\delta}\in\Delta_n} \frac{\mathbb{E}^k_n[\beta_{0,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \beta_{0,i}(1-D_i)\tilde{\delta}_i]}{\mathbb{E}^k_n[\phi_1(Y_i;\hat{\eta}_{-k}) + (1-D_i)\tilde{\delta}_i]}, \tag{20}$$

where $\Delta_n = \left\{\tilde{\delta} \in \mathbb{R}^n : \underline{\delta}(X_i) \leq \tilde{\delta}_i \leq \overline{\delta}(X_i) \text{ for } i = 1, \ldots, n\right\}$. Our estimator then averages these fold-specific estimates

$$\widehat{\overline{\text{perf}}}_+(s; \beta, \Delta_n) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\overline{\text{perf}}}^k_+(s; \beta, \Delta_n). \tag{21}$$

Analogously, we construct fold-specific estimates of the lower-bound $\widehat{\underline{\text{perf}}}^k_+(s; \beta, \Delta)$ by solving the corresponding minimization problem, and estimate the lower-bound on positive-class predictive performance as $\widehat{\underline{\text{perf}}}_+(s; \beta, \Delta_n) = K^{-1} \sum_{k=1}^{K} \widehat{\underline{\text{perf}}}^k_+(s; \beta, \Delta)$. Algorithm 3 summarizes this procedure.

Algorithm 3: Pseudo-algorithm for positive-class predictive performance bounds estimators.

Input: Data $\mathcal{O} = \{(O_i)\}_{i=1}^n$ where $O_i = (X_i, D_i, Y_i)$, number of folds K.

for $k = 1, \ldots, K$ **do**
- Estimate $\hat{\eta}_{-k} = (\hat{\pi}_{1,-k}, \hat{\mu}_{1,-k})$.
- Set $\widehat{\overline{\text{perf}}}^k_+(s; \beta, \Delta_n)$ by solving (20).
- Set $\widehat{\underline{\text{perf}}}^k_+(s; \beta, \Delta_n)$ by solving the corresponding minimization.

Output: Estimates $\widehat{\overline{\text{perf}}}_+(s; \beta, \Delta_n) = K^{-1} \sum_{k=1}^{K} \widehat{\overline{\text{perf}}}^k_+(s; \beta, \Delta)$,
$\widehat{\underline{\text{perf}}}_+(s; \beta, \Delta) = K^{-1} \sum_{k=1}^{K} \widehat{\underline{\text{perf}}}^k_+(s; \beta, \Delta_n)$.

At first glance, $\widehat{\overline{\text{perf}}}^k_+(s; \beta, \Delta_n)$, $\widehat{\underline{\text{perf}}}^k_+(s; \beta, \Delta_n)$ may appear to be challenging optimization problems, but there is important structure to exploit. Since both are linear-fractional programs, they can be equivalently expressed as linear programs by applying the Charnes-Cooper transformation (Charnes and Cooper, 1962).

Lemma 5.1. *For any fold k, define $n^k = \sum_{i=1}^n 1\{K_i = k\}$, $\hat{c}^k = \mathbb{E}_n^k[(\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1 - D_i)\underline{\delta}_i)]$, $\hat{d}^k = \mathbb{E}_n^k[(\phi_1(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\underline{\delta}_i)]$. Also define $\hat{\alpha}_i = \beta_{0,i}(1 - D_i)(\overline{\delta}_i - \underline{\delta}_i)$, $\hat{\gamma}_i = (1 - D_i)(\overline{\delta}_i - \underline{\delta}_i)$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)$, $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_n)$. Then,*

$$\begin{aligned} \widehat{\overline{perf}}^k_+(s; \beta, \Delta_n) = \max_{\tilde{U} \in \mathbb{R}^{n^k}, \tilde{V} \in \mathbb{R}} \quad & \hat{\alpha}'\tilde{U} + \hat{c}^k \tilde{V} \\ \text{s.t.} \quad & 0 \leq \tilde{U}_i \leq \tilde{V} \text{ for } i = 1, \ldots n_k, \\ & 0 \leq \tilde{V}, \ \hat{\gamma}'\tilde{U} + \tilde{V}\hat{d}^k = 1. \end{aligned}$$

$\widehat{\underline{perf}}^k_+(s; \beta, \Delta_n)$ is optimal value of the corresponding minimization problem.

The calculation of our proposed estimators of the bounds on positive-class predictive performance simply requires the user to find the solution to K linear programs.

We can further exploit the optimization structure of these estimators in order to analyze their convergence rates to the true bounds under the MOSM. For exposition, we only derive the convergence rate for the our estimator of the upper bound $\widehat{\overline{\text{perf}}}_+(s; \beta, \Delta_n)$, but the analogous result applies to our estimator of the lower bound. As a first step, we show that optimization over the set of bounding functions Δ under the MOSM is equivalent to optimization over a special subclass of bounding functions both in the population and the sample optimization problems. This result exploits the linear-fractional structure of the optimizations, and such a reduction has been noted before in other sensitivity analysis models such as Aronow and Lee (2013); Kallus, Mao and Zhou (2018); Zhao, Small and Bhattacharya (2019); Kallus and Zhou (2021).

Lemma 5.2. *Define $\mathcal{U}$ to be the set of monotone, non-decreasing functions $u(\cdot)\colon \mathbb{R} \to [0,1]$, $\Delta^M := \left\{\delta(x) = \underline{\delta}(x) + (\overline{\delta}(x) - \underline{\delta}(x))u(\beta_0(x)) \text{ for } u(\cdot) \in \mathcal{U}\right\}$ and $\Delta_n^M = \left\{(\delta(X_1), \ldots, \delta(X_n))\colon \delta \in \Delta^M\right\}$. Then,*

$$\overline{perf}_+(s; \beta, \Delta) := \sup_{\delta \in \Delta^M} perf_+(s; \beta, \delta),$$
$$\widehat{\overline{perf}}^k_+(s; \beta, \Delta_n) := \max_{\tilde{\delta} \in \Delta_n^M} \frac{\mathbb{E}_n^k[\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i]}{\mathbb{E}_n^k[\phi_1(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i]} \text{ for any } k.$$

That is, optimization over the set of Δ in the MOSM is equivalent to optimization over the class of monotone, non-decreasing functions on the real line Δ^M. Intuitively, the extremal bounding function that achieves the bounds is either equal to the lower bounding function $\underline{\delta}(x)$ everywhere, the upper bounding function $\overline{\delta}(x)$ everywhere, or can be represented as a non-decreasing step-function that jumps from the lower bounding function to the upper bounding function depending on the value of $\beta_{0,i}$. Lemma 5.2 establishes this formally. Since the class of functions Δ^M is a sufficiently simple function class, we can apply uniform concentration inequalities to derive the convergence rate of $\widehat{\text{perf}}(s; \beta, \Delta_n)$ to the true bound under the MOSM.

Theorem 5.2. *Define the remainder $R_{1,n}^k = \|\hat{\mu}_{1,k}(\cdot) - \mu_{1,k}(\cdot)\|\|\hat{\pi}_{1,k}(\cdot) - \pi_{1,k}(\cdot)\|$ for each fold $k = 1, \ldots, K$. Assume that (i) there $\delta > 0$ such that $\mathbb{P}(\pi_1(X_i) \geq \delta) = 1$; (ii) there exists $\epsilon > 0$ such that $\mathbb{P}(\hat{\pi}(X_i) \geq \epsilon) = 1$; and (iii) $\|\hat{\mu}_{1,k}(\cdot) - \mu_{1,k}(\cdot)\| = o_P(1)$ and $\|\hat{\pi}_{1,k}(\cdot) - \pi_{1,k}(\cdot)\| = o_P(1)$ for each fold $k = 1, \ldots, K$. Then,*

$$\left\|\widehat{\overline{perf}}_+(s; \beta, \Delta_n) - \overline{perf}_+(s; \beta, \Delta)\right\| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R_{1,n}^k\right)$$
$$\left\|\widehat{\underline{perf}}_+(s; \beta, \Delta_n) - \overline{perf}_+(s; \beta, \Delta)\right\| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R_{1,n}^k\right).$$

Theorem 5.2 shows that the errors associated with our proposed estimators of the bounds on positive-class predictive performance under the MOSM consist of two remainders. The first remainder $R_{1,n}$ is the same doubly-robust remainder that we encountered in Theorem 5.1. The use of efficient influence functions and cross-fitting in the construction of our estimators again means that we can effectively control bias from the nonparametric estimation of nuisance parameters. The bounded complexity of Δ^M implies that we pay no penalty in terms of the rate of convergence for the inherent optimization. Our proposed estimator continues to converge at fast rates to the true bounds.

Finally, in Appendix C, we also develop estimators for the bounds on the positive-class predictive disparities of the risk assessment $s(X_i)$ under the MOSM. By a similar argument as given in the proof of Theorem 5.2, we again show that these estimators converge at a fast rate.

5.2.1 Incorporating estimated bounding functions under the MOSM

Consider the case in which the user specifies nonparametric outcome regression bounds under the MOSM, and so must construct estimates of the bounding functions $\hat{\overline{\delta}}(\cdot), \hat{\underline{\delta}}(\cdot)$. We now analyze how using estimated bounding functions affects the convergence rate of our estimated of the upper-bound on positive class predictive performance. In this case, the upper-bound on positive-class predictive performance is estimated by solving the maximization problem in each fold

$$\widehat{\overline{\text{perf}}}^k_+(s;\beta,\hat{\Delta}_n) := \max_{\tilde{\delta}\in\hat{\Delta}_n} \frac{\mathbb{E}^k_n[\beta_{0,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \beta_{0,i}(1-D_i)\tilde{\delta}_i \mid \mathcal{O}_{-k}]}{\mathbb{E}^k_n[\phi_1(Y_i;\hat{\eta}_{-k}) + (1-D_i)\tilde{\delta}_i \mid \mathcal{O}_{-k}]},$$

where $\hat{\Delta}_n := \left\{\delta \in \mathbb{R}^n : \hat{\underline{\delta}}(X_i) \le \delta_i \le \hat{\overline{\delta}}(X_i) \text{ for } i = 1,\dots,n\right\}$. In order to analyze the convergence rate of $\widehat{\overline{\text{perf}}}(s;\beta,\hat{\Delta}_n)$, notice that

$$\|\widehat{\overline{\text{perf}}}^k_+(s;\beta,\hat{\Delta}_n) - \overline{\text{perf}}_+(s;\beta,\Delta)\| \le \|\widehat{\overline{\text{perf}}}^k_+(s;\beta,\hat{\Delta}_n) - \widehat{\overline{\text{perf}}}^k_+(s;\beta,\Delta_n)\| + \|\widehat{\overline{\text{perf}}}^k_+(s;\beta,\Delta_n) - \overline{\text{perf}}_+(s;\beta,\Delta)\|.$$

It is therefore sufficient to bound the extent to which the fold-specific estimator using the estimated bounds $\hat{\Delta}_n$ affects our estimator relative to the oracle bounds Δ_n. Our next result shows that this is bounded by the mean squared error of the estimated bounds.

Proposition 5.3. *Assume the same conditions as Theorem 5.2 and the estimated bounding functions satisfy $P(\phi_1(Y_i;\hat{\eta}_{-k}) + (1-D_i)\hat{\underline{\delta}}_1) > c) = 1$ for some $c > 0$. Then, for each fold k,*

$$\|\widehat{\overline{perf}}^k_+(s;\beta,\hat{\Delta}_n) - \widehat{\overline{perf}}^k_+(s;\beta,\Delta_n)\| \lesssim \sqrt{\frac{1}{n}\sum_{i=1}^{n_k}(\hat{\underline{\delta}}_i - \underline{\delta}_i)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n_k}(\hat{\overline{\delta}}_i - \overline{\delta}_i)^2},$$

where $a \lesssim b$ means $a \le Cb$ for some constant C.

This result immediately implies convergence rates for our positive-class estimator using estimated bounding functions. Since our estimated bounding functions are based on efficient influence functions, the main conclusions of Theorem 5.2 continue to hold as their mean-squared errors are $o_{\mathbb{P}}(1)$ provided the product of nuisance functions are estimated accurately. The following corollary is an immediate consequence. For nonparametric outcome regression bounds, the estimated bounding functions are defined to be $\hat{\underline{\delta}}_i = (\underline{\Gamma} - 1)\phi_1(Y_i;\hat{\eta}_{-k})$, $\hat{\overline{\delta}}_i = (\overline{\Gamma} - 1)\phi(Y_i;\hat{\eta}_{-k})$.

Corollary 5.1. *Suppose the user specifies nonparametric outcome regression bounds for some $\underline{\Gamma}, \overline{\Gamma} > 0$. Under the same conditions as Theorem 5.2, then*

$$\left\|\widehat{\overline{perf}}_+(s;\beta,\hat{\Delta}_n(\Gamma)) - \overline{perf}_+(s;\beta,\Delta(\Gamma))\right\| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K} R^k_{1,n}\right),$$

and analogously for $\widehat{\underline{perf}}_+(s;\beta,\hat{\Delta}_n(\Gamma))$.

Remark 1. Our analysis can be directly extended to the case in which the user specifies instrumental variable bounds under the MOSM by redefining our estimator of the bounds on positive-class predictive performance. In this case, we construct a fold-specific estimate of the upper bound by instead solving

$$\widehat{\text{perf}}(s; \beta; \Delta_n(z)) := \max_{\tilde{\delta} \in \Delta_n(z)} \frac{\mathbb{E}_n^k[\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}\tilde{\delta}_i]}{\mathbb{E}_n^k[\phi_1(Y_i; \hat{\eta}_{-k}) + \tilde{\delta}_i]},$$

for $\Delta_n = \{\tilde{\delta} \in \mathbb{R}^n \colon \underline{\delta}_z(X_i) \leq \tilde{\delta}_i \leq \overline{\delta}_z(X_i)\}$, where $\underline{\delta}_z(x), \overline{\delta}_z(x)$ are defined earlier in Section 3.3. We define $\widehat{\text{perf}}(s; \beta; \widehat{\Delta}_n(z)$ as before for $\widehat{\Delta}_n = \{\widehat{\underline{\delta}}_z(X_i) \leq \tilde{\delta}_i \leq \widehat{\overline{\delta}}_z(X_i)\}$. This allows us to express the estimated instrumental variable bounding functions $\widehat{\underline{\delta}}_z(X_i), \widehat{\overline{\delta}}_z(X_i)$ as being linear in the nuisance functions and extend Proposition 5.3 using the same arguments.

6 Monte Carlo simulations

In this section, we conduct Monte Carlo simulations to show that the theoretical guarantees for our proposed DR-Learners for the MOSM bounds on the target regression and proposed estimators for robust audits under the MOSM hold well for reasonable choices of the sample size n and dimension of the features d.

6.1 Integrated mean-square error of DR-Learners for target regression bounds

We first compare the integrated mean-square error of the DR-Learners for the target regression bounds against the integrated mean-square error of the oracle DR-Learner and a plug-in learner as the sample size n and dimension of the features d vary.

Simulation design: We generate data satisfying the MOSM with nonparametric outcome bounds (Section 2.3). We simulate the features $X_i = (X_{i,1}, \dots, X_{i,d})' \sim N(0, I_d)$. We simulate the intervention $D_i \in \{0,1\}$ conditional on X_i according to $\mathbb{P}(D_i = 1 \mid X_i = x) = \sigma\left(\frac{1}{2\sqrt{d_\pi}} \sum_{d=1}^{d_\pi} X_{i,d}\right)$ for some $d_\pi \in \{1, \dots, d\}$, where $\sigma(a) = \frac{\exp(a)}{1+\exp(a)}$. We simulate the potential outcomes $(Y_i(0), Y_i(1))$ conditional on (D_i, X_i) according to

$$\mathbb{P}(Y_i(1) = 1 \mid D_i = 1, X_i = x) = \sigma\left(\frac{1}{2\sqrt{d_\mu}} \sum_{d=1}^{d_\mu} X_{i,d}\right)$$
$$\mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i = x) = \Gamma_{true}\sigma\left(\frac{1}{2\sqrt{d_\mu}} \sum_{d=1}^{d_\mu} X_{i,d}\right),$$

for some $d_\mu \in \{1, \dots, d\}$. By construction, this data generating process satisfies the MOSM with nonparametric outcome bounds at $\underline{\Gamma} \leq \Gamma_{true} \leq \overline{\Gamma}$, and we set $\Gamma_{true} = 0.75$. Under this data-generating process, we simulate training datasets (X_i, D_i, Y_i) for $i = 1, \dots, n$.

We compare three different estimators for the MOSM bounds on the target regression: first, our DR-Learners proposed in Section 3.1; second, the infeasible oracle, which use the true outcome regression and propensity score as defined in Section 3.2; and finally, a plug-in learner, which simply plugs-in nuisance function estimates based on the same fold as the second-stage nonparametric regression and does not use pseudo-outcomes based on the efficient influence function. Recall Theorem 3.1 established that the integrated mean-square error of our proposed DR-Learners converge at fast rates to the integrated mean-square error of the infeasible oracle DR-Learners. In contrast, by not using sample-splitting, the plug-in estimator may inherit the errors from estimating the individual nuisance functions.

Simulation results: We first analyze the performance of the DR-Learner for the upper-bound on the target regression for a fixed choice $\underline{\Gamma} = 2/3$, $\overline{\Gamma} = 3/2$, and evaluate how well it recovers the true upper bound $\overline{\mu}(\cdot;\Delta(\Gamma))$. Across 1000 simulated datasets of varying size $n \in \{1000, 2500, 5000\}$ and dimension $d \in \{100, 500\}$, we calculate the DR-Learner $\widehat{\overline{\mu}}(\cdot;\Delta(\Gamma))$, the oracle learner $\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(\Gamma))$, and the plug-in learner $\widehat{\overline{\mu}}_{plugin}(\cdot;\Delta(\Gamma))$. Since the outcome regression and propensity score models are known under the simulation design, we can directly calculate the true upper bound $\overline{\mu}(\cdot;\Delta(\Gamma))$ for any choice of $\underline{\Gamma}, \overline{\Gamma}$. Throughout we set $d_\pi = 20$ and $d_\mu = 25$. The estimators are constructed using a single split of the evaluation data (except for the plug-in learner), and the first-stage nuisance functions $\eta = (\pi_1(X_i), \mu_1(X_i))$ are estimated using cross-validated Lasso. We find analogous results for the estimators of the lower bound on the target regression.

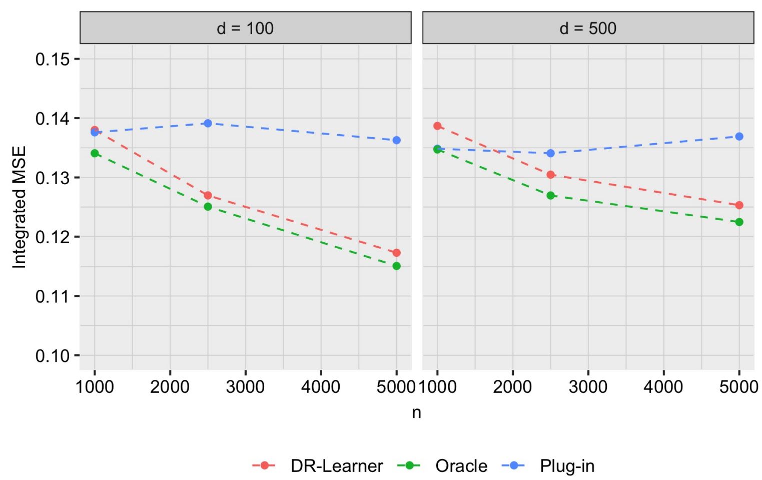


Figure 1: Average integrated mean square of DR-Learner, oracle learner, and plug-in learner across Monte Carlo simulations with nonparametric outcome bounds.

Notes: This figure plots the average integrated mean square error of the DR-Learner, oracle learner, and plug-in learner for the upper bound on the target regression $\overline{\mu}(\cdot;\Delta(\Gamma))$ across Monte Carlo simulations. We report these results for $n \in \{1000, 2500, 5000\}$. The DR-Learner is constructed using a single split. The nuisance functions are estimated using cross-validated Lasso for all estimators, and all estimators assume $\underline{\Gamma} = 2/3$, $\overline{\Gamma} = 3/2$. The results are computed over 1,000 simulations. See Section 6.1 for further details on the simulation design.

We report the average integrated mean square error of each estimator for the true upper bound $\overline{\mu}(\cdot;\Delta(\Gamma))$. The results are summarized in Figure 1. As n grows large, the integrated mean square error of the DR-Learner converges to zero with the integrated mean square error of the oracle learner as expected from Theorem 3.1. While it is competitive with the DR-Learner for smaller sample sizes ($n = 1000$), the integrated mean square for the plug-in learner is relatively constant across sample sizes n and dimension of the features d. As a result, as n grows larger, the plug-in learner performs poorly relative to the DR-Learner – for example, at $n = 5000$ and $d = 500$, the DR-Learner's integrated mean square error is only 1.9% larger than that of the oracle learner, whereas the plug-in learner's integrated mean square error is 18.4% larger. In contrast, by leveraging both sample-splitting and pseudo-outcomes based on efficient influence functions, the DR-Learner quickly converges to the true target regression bound and improves upon simple plug-in estimation approaches.

6.2 Finite sample behavior of estimators for robust audits

We show that our proposed estimators for the bounds on overall predictive performance under the MOSM converge quickly to the true bounds and that the associated confidence intervals for the bounds based on the derived asymptotic normal approximation have good coverage properties for reasonable choices of the sample size n and dimension of the features d. Appendix D.1 shows that our proposed estimators for the bounds on the true positive rate and false positive rate converge quickly to the true bounds as well.

Simulation design: We again generate data satisfying the MOSM with nonparametric outcome bounds. We now simulate the features $X_i \sim U([0,1]^d)$, and, for a randomly drawn coefficient vector μ, we simulate the intervention $D_i \in \{0,1\}$ conditional on X_i according to $\mathbb{P}(D_i = 1 \mid X_i = x) = \sigma(X_i'\mu)$. For coefficient vectors β_0, β_1 and some $\Gamma_{true} > 0$, we finally simulate the potential outcomes $(Y_i(0), Y_i(1))$ conditional on D_i, X_i according to

$$\mathbb{P}(Y_i(1) = 1 \mid D_i = 1, X_i = x) = \sigma(X_i'\beta_1),$$
$$\mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i = x) = \Gamma_{true}\sigma(X_i'\beta_1).$$

This data generating process satisfies the MOSM with nonparametric outcome bounds at $\underline{\Gamma} \leq \Gamma_{true} \leq \overline{\Gamma}$, and we again set $\Gamma_{true} = 0.75$.

Under this data generating process, we first simulate a training dataset (X_i, D_i, Y_i) for $i = 1, \ldots, n_{train}$, and estimate a risk score $s(\cdot)$ that predicts $Y_i = 1$ only on the selected data $D_i = 1$. Keeping this estimated risk score fixed, we then robustly audit the overall performance of the risk score under various assumptions on the strength of unmeasured confounding using evaluation data (X_i, D_i, Y_i) for $i = 1, \ldots, n$ simulated from the same data generating process. We robustly audit the risk score's accuracy $\text{perf}(s; \beta_{acc})$ and mean square error $\text{perf}(s; \beta_{mse})$ (as defined in Example 1) under nonparametric outcome bounds $\Delta(\Gamma)$ under alternative choices of $\Gamma = (\underline{\Gamma}, \overline{\Gamma})$.

Simulation results: We first audit the overall predictive performance of the risk score for a fixed choice $\underline{\Gamma} = 2/3$, $\overline{\Gamma} = 3/2$, and evaluate how well our proposed estimators recover the true bounds $[\underline{\text{perf}}(s; \beta, \Delta(\Gamma)), \overline{\text{perf}}(s; \beta, \Delta(\Gamma))]$. Across 1000 simulated evaluation datasets of varying size $n \in \{500, 1000, 2500\}$, we calculate the estimates $[\widehat{\underline{\text{perf}}}(s; \beta, \Delta(\Gamma)), \widehat{\overline{\text{perf}}}(s; \beta, \Delta(\Gamma)]$. The estimators are constructed using a single split of the evaluation data, and we estimate the first-stage nuisance functions $\eta = (\pi_1(X_i), \mu_1(X_i))$ using random forests. Since the outcome regression and propensity score models are known under the simulation design, we can directly calculate the true bounds $[\underline{\text{perf}}(s; \beta, \Delta(\Gamma)), \overline{\text{perf}}(s; \beta, \Delta(\Gamma))]$ for the chosen values of $\underline{\Gamma}, \overline{\Gamma}$. We report the average bias of our estimators for the true bounds on overall performance as well as the estimated coverage rate of a 95% nominal confidence interval for the true bounds based on the asymptotic normal approximation derived in Proposition 5.1.

Table 1(a) summarizes these results for our estimator of the upper bound on the risk score's mean square error, and Table 1(b) summarizes these results for our estimator of the lower bound on the risk score's accuracy. We find analogous results for our estimator of the lower bound on the risk score's mean square error and estimator of the upper bound on the risk score's accuracy. Our proposed estimators are approximately unbiased for the true bounds. Their estimated standard errors slightly underestimate the true standard errors when the size of the evaluation dataset is small, but are quite accurate for $n \geq 1000$. Consequently, confidence intervals based on the asymptotic normal approximation for our proposed estimators have approximately 95% coverage for both the upper bound on mean square error and lower bound on accuracy (up to simulation error). Figure 2 depicts that our proposed estimators are approximately normally distributed in finite samples and concentrate around the true bounds quickly as the size of the evaluation data grows large. Altogether, these numerical results indicate our theoretical analysis of the limiting distribution of our estimators for the bounds on overall predictive performance provide good guidance about their finite sample behavior.

n	Bias of $\widehat{\overline{\text{perf}}}(s; \beta_{mse}, \Delta(\Gamma))$	SD. of $\widehat{\overline{\text{perf}}}(s; \beta_{mse}, \Delta(\Gamma))$	$\widehat{\sigma}$	Coverage
500	0.000	0.016	0.015	0.942
1000	0.000	0.010	0.010	0.945
2500	0.000	0.006	0.006	0.944

(a) Upper bound on mean square error

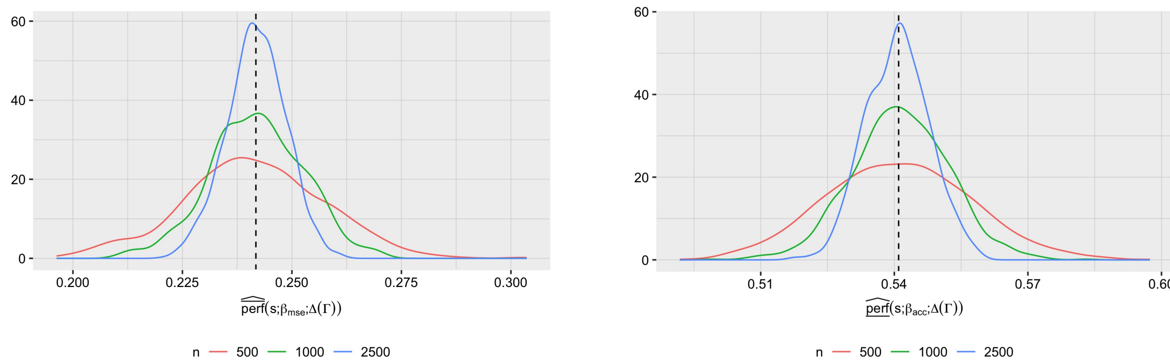
n	Bias of $\widehat{\underline{\text{perf}}}(s; \beta_{acc}, \Delta(\Gamma))$	SD. of $\widehat{\underline{\text{perf}}}(s; \beta_{acc}, \Delta(\Gamma))$	$\widehat{\sigma}$	Coverage
500	0.001	0.015	0.015	0.941
1000	0.001	0.010	0.010	0.946
2500	0.000	0.006	0.006	0.947

(b) Lower bound on accuracy

Table 1: Bias and coverage properties of overall performance estimators of a risk score $s(\cdot)$ with nonparametric outcome bounds.

Notes: This table summarizes the average bias of our estimator $\widehat{\overline{\text{perf}}}(s; \beta_{mse}, \Delta(\Gamma))$ for the upper bound on MSE and the lower bound on accuracy $\widehat{\underline{\text{perf}}}(s; \beta_{acc}, \Delta(\Gamma))$, the standard deviation of our estimator (SD. of $\widehat{\overline{\text{perf}}}(s; \beta_{MSE}, \Delta(\Gamma))$ and $\widehat{\underline{\text{perf}}}(s; \beta_{acc}, \Delta(\Gamma))$), the average estimated standard error of our estimator ($\widehat{\sigma}$), and the coverage rate of nominal 95% confidence intervals based on the asymptotic normal approximation in Proposition 5.1. The overall performance estimators are constructed using a single split, and the nuisance functions are estimated using random forests. The overall performance estimators assume that $\underline{\Gamma} = 2/3$, $\overline{\Gamma} = 3/2$. The results are computed over 1,000 simulations. See Section 6.2 for further details on the simulation design.

Finally, we examine how the performance of our proposed estimators vary as our assumptions on the magnitude of unobserved confounding vary (i.e., the choice $\underline{\Gamma}, \overline{\Gamma}$). To do so, we set $\underline{\Gamma} = 1/\widetilde{\Gamma}$, $\overline{\Gamma} = \widetilde{\Gamma}$ for $\widetilde{\Gamma} \geq 1$, and report results varying $\widetilde{\Gamma} \in \{1, \ldots, 2.5\}$. For each choice of $\widetilde{\Gamma}$, we again simulate 1,000 evaluation datasets of size $n = 2500$ and calculate estimates the $[\widehat{\underline{\text{perf}}}(s; \beta, \Delta(\Gamma)), \widehat{\overline{\text{perf}}}(s; \beta, \Delta(\Gamma)]$. Across all choices of $\widetilde{\Gamma}$, the estimated coverage of 95% nominal confidence intervals for the upper bound on mean square error never dips below 94.5%, 93.3% for the lower bound on mean square error, 93.9% for the upper bound on accuracy, and 94.5% for the lower bound on accuracy. Our theoretical analysis therefore also provides reasonable guidance as assumptions on the magnitude of unmeasured confounding varies.



(a) Mean square error, $\widehat{\overline{\text{perf}}}(s;\beta_{mse},\Delta(\Gamma))$

(b) Accuracy, $\widehat{\underline{\text{perf}}}(s;\beta_{acc},\Delta(\Gamma))$

Figure 2: Distribution of overall performance estimators across Monte Carlo simulations with nonparametric outcome bounds.

Notes: This figure plots the distribution of the overall performance estimator for the upper bound on the mean square error $\overline{\text{perf}}(s;\beta_{mse},\Delta(\Gamma))$ (A) and the lower bound on the accuracy $\underline{\text{perf}}(s;\beta_{acc},\Delta(\Gamma))$ (B) of a risk score $s(\cdot)$ across Monte Carlo simulations. We report these results for $n \in \{500, 1000, 1500\}$ (color). The vertical dashed lines show the true upper bound on mean square error $\overline{\text{perf}}(s;\beta_{mse},\Delta(\Gamma))$ and the true lower bound on accuracy $\underline{\text{perf}}(s;\beta_{acc},\Delta(\Gamma))$. The overall performance estimators are constructed using a single split, and the nuisance functions are estimated using random forests. The overall performance estimators assume that $\underline{\Gamma} = 2/3$, $\overline{\Gamma} = 3/2$. We report these results for $n \in \{500, 1000, 2500\}$ (colors). The results are computed over 1,000 simulations. See Section 6.2 for further details on the simulation design.

7 Empirical illustration: consumer lending

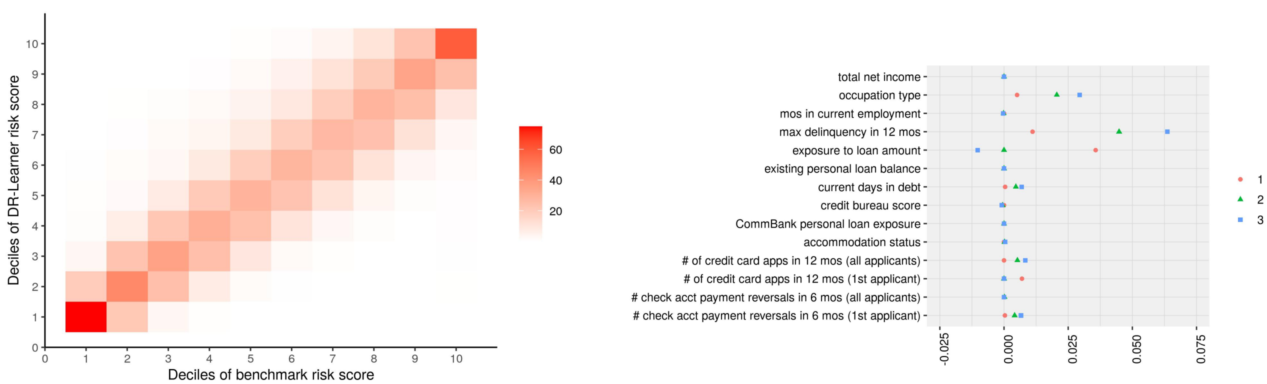
A financial institution wishes to construct a credit risk score that predicts whether an application will default $Y_i(1) \in \{0,1\}$ based on application-level features, and audit the predictive performance of an existing credit risk score. The financial institution observes historical data on past, submitted applications, but it only observes whether a past application defaulted on their loan *if* the application was "funded" (i.e., approved by the financial institution *and* the offered terms were accepted by the applicant). As a result, the funding decision $D_i \in \{0,1\}$ may be subject to unobserved confounding as applicants may differ in unobserved ways that jointly affect their default risk and likelihood of accepting an offered loan. For example, an applicant's decision to accept an offered loan may depend on whether they secured another credit offer at a competing financial institution. We now illustrate empirically how our framework can be used to learn and audit a credit risk score that is robust to such unobserved confounding.

We use data from Commonwealth Bank of Australia ("CommBank"), a large financial institution in Australia, on 372,346 applications submitted from July 2017 to July 2019. As discussed in Coston, Rambachan and Chouldechova (2021), personal loans are paid back in monthly installments, and used, for example, to purchase a used car, refinance existing debt, pay for funeral and wedding expenses among many other purposes. In the time period considered, personal loan sizes included amounts up to AU\$50,000 with a median of AU\$10,000, and the median offered interest rate was 14.9% per annum in our sample of applications. We observe rich application-level features such as, for example, the applicant's reported income, the applicant's occupation, and their credit history at CommBank. We only observe whether an applicant defaulted on the personal loan within 5 months $Y_i(1) \in \{0,1\}$ if the application was funded $D_i = 1$ ($Y_i(0) := 0$ since unfunded applications cannot default). In our sample of applications, approximately one-third of the applications were funded, and 2.0% of all funded loans defaulted within 5 months.

7.1 Bounding individual default risk

We first construct a credit risk score to predict individual default risk $Y_i(1)$ as a function of application-level features X_i and is robust to unobserved confounding. To do so, we use the DR-Learner to construct the upper bound on the target regression under the MOSM with nonparametric outcome regression bounds, $\overline{\mu}(\cdot; \Delta(\Gamma))$. We set $\underline{\Gamma} = 1$ and report results as $\overline{\Gamma} \in \{1, 2, 3\}$ varies. This implies that, conditional on observable application-level features X_i, we assume that unfunded applications are at least as likely to default as funded applications but may be no more than $\overline{\Gamma}$ times as likely to default as funded applications.

To construct the DR-Learner, we first split our sample of applications into two folds. In the first fold, we construct estimates of the nuisance functions $\pi_1(\cdot)$, $\mu_1(\cdot)$ using random forests. On the second fold, we regress the pseudo-outcome $\phi_1(Y_i; \hat{\eta}) + \overline{\Gamma}\phi(\pi_0(X_i)\mu_1(X_i); \hat{\eta})$ on the features X_i using cross-validated logistic regression with a Lasso penalty. This yields the estimated DR-Learner for the upper bound on 5-month default risk, $\widehat{\overline{\mu}}(X_i; \Delta(\Gamma))$.



(a) Joint distribution of benchmark risk score vs. DR-Learner.

(b) Coefficients in risk score as $\overline{\Gamma}$ varies.

Figure 3: Estimated personal loan credit risk scores as assumptions on unobserved confounding vary.

Notes: The left panel summarizes the joint distribution of a benchmark credit risk score's predictions of default risk against the DR-Learner's predictions of the default risk. Among applications in each decile of the benchmark credit score's predicted risk distribution, the left panel plots the percentage of applications at each decile of the DR-Learner's predicted risk distribution. The DR-Learner is constructed assuming $\underline{\Gamma} = 1$, $\overline{\Gamma} = 2$, and the benchmark credit score predicts default risk among only funded applications (i.e., it is the DR-Learner constructed assuming $\underline{\Gamma} = \overline{\Gamma} = 1$). The right panel summarizes how the coefficients on a subset of application-level characteristics vary as our assumptions on unobserved confounding varies $\overline{\Gamma} \in \{1, 2, 3\}$. The value $\overline{\Gamma} = 1$ corresponds to the benchmark credit score. Table A3 provides a detailed description of the variable names in the right panel. See Section 7.1 for further discussion.

In order to understand how varying our assumptions on unobserved confounding affects the resulting credit risk score, we compare the DR-Learner's predictions of default risk against a benchmark credit score that simply predicts default risk among only funded applications. The left panel of Figure 3 summarizes the joint distribution of the benchmark credit score's predictions and the DR-Learner's predictions, where the DR-Learner is constructed assuming $\underline{\Gamma} = 1, \overline{\Gamma} = 2$. Among applications in each decile of the benchmark credit score's predicted risk distribution, the left panel plots the percentage of applications at each decile of the DR-Learner's predicted risk distribution. We find that among applications at any particular decile of the benchmark credit score's predicted risk distribution, there exists meaningful variation in the DR-Learner's risk predictions. Among applications in the 3rd decile and 8th decile of the benchmark risk score's predicted risk distribution, 13.8% of applications are in the top half of the DR-Learner's predicted risk distribution and 4.7% of applications are in the bottom half of the DR-Learner's predicted risk distribution respectively. Similarly, among applications in the 5th decile of the benchmark risk score's predicted risk distribution, 18.7% of applications are flipped to either the bottom (deciles 1-3) or top (deciles 8-10) of the DR-Learner's predicted risk distribution.

The right panel of Figure 3 investigates what drives these differences in predicted default risk by comparing how the estimated coefficients on a subset of application-level characteristics vary as our assumptions on unobserved confounding $\overline{\Gamma}$ vary. Table A3 provides a more detailed description of the variable names in the right panel.[11] Interestingly, altering our assumptions on unobserved confounding leads to effectively no difference in the estimated coefficients for some notable application-level characteristics. For example, the estimated coefficient on the application's total net income and credit bureau score do not vary as $\overline{\Gamma}$ varies and are always equal to zero. In contrast, for some characteristics such as the number of credit card applications submitted by all applicants on the application, the benchmark risk score places no weight them, whereas all DR-Learners with $\overline{\Gamma} > 1$ incorporate them into the model and assign a non-zero weight. For other characteristics like the the applicant's occupation type or their maximum delinquency over the last 12 months, varying our assumptions on unobserved confounding lead to large changes in the magnitudes of the estimated coefficients. Altogether, these results highlight that explicitly accounting for unobserved confounding may lead to substantive differences in the resulting credit risk score.

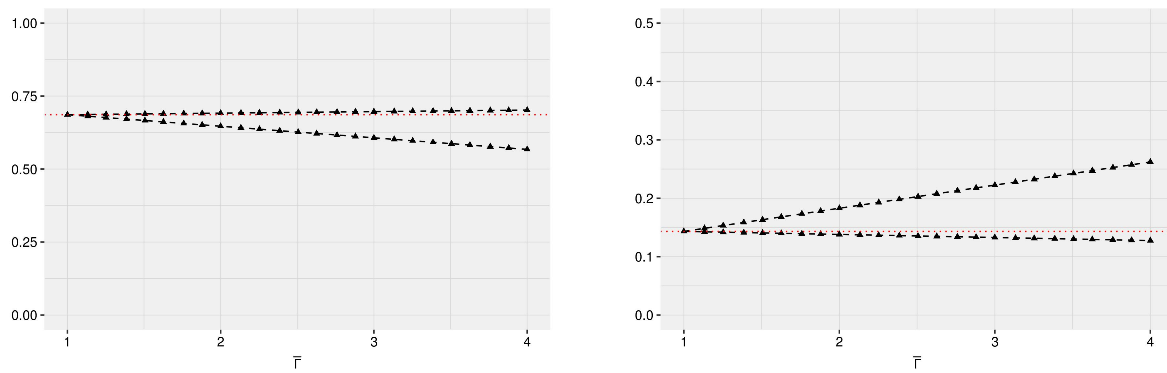
7.2 Robust audits of a credit score

We next robustly audit an existing credit risk score for its overall predictive performance and its predictive performance across various subgroups of interest. To do so, we again split our sample of applications into two subsets, which we will refer to now as the training data and evaluation data. In the training data, we estimate a benchmark risk score that predicts default only among funded applications using a random forest estimated on application-level characteristics. Keeping this benchmark risk score fixed, we then use the evaluation data to analyze its predictive performance. We implement our estimators by splitting the evaluation data into two folds; on the first fold, we construct estimates of the nuisance functions $\pi_1(\cdot), \mu_1(\cdot)$, and plug them into our proposed estimators from Section 5 in the second fold. We again bound unobserved confounding using the MOSM with nonparametric outcome regression bounds, setting $\underline{\Gamma} = 1$ and reporting results as $\overline{\Gamma} \in [1, 4]$ varies.

We first investigate how the bounds on the benchmark risk score's overall predictive performance varies as our assumptions on unobserved confounding vary. We evaluate its overall accuracy and mean square error; the results are summarized in Figure 4. Panel (A) illustrates how the upper and lower bounds (black) on the benchmark risk score's accuracy varies as $\overline{\Gamma}$ varies, and Panel (B) reports the analogous results for its mean square error. When $\overline{\Gamma} = 1$, the MOSM with nonparametric outcome regression bounds is equivalent to assuming that there is no unmeasured confounding; therefore, the upper and lower bounds both equal the benchmark risk score's predictive performance over only funded

[11] Personal loans applications can have multiple listed applicants, and so some variables refer to just the 1st listed applicant.

applications (dashed red line). As $\overline{\Gamma}$ increases, the MOSM with nonparametric outcome regression bounds now allows default rates among unfunded applications to now differ, and so the estimated bounds widen. Importantly, as $\overline{\Gamma}$ increases, the bounds on the benchmark risk score's accuracy and mean square error remain informative. For example, at $\overline{\Gamma} = 3$, which allows unfunded applications to be no more than 3 times as likely to default as funded applications conditional on observed application-level features, the lower bound on accuracy is 0.60 and the upper bound is 0.69.



(a) Accuracy

(b) Mean square error

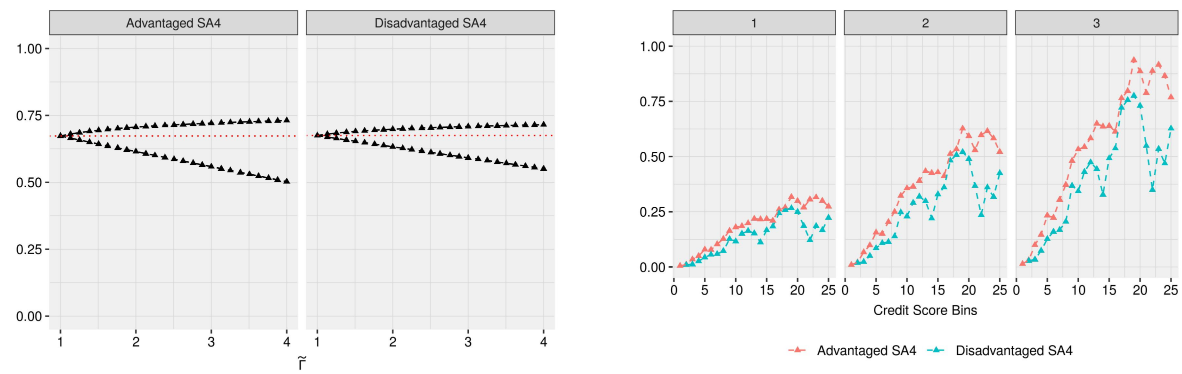
Figure 4: Overall predictive performance of benchmark risk score as assumptions on unobserved confounding varies.

Notes: The left panel summarizes how the bounds on the benchmark risk score's overall accuracy varies as the our assumptions on unobserved confounding varies. The right panel summarizes how the bounds on its overall mean square error varies. The bounds on overall predictive performance (black) are constructed under the MOSM with nonparametric outcome regression bounds, setting $\underline{\Gamma} = 1$ and varying $\overline{\Gamma} \in [1, 4]$. The overall predictive performance over only funded applications is plotted in the red dashed line. The estimated bounds on overall predictive performance across constructed using a single split of the evaluation data, and the nuisance functions are estimated using random forests. The benchmark risk score is constructed by predicting default risk among only funded applications in the training data. See Section 7.2 for further details.

Finally, we use our estimators to investigate the benchmark risk score's predictive disparities. Following Coston, Rambachan and Chouldechova (2021), we focus on disparities in predictive performance across SA4 geographic regions in Australia. An SA4 geographic region is a statistical area defined by the Australian Bureau of Statistics, and we define an SA4 region to be "socioeconomically disadvantaged" if it falls in the top quartile of SA4 regions based on the Australian Bureau of Statistics Index of Relative Socioeconomic Disadvantage (IRSD).[12] The left panel of Figure 5 summarizes how the bounds on the benchmark risk score's accuracy vary across socioeconomically advantaged and disadvantaged SA4 regions. We find qualitatively similar results for the benchmark risk score's mean square error. Among only funded applications, we observe that the benchmark risk score's accuracy is approximately the same across advantaged and disadvantaged SA4 regions in the left panel. As before, the bounds on the benchmark risk score's accuracy widen for both groups as $\overline{\Gamma}$ grows larger. Interestingly, however, the bounds are wider over advantaged SA4 regions than disadvantaged SA4 regions at each choice of $\overline{\Gamma}$,

[12]The Index of Relative Socioeconomic Disadvantage is an index constructed by the Australian Bureau of Statistics that summarizes census data related to socioeconomic disadvantage, such as the average household income and the fraction of households without internet access. See Australian Bureau of Statistics (2016) for complete details on the construction of the IRSD. The IRSD is constructed at a more granular geographic level (SA2 regions), and so we aggregate the IRSD to the SA4 region level by taking its population-weighted average across all SA2 regions within each SA4 region.

suggesting that the benchmark risk score's accuracy is more sensitive to assumptions about unobserved confounding on advantaged SA4 regions. The right panel summarizes how the upper bound on the benchmark risk score's calibration varies as we vary $\overline{\Gamma} \in \{1, 2, 3\}$ across SA4 regions. We report the benchmark risk score's calibration over 25 equal-sized bins of predicted risk. At each credit score bin and each choice of $\overline{\Gamma}$, the upper bound on the risk score's calibration is larger among socioeconomically advantaged geographic regions than socioeconomically disadvantaged geographic regions. Furthermore, as $\overline{\Gamma}$ grows larger, the gap in the risk score's calibration across geographic regions widens.



(a) Accuracy by SA4 region

(b) Calibration by SA4 region

Figure 5: Accuracy and calibration of estimated risk score across SA4 regions as assumptions on unobserved confounding vary.

Notes: The left panel summarizes how the bounds on the benchmark risk score's accuracy varies as our assumptions on unobserved confounding varies across socioeconomically advantaged and disadvantaged SA4 geographic regions. The right panel summarizes how the upper bound on its calibration varies as our assumptions on unobserved confounding varies across socioeconomically advantaged and disadvantaged SA4 geographic regions. The bounds on predictive performance are constructed under the MOSM with nonparametric outcome regression bounds, setting $\underline{\Gamma} = 1$ and varying $\overline{\Gamma} \in [1, 4]$. In the left panel, the bench mark risk score's accuracy over only funded applications is plotted in the red dashed line. The estimated bounds on predictive performance across constructed using a single split of the evaluation data, and the nuisance functions are estimated using random forests. The benchmark risk score is constructed by predicting default risk among only funded applications in the training data. See Section 7.2 for further details.

Altogether, these results highlight that explicitly accounting for unobserved confounding meaningfully affects how we assess the predictive performance of credit risk scores, and our estimators enable users to tractably report sensitivity analyses that vary their assumptions on unobserved confounding.

8 Connections to existing sensitivity analysis models

We now formally relate the MOSM to existing approaches to modelling unobserved confounding in the causal inference literature. We discuss how existing approaches imply the MOSM using non-parametric outcome regression bounds for particular choices of $\underline{\Gamma}, \overline{\Gamma} > 0$. In this sense, the MOSM places weaker restrictions on unobserved confounding than these existing approaches, but our methods nonetheless enable users to robustly learn and evaluate risk assessments in high-stakes settings.

We emphasize that we do not view the MOSM as being in competition with these existing sensitivity analysis models. In contrast, users must have a suite of options that can be used depending on what is most intuitive to them.

8.1 Marginal sensitivity model

A recently popular model used for sensitivity analysis is the *marginal sensitivity model* (MSM), which is a nonparametric relaxation of unconfoundedness that restricts the extent to which unobserved confounders may impact the odds of being treated vs. untreated. The MSM specifies that, for some $\Lambda \geq 1$, $(X_i, D_i, Y_i(0), Y_i(1)) \sim \mathbb{P}(\cdot)$ satisfies

$$\Lambda^{-1} \leq \frac{\mathbb{P}(D_i = 1 \mid X_i, Y_i(0), Y_i(1))}{\mathbb{P}(D_i = 0 \mid X_i, Y_i(0), Y_i(1))} \frac{\mathbb{P}(D_i = 0 \mid X_i)}{\mathbb{P}(D_i = 1 \mid X_i)} \leq \Lambda \tag{22}$$

with probability one. The MSM nests the special case of unconfoundedness by setting $\underline{\Lambda} = \overline{\Lambda} = 1$. Notice that for the odds ratio in (22) to be well-defined requires overlap to hold conditional both on $(X_i, Y_i(0), Y_i(1))$ and X_i. The MSM was originally proposed by Tan (2006), and has since received substantial attention among researchers (e.g., see Zhao, Small and Bhattacharya, 2019; Kallus, Mao and Zhou, 2018; Dorn and Guo, 2021; Dorn, Guo and Kallus, 2021; Kallus and Zhou, 2021).

We now state a simple proposition showing that relates the MSM to the nonparametric outcome regression bounds on $\delta(x) = \mathbb{E}[Y_i(1) \mid D_i = 0, X_i] - \mathbb{E}[Y_i(1) \mid D_i = 1, X_i]$ under the MOSM.

Proposition 8.1.

i. Suppose that $(X_i, D_i, Y_i(0), Y_i(1)) \sim \mathbb{P}(\cdot)$ satisfies the MSM (22) for some $\Lambda \geq 1$. Then, $\mathbb{P}(\cdot)$ satisfies the MOSM (Assumption 2.1) with $\underline{\delta}(x) = (\Lambda^{-1} - 1)\mu_1(x)$ and $\overline{\delta}(x) = (\Lambda - 1)\mu_1(x)$.

ii. Suppose that $(X_i, D_i, Y_i(0), Y_i(1)) \sim \mathbb{P}(\cdot)$ satisfies $Y_i(0) = 0$ and the MOSM (Assumption 2.1) with nonparametric outcome regression bounds for some $\underline{\Gamma}, \overline{\Gamma} > 0$. Then, $\mathbb{P}(\cdot)$ satisfies

$$\overline{\Gamma}^{-1} \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = 1, X_i)\mathbb{P}(D_i = 0 \mid X_i)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = 1, X_i)\mathbb{P}(D_i = 1 \mid X_i)} \leq \underline{\Gamma}^{-1}, \text{ and}$$

$$\frac{\underline{\Gamma} - 1}{\underline{\Gamma}(1 - \underline{\Gamma}\mu_1(x))} + \underline{\Gamma}^{-1} \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = 0, X_i)\mathbb{P}(D_i = 0 \mid X_i)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = 0, X_i)\mathbb{P}(D_i = 1 \mid X_i)} \leq \frac{\overline{\Gamma} - 1}{\overline{\Gamma}(1 - \overline{\Gamma}\mu_1(x))} + \overline{\Gamma}^{-1}.$$

This first result relates to Proposition 3 in Dorn, Guo and Kallus (2021), which establishes that the MSM implies a bound on $\mathbb{E}[Y_i(1) \mid D_i = 0, X_i]$ via the solution to a conditional value-at-risk problem for general outcomes. For our binary outcome setting, we show in the proof that the MSM directly implies a bound on $\mathbb{E}[Y_i(1) \mid D_i = 0, X_i]$ by an application of Bayes' rule. By an analogous argument, our second result establishes a partial converse, showing that the MOSM implies an MSM-like bound for sample-selection models in which $Y_i(0) \equiv 0$ (e.g., our running credit lending and pretrial release examples). A user that specifies the MSM (22) for conducting sensitivity analyses can, therefore, use our methods to bound the target regression, construct robust decision rules, or conduct robust audits of risk assessments under the MOSM.

8.2 Rosenbaum's Γ-sensitivity model

Another famous framework for conducting sensitivity analysis is Rosenbaum's Γ-*sensitivity analysis model*, which summarizes the violation of the unconfoundedness by bounding the extent to which the odds of being treated vs. untreated may vary across different values of the unobservables (e.g., Rosenbaum, 1987, 2002). The Γ-sensitivity analysis model specifies that for some $\Gamma \geq 1$, $(X_i, D_i, Y_i(0), Y_i(1)) \sim$

$\mathbb{P}(\cdot)$ satisfies

$$\Gamma^{-1} \leq \frac{\mathbb{P}(D_i = 1 \mid X_i, Y_i(1) = y_1, Y_i(0) = y_0)}{\mathbb{P}(D_i = 0 \mid X_i, Y_i(1) = y_1, Y_i(0) = y_0)} \frac{\mathbb{P}(D_i = 0 \mid X_i, Y_i(1) = \tilde{y}_1, Y_i(0) = \tilde{y}_0)}{\mathbb{P}(D_i = 1 \mid X_i, Y_i(1) = \tilde{y}_1, Y_i(0) = \tilde{y}_0)} \leq \Gamma \tag{23}$$

for all $y_0, y_1, \tilde{y}_0, \tilde{y}_1 \in \{0, 1\}$ and with X_i-probability one. Notice that $\Gamma = 1$ again nests the special case of unconfoundedess. As discussed in Section 7 of Zhao, Small and Bhattacharya (2019), Rosenbaum's Γ-sensitivity analysis model was originally proposed to conduct sensitivity analysis on observational experiments conducted on finite populations that have a paired or grouped design, ignoring sampling uncertainty.[13,14] Recently, Yadlowsky et al. (2018) applies Rosenbaum's Γ-sensitivity analysis model to observational settings like we consider, deriving bounds on the conditional average treatment effect under the model, and developing methods for conducting inference on the average treatment effect.

For our purposes, it is sufficient to state a simple proposition that relates Rosenbaum's Γ-sensitivity analysis model to the MOSM with nonparametric outcome regression bounds.

Proposition 8.2.

i. Suppose $(X_i, D_i, Y_i(0), Y_i(1)) \sim \mathbb{P}(\cdot)$ satisfies Rosenbaum's sensitivity analysis model (23) for some $\Gamma > 1$. Then $P(\cdot)$ satisfies the MOSM (Assumption 2.1) with $\underline{\delta}(x) = (\Gamma^{-1} - 1)\mu_1(x)$ and $\overline{\delta}(x) = (\Gamma - 1)\mu_1(x)$.

ii. Suppose $(X_i, D_i, Y_i(0), Y_i(1)) \sim \mathbb{P}(\cdot)$ satisfies $Y_i(0) = 0$ and the MOSM (Assumption 2.1) with nonparametric outcome regression bounds for some $\underline{\Gamma}, \overline{\Gamma} > 0$. Then, $P(\cdot)$ satisfies

$$\frac{\underline{\Gamma} - 1}{1 - \underline{\Gamma}\mu_1(x)} + 1 \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = 0, X_i)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = 0, X_i)} \frac{\mathbb{P}(D_i = 0 \mid X_i, Y_i(1) = 1)}{\mathbb{P}(D_i = 1 \mid X_i, Y_i(1) = 1)} \leq \frac{\overline{\Gamma} - 1}{1 - \overline{\Gamma}\mu_1(x)} + 1.$$

To show the first result, we show that Rosenbaum's Γ-sensitivity model implies a marginal sensitivity model in our binary outcome setting. This in turn implies a MOSM with nonparametric outcome regression bounds. This relates to Lemma 2.2 in Yadlowsky et al. (2018), which shows that a version of Rosenbaum's sensitivity analysis model implies a bound on $\mathbb{E}[Y_i(1) \mid D_i = 0, X_i]$ via the solution to an estimating equation for general outcomes. The second result again establishes a partial converse – the MOSM implies an Rosenbaum-style bound for sample-selection models in which $Y_i(0) \equiv 0$, where the bounds on the odds ratio vary based on the features X_i but only through the identified outcome regression. As a consequence, a user that specifies Rosenbaum's sensitivity analysis model (23) for conducting sensitivity analyses can, therefore, again use our methods to bound the target regression, construct robust decision rules, or conduct robust audits of risk assessments under the MOSM.

8.3 Sensitivity analysis via outcome modelling

Finally, a large literature conducts sensitivity analyses in missing data problems via outcome modelling. A popular approach is to specify flexible parametric models for the difference between the unobserved conditional distribution $Y_i(1) \mid \{X_i, D_i = 0\}$ and the observed conditional distribution

[13] We refer the reader to Section 7 of Zhao, Small and Bhattacharya (2019) for an in-depth comparison of the marginal sensitivity model and Rosenbaum's Γ-sensitivity model.

[14] See also Aronow and Lee (2013) and Miratrix, Wager and Zubizarreta (2018) which also use a version of Rosenbaum's Γ-sensitivity model to construct bounds on a finite-population from a random sample with unknown selection probabilities.

$Y_i(1) \mid \{X_i, D_i = 1\}$, or between $Y_i(1) \mid \{X_i, D_i = 0\}$ and $Y_i(1) \mid X_i$ (e.g., Rotnitzky et al., 2001; Birmingham, Rotnitzky and Fitzmaurice, 2003; Brumback et al., 2004; Franks, Airoldi and Rubin, 2020). For example, Robins, Rotnitzky and Scharfstein (2000*b*); Franks, D'Amour and Feller (2019); Scharfstein et al. (2021) consider a sensitivity analysis model that assumes $\mathbb{P}(Y_i(1) \mid D_i = 0, X_i) = \mathbb{P}(Y_i(1) \mid D_i = 1, X_i)\frac{\exp(\gamma_t s_t(Y_i(1)))}{C(\gamma_t; X_i)}$, where γ_t is a parameter chosen by the user and $s_t(\cdot)$ is a "tilting function" that is also specified by the user. For particular fixed choices of γ_t, $s_t(\cdot)$, such a model is sufficient to point identify various quantities of interest such as the target regression $\mu^*(x)$, the difference $\delta(x) = \mathbb{P}(Y_i(1) = 1 \mid D_i = 0, X_i) - \mathbb{P}(Y_i(1) = 1 \mid D_i = 1, X_i)$ or the predictive performance measures we consider. The literature then recommends that researchers report a sensitivity analysis that summarizes how their conclusions vary for alternative choices of γ_t or $s_t(\cdot)$. In practice, however, it may be difficult, for the user to specify domain-specific knowledge that completely summarizes the relationship between these conditional distributions. Furthermore, any particular choice of the sensitivity analysis parameter γ_t and tilting function $s_t(\cdot)$ may be mis-specified, and it is common that users only report results for a few choices. This outcome modelling approach may not encompass all possible values of the unidentified quantities that are consistent with the user's domain-knowledge.

An alternative approach places bounds on the mean difference in potential outcomes under treatment and control Luedtke, Diaz and van der Laan (2015); Díaz and van der Laan (2013); Díaz, Luedtke and van der Laan (2018). Our MOSM extends this approach by placing bounds on the *covariate-conditional* difference in means. That is, the MOSM considers all joint distributions $(X_i, D_i, Y_i(0), Y_i(1)) \sim \mathbb{P}(\cdot)$ that are consistent with the observable data and the user's specified bounds on the mean difference $\delta(x)$. This requires the user to specify intuitive domain knowledge, such as how much the probability of default can vary between accepted and rejected applicants in credit lending or how much the failure to appear rate can differ between released and detained defendants in pretrial release. Furthermore, as we showed earlier, such bounds natural arise from popular quasi-experimental methods such as instrumental variables.

9 Conclusion

This paper developed counterfactual methods for learning and evaluating statistical risk assessments that are robust to unmeasured confounding. We proposed the mean outcome sensitivity model for unobserved confounding that bounds the extent to which unmeasured confounders can affect outcomes on average in the population. Under the MOSM, we derived their sharp identified sets for the conditional likelihood of the outcome under the proposed decision, popular predictive performance metrics, and commonly-used predictive disparities are partially identified.

We solved three tasks essential to deploying counterfactual risk assessments in high-stakes settings. First, we proposed a doubly-robust learning procedure for the bounds on the conditional likelihood of the outcome under the proposed decision. Second, we translated our estimated bounds on the conditional likelihood of the outcome under the proposed decision can be translated into a robust recommendation rule. Third, we developed estimators for the bounds on the predictive performance metrics of existing statistical risk assessments based on efficient influence functions and cross-fitting.

The safe and reliable use of statistical risk assessments in high-stakes settings requires taking the violations of unconfoundedness seriously and a suite of frameworks for modeling such violations.

Providing practitioners with a range of alternative sensitivity analysis models gives them flexibility to choose the framework that is most intuitive for their own setting – some may find sensitivity analysis frameworks like the MOSM that bound how unmeasured confounders affect outcomes to be more natural, whereas others may prefer those that bound how unmeasured confounders affect historical decisions like the marginal sensitivity model. There is room for more work on proposing intuitive models for unobserved confounding, and developing the associated suite of tools needed for robust learning and evaluation of statistical risk assessments.

References

Agarwal, Alekh, Miroslav Dudík, and Zhiwei Steven Wu. 2019. "Fair Regression: Quantitative Definitions and Reduction-Based Algorithms."

Arnold, David, Will Dobbie, and Peter Hull. 2020*a*. "Measuring Racial Discrimination in Algorithms." NBER Working Paper No. 28222.

Arnold, David, Will Dobbie, and Peter Hull. 2020*b*. "Measuring Racial Discrimination in Bail Decisions." NBER Working Paper No. 26999.

Aronow, Peter M., and Donald K. K. Lee. 2013. "Interval estimation of population means under unknown but bounded probabilities of sample selection." *Biometrika*, 100(1): 235–240.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica*, 89(1): 133–161.

Australian Bureau of Statistics. 2016. "Socio-Economic Indexes for Areas (SEIFA) Technical Paper."

Balke, Alexander, and Judea Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association*, 92(439): 1171–1176.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning.* fairmlbook.org. http://www.fairmlbook.org.

Ben-Michael, Eli, D. James Greiner, Kosuke Imai, and Zhichao Jiang. 2021. "Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment." arXiv preprint, arXiv 2109.11679.

Ben-Michael, Eli, Kosuke Imai, and Zhichao Jiang. 2022. "Policy learning with asymmetric utilities." arXiv preprint, arXiv:2206.10479.

Berk, Richard. 2012. *Criminal justice forecasts of risk: a machine learning approach.* Springer Science & Business Media.

Birmingham, Jolene, Andrea Rotnitzky, and Garrett M. Fitzmaurice. 2003. "Pattern-Mixture and Selection Models for Analysing Longitudinal Data with Monotone Missing Patterns." *Journal of the Royal Statistical Society (Series B)*, 65(1): 275–297.

Blattner, Laura, and Scott T. Nelson. 2021. "How Costly is Noise?" arXiv preprint, arXiv:arXiv:2105.07554.

Brumback, Babette A, Miguel A Hernán, Sebastien JPA Haneuse, and James M Robins. 2004. "Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures." *Statistics in medicine*, 23(5): 749–767.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." 1721–1730.

Charnes, A., and W. W. Cooper. 1962. "Programming with linear fractional functionals." *Naval Research Logistics Quarterly*, 9(3-4): 181–186.

Chen, Irene Y., Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. 2020. "Probabilistic Machine Learning for Healthcare." arXiv preprint, arXiv:2009.11087.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, 21(1): C1–C68.

Choi, Edward, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." Vol. 56, 301–318.

Chouldechova, Alexandra. 2017. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big data*, 5(2): 153–163.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." Vol. 81 of *Proceedings of Machine Learning Research*, 134–148.

Coston, Amanda, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. "Counterfactual Risk Assessments, Evaluation and Fairness." *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 582—593.

Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova. 2021. "Characterizing Fairness Over the Set of Good Models Under Selective Labels." *Proceedings of the 38th International Conference on Machine Learning (ICML 2021).*

Díaz, Iván, Alexander R Luedtke, and Mark J van der Laan. 2018. "Sensitivity Analysis." In *Targeted Learning in Data Science*. 511–522. Springer.

Díaz, Iván, and Mark J van der Laan. 2013. "Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems." *The international journal of biostatistics*, 9(2): 149–160.

Dorn, Jacob, and Kevin Guo. 2021. "Sharp Sensitivity Analysis for Inverse Propensity Weighting via Quantile Balancing." arXiv preprint, arXiv:2102.04543.

Dorn, Jacob, Kevin Guo, and Nathan Kallus. 2021. "Doubly-Valid/Doubly-Sharp Sensitivity Analysis for Causal Inference with Unmeasured Confounding." arXiv preprint, arXiv:2112.11449.

Einav, Liran, Mark Jenkins, and Jonathan Levin. 2013. "The impact of credit scoring on consumer lending." *Rand Journal of Economics*, 44(2): 249—-274.

Foster, Dylan J., and Vasilis Syrgkanis. 2020. "Orthogonal Statistical Learning." arXiv preprint, arXiv:1901.09036.

Franks, Alexander, Alexander D'Amour, and Avi Feller. 2019. "Flexible sensitivity analysis for observational studies without observable implications." *Journal of the American Statistical Association.*

Franks, Alexander M., Edoardo M. Airoldi, and Donald B. Rubin. 2020. "Nonstandard conditionally specified models for nonignorable missing data." *Proceedings of the National Academy of Sciences*, 117(32): 19045–19053.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably unequal? The effects of machine learning on credit markets." *The Journal of Finance*, 77(1): 5–47.

Hand, D. J., and W. E. Henley. 1993. "Can reject inference ever work?" *IMA Journal of Management Mathematics*, 5(1): 45–55.

Hardt, Moritz, Eric Price, and Nati Srebro. 2016. "Equality of opportunity in supervised learning." 3315–3323.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. "Demystifying Statistical Learning Based on Efficient Influence Functions." *The American Statistician*, 76(3): 292–304.

Imbens, Guido, and Charles Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica*, 72(6): 1845–1857.

Jin, Ying, Zhimei Ren, and Emmanuel J. Candès. 2021. "Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach." arXiv preprint, arXiv:2111.12161.

Jin, Ying, Zhimei Ren, and Zhengyuan Zhou. 2022. "Sensitivity Analysis under the f-Sensitivity Models: Definition, Estimation and Inference." arXiv preprint, arXiv:2203.04373.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2020*a*. "Simple rules to guide expert classifications." *Journal of the Royal Statistical Society Series A*, 183(3): 771–800.

Jung, Jongbin, Ravi Shroff, Avi Feller, and Sharad Goel. 2020*b*. "Bayesian Sensitivity Analysis for Offline Policy Evaluation." *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 64–70.

Kallus, Nathan, and Angela Zhou. 2021. "Minimax-Optimal Policy Learning Under Unobserved Confounding." *Management Science*, 67(5): 2870–2890.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou. 2018. "Interval Estimation of Individual-Level Causal Effects Under Unobserved Confounding."

Kennedy, Edward H. 2022*a*. "Semiparametric doubly robust targeted double machine learning: a review." arXiv preprint, arXiv:2203.06469.

Kennedy, Edward H. 2022*b*. "Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects." arXiv preprint, arXiv:2004.14497.

Kennedy, Edward H., Sivaraman Balakrishnan, and Max G'Sell. 2020. "Sharp Instruments for Classifying Compliers and Generalizing Causal Effects." *The Annals of Statistics*, 48(4): 2008 – 2030.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance*, 34(11): 2767 – 2787.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica*, 86(2): 591–616.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

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

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Luedtke, Alexander R, Ivan Diaz, and Mark J van der Laan. 2015. "The statistics of sensitivity analyses."

Manski, Charles F. 1994. "The Selection Problem." In *Advances in Econometrics: Sixth World Congress*. Vol. 1, , ed. Christopher Sims, 143–170. Cambridge University Press.

Manski, Charles F. 2007. *Identification for Prediction and Decision.* Harvard University Press.

Miratrix, Luke W., Stefan Wager, and Jose R. Zubizarreta. 2018. "Shape-constrained partial identification of a population mean under unknown probabilities of sample selection." *Biometrika*, 105(1): 103–114.

Mishler, Alan, and Edward Kennedy. 2021. "FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes." arXiv preprint, arXiv:2109.00173.

Mishler, Alan, Edward H. Kennedy, and Alexandra Chouldechova. 2021. "Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds." *FAccT '21*, 386–400.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2019. "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions." arXiv Working Paper, arXiv:1811.07867.

Nie, Xinkun, and Stefan Wager. 2020. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika*, 108(2): 299–319.

Nie, Xinkun, Guido Imbens, and Stefan Wager. 2021. "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials across Locations." *arXiv preprint arXiv:2112.04723*.

Pu, Hongming, and Bo Zhang. 2021. "Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2): 318–345.

Qiu, Hongxiang, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C. Kessler, and Alex Luedtke. 2021. "Optimal Individualized Decision Rules Using Instrumental Variable Methods." *Journal of the American Statistical Association*, 116(533): 174–191.

Rambachan, Ashesh. 2021. "Identifying Prediction Mistakes in Observational Data." Harvard University.

Robins, James M, Andrea Rotnitzky, and Daniel O Scharfstein. 2000*a*. "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models." In *Statistical models in epidemiology, the environment, and clinical trials*. 1–94. Springer.

Robins, James M., Andrea Rotnitzky, and Daniel O. Scharfstein. 2000*b*. "Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models." 1–94. New York, NY:Springer New York.

Robins, James M., Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. 2008. "Higher order influence functions and minimax estimation of nonlinear functionals." *Probability and Statistics: Essays in Honor of David A. Freedman*, 335–421.

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

Rosenbaum, Paul R. 2002. *Observational Studies.* Springer.

Rotnitzky, Andrea, Daniel Scharfstein, Ting-Li Su, and James Robins. 2001. "Methods for Conducting Sensitivity Analysis of Trials with Potentially Nonignorable Competing Causes of Censoring." *Biometrics*, 57(1): 103–113.

Sahoo, Roshni, Lihua Lei, and Stefan Wager. 2022. "Learning from a Biased Sample." *arXiv preprint arXiv:2209.01754.*

Sansone, Dario. 2019. "Beyond Early Warning Indicators: High School Dropout and Machine Learning." *Oxford Bulletin of Economics and Statistics*, 81: 456–485.

Saxena, Devansh, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. "A Human-Centered Review of Algorithms used within the US Child Welfare System." 1–15.

Scharfstein, Daniel O, Razieh Nabi, Edward H Kennedy, Ming-Yueh Huang, Matteo Bonvini, and Marcela Smid. 2021. "Semiparametric Sensitivity Analysis: Unmeasured Confounding In Observational Studies." *arXiv preprint arXiv:2104.08300.*

Schulam, Peter, and Suchi Saria. 2017. "Reliable decision support using counterfactual models." 1697–1708.

Shalit, Uri, Fredrik D. Johansson, and David Sontag. 2017. "Estimating Individual Treatment Effect: Generalization Bounds and Algorithms." 3076–3085.

Smith, Vernon C., Adam Lange, and Daniel R Huston. 2012. "Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses." *Journal of Asynchronous Learning Networks*, 16(3): 51–61.

Stoye, Jorg. 2009. "More on Confidence Intervals for Partially Identified Parameters." *Econometrica*, 77(4): 1299–1315.

Tan, Zhiqiang. 2006. "A Distributional Approach for Causal Inference Using Propensity Scores." *Journal of the American Statistical Association*, 101(476): 1619–1637.

van der Vaart, Aad W., and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes With Applications to Statistics.* Springer Series in Statistics.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint.* Vol. 48, Cambridge University Press.

Yadlowsky, Steve, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. 2018. "Bounds on the conditional and average treatment effect with unobserved confounding factors." *arXiv preprint arXiv:1808.09521.*

Zeng, Guoping, and Qi Zhao. 2014. "A rule of thumb for reject inference in credit scoring." *Mathematical Finance Letters.*

Zhang, Bo, Jordan Weiss, Dylan S. Small, and Qingyuan Zhao. 2020. "Selecting and Ranking Individualized Treatment Rules With Unmeasured Confounding." *Journal of the American Statistical Association.*

Zhang, Yi, Eli Ben-Michael, and Kosuke Imai. 2022. "Safe Policy Learning under Regression Discontinuity Designs." arXiv preprint, arXiv:2208.13323.

Zhao, Qingyuan, Dylan S. Small, and Bhaswar B. Bhattacharya. 2019. "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4): 735–761.

Zheng, Wenjing, and Mark J. van der Laan. 2011. "Cross-Validated Targeted Minimum-Loss-Based Estimation." *Targeted Learning: Causal Inference for Observational and Experimental Data*, 459–474. New York, NY:Springer New York.

Counterfactual Risk Assessments under Unmeasured Confounding

Online Appendix

Amanda Coston   Ashesh Rambachan   Edward Kennedy

This online appendix contains proofs and additional theoretical results for the paper "Counterfactual Risk Assessments under Unmeasured Confounding" by Amanda Coston, Ashesh Rambachan and Edward Kennedy. Section A contains proofs for results stated in the main text. Section B contains auxiliary lemmas used in the proofs for results stated in the main text. Section C contains additional theoretical results for variance estimation of our overall predictive performance estimators and analyzes the bounds on predictive disparity measures under the MOSM. Section D contains additional Monte Carlo simulations.

A Omitted proofs

A.1 Section 2: the mean outcome sensitivity model

A.1.1 Proof of Proposition 2.1

Proof. For any $z \in \mathcal{Z}$, note that $\mu^*(x,z) = \mathbb{E}[Y_i(1)D_i \mid X_i = x, Z_i = z] + \mathbb{E}[Y_i(1)(1-D_i) \mid X_i = x, Z_i = z]$, where $\mathbb{E}[Y_i(1)D_i \mid X_i = x, Z_i = z] = \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z]$ and $\mathbb{E}[Y_i(1)(1-D_i) \mid X_i = x, Z_i = z] \in [0, \pi_0(x,z)]$. Therefore, $\mu^*(x,z)$ satisfies

$$\mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] \leq \mu^*(x,z) \leq \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] + \pi_0(x,z).$$

Further, since $(Y_i(0), Y_i(1)) \perp Z_i \mid X_i$, $\mu^*(x,z) = \mu^*(x)$, this in turn implies that

$$\mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] \leq \mu^*(x) \leq \mathbb{E}[Y_i D_i \mid X_i = x, Z_i = z] + \pi_0(x,z).$$

The result then follows by noting that $\mu^*(x) = \mu_1(x) + \delta(x)\pi_0(x)$ and rearranging. □

A.1.2 Proof of Lemma 2.1

Proof. The statements for $\mathcal{H}(\mu^*(x);\Delta)$ and $\mathcal{H}(\text{perf}(s;\beta);\Delta)$ follow immediately since (i) both $\mu^*(x)$ and $\text{perf}(s;\beta)$ are linear in $\delta(\cdot)$, and (ii) Δ is convex.

To prove the statement for $\text{perf}_+(s;\beta)$, we introduce the convenient shorthand

$$\text{perf}_+(s;\beta,\delta) := \frac{\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i)]}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i)]}.$$

Observe that if $\widetilde{\text{perf}}_+(s;\beta) \in \mathcal{H}(\text{perf}_+(s;\beta),\Delta)$, then there exists some $\tilde{\delta} \in \Delta$ such that $\widetilde{\text{perf}}_+(s;\beta) = \text{perf}_+(s;\beta,\tilde{\delta})$. It follows immediately that $\widetilde{\text{perf}}_+(s;\beta) \in [\underline{\text{perf}}_+(s;\beta,\Delta), \overline{\text{perf}}_+(s;\beta,\Delta)]$. All that remains to show is that every value in the interval $[\underline{\text{perf}}_+(s;\beta,\Delta), \overline{\text{perf}}_+(s;\beta,\Delta)]$ is achieved by some $\delta(\cdot) \in \Delta$.

Towards this, we apply a one-to-one change-of-variables. Let $U(\cdot)\colon \mathcal{X} \to [0,1]$ be defined as $U(x) = \frac{\delta(x) - \underline{\delta}(x)}{\overline{\delta}(x) - \underline{\delta}(x)}$. For any $\delta(\cdot) \in \Delta$, there exists $U(\cdot) \in [0,1]$ such that $\text{perf}_+(s;\beta,\delta) = \text{perf}_+(s;\beta,U)$, where

$$\text{perf}_+(s;\beta,U) := \frac{\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\underline{\delta}(X_i) + \beta_{0,i}\pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))U(X_i)]}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\underline{\delta}(X_i) + \pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))U(X_i)]}.$$

Conversely, for any $U(\cdot) \in [0,1]$, there exists a corresponding $\delta(\cdot) \in \Delta$ such that $\text{perf}_+(s;\beta,U) = \text{perf}_+(s;\beta,\delta)$, where $\delta(x) = \underline{\delta}(x) + (\overline{\delta}(x) - \underline{\delta}(x))U(x)$.

Next, apply the Charnes-Cooper transformation with

$$\tilde{V} = \frac{1}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\underline{\delta}(X_i) + \pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))U(X_i)]}$$

$$\tilde{U}(\cdot) = \frac{U(\cdot)}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\underline{\delta}(X_i) + \pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))U(X_i)]}.$$

So, for any $U(\cdot) \in [0,1]$, there exists $\tilde{V}, \tilde{U}(\cdot)$ satisfying $\tilde{U}(\cdot) \in [0, \tilde{V}]$, $\tilde{V} \geq 0$ and $\mathbb{E}[\mu_1(X_i)+\pi_0(X_i)\underline{\delta}(X_i)]\tilde{V} + \mathbb{E}[\pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))\tilde{U}(X_i)] = 1$ such that $\text{perf}_+(s;\beta,U) = \text{perf}_+(s;\beta,\tilde{U},\tilde{V})$, where

$$\text{perf}_+(s;\beta,\tilde{U},\tilde{V}) = \mathbb{E}[\beta_0(X_i)\mu_1(X_i) + (1-D_i)\beta_0(X_i)\underline{\delta}(X_i)]\tilde{V} + \mathbb{E}[\beta_0(X_i)\pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))\tilde{U}(X_i)].$$

Conversely, for any such $\tilde{U}(\cdot), \tilde{V}$, there exists $U(\cdot) \in [0,1]$ such that $\text{perf}_+(s;\beta,\tilde{U},\tilde{V}) = \text{perf}_+(s;\beta,U)$.

Now consider any $\tilde{p} \in [\underline{\text{perf}}_+(s;\beta,\Delta), \overline{\text{perf}}_+(s;\beta,\Delta)]$, which satisfies for some $\lambda \in [0,1]$

$$\tilde{p} = \lambda\underline{\text{perf}}_+(s;\beta,\Delta) + (1-\lambda)\overline{\text{perf}}_+(s;\beta,\Delta).$$

Let $\underline{\delta}(\cdot), \overline{\delta}(\cdot)$ be the functions achieving the infimum and supremum respectively

$$\underline{\delta}(\cdot) \in \arg\min_{\delta\in\Delta} \text{perf}_+(s;\beta,\delta), \;\; \overline{\delta}(\cdot) \in \arg\max_{\delta\in\Delta} \text{perf}_+(s;\beta,\delta).$$

By the change-of-variables, there exists $\underline{\tilde{V}}, \underline{\tilde{U}}(\cdot)$ and $\overline{\tilde{V}}, \overline{\tilde{U}}(\cdot)$ such that

$$\underline{\text{perf}}_+(s;\beta,\Delta) = \text{perf}_+(s;\beta,\underline{\tilde{U}}(\cdot),\underline{\tilde{V}}), \;\; \overline{\text{perf}}_+(s;\beta,\Delta) = \text{perf}_+(s;\beta,\overline{\tilde{U}}(\cdot),\overline{\tilde{V}}).$$

Therefore, $\tilde{p} = \lambda\text{perf}_+(s;\beta,\underline{\tilde{U}}(\cdot),\underline{\tilde{V}}) + (1-\lambda)\text{perf}_+(s;\beta,\overline{\tilde{U}}(\cdot),\overline{\tilde{V}})$. Since $\text{perf}_+(s;\beta,\tilde{U},\tilde{V})$ is linear in $\tilde{U}, \tilde{V}$, we also have that

$$\tilde{p} = \text{perf}_+(s;\beta,\lambda\underline{\tilde{U}} + (1-\lambda)\overline{\tilde{U}}, \lambda\underline{\tilde{V}} + (1-\lambda)\overline{\tilde{V}}).$$

We can therefore apply the change-of-variables in the other direction to construct the corresponding $\tilde{\delta}(\cdot) \in \Delta$, which satisfies $\tilde{p} = \text{perf}_+(s;\beta,\tilde{\delta})$ by construction. ☐

A.2 Section 3: bounding the target regression under the outcome sensitivity model

A.2.1 Proof of Theorem 3.1

Proof. We prove this result for the estimator of the upper-bound, and the same argument applies to the estimator of the lower-bound. Observe that

$$\begin{aligned}
\|\widehat{\overline{\mu}}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\| &\leq \|\widehat{\overline{\mu}}(\cdot;\Delta) - \widehat{\overline{\mu}}_{oracle}(\cdot;\Delta)\| + \|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\| \\
&\leq \|\widehat{\overline{\mu}}(\cdot;\Delta) - \widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \tilde{b}(\cdot)\| + \|\tilde{b}(\cdot)\| + \|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \overline{\mu}^*(\cdot;\Delta)\|
\end{aligned}$$

for $\tilde{b}(x) = \widehat{\mathbb{E}}_n[\hat{b}(X_i) \mid X_i = x]$ is the smoothed bias and $\hat{b}(x) = \mathbb{E}[\phi_1(Y_i;\hat{\eta}) - \phi_1(Y_i;\eta) \mid \mathcal{O}_1, X_i = x]$ is the conditional bias of the estimated pseudo-outcome. Under Assumption B.1, Lemma B.1 implies that $\|\widehat{\overline{\mu}}(\cdot;\Delta) - \widehat{\overline{\mu}}_{oracle}(\cdot;\Delta) - \tilde{b}(\cdot)\| = o_{\mathbb{P}}(R_{oracle})$. Furthermore,

$$\begin{aligned}
\hat{b}(x)^2 &= \left\{\frac{\pi_1(x) - \hat{\pi}_1(x)}{\hat{\pi}_1(x)}(\mu_1(x) - \hat{\mu}_1(x))\right\}^2 \\
&\leq \frac{1}{\epsilon^2}\left\{(\pi_1(x) - \hat{\pi}_1(x))(\mu_1(x) - \hat{\mu}_1(x))\right\}^2,
\end{aligned}$$

where the first equality applies iterated expectations, and the second applies the assumption of bounded propensity score. Putting this together yields

$$\|\widehat{\overline{\mu}}(\cdot;\Delta)-\overline{\mu}^*(\cdot;\Delta)\| \leq \|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta)-\overline{\mu}^*(\cdot;\Delta)\|+\epsilon^{-1}\|\tilde{R}(\cdot)\|+o_{\mathbb{P}}(R_{oracle})$$

as desired. □

A.2.2 Proof of Proposition 3.1

Proof. We prove the result for the DR-Learner of the upper bound, and the same argument applies for the DR-Learner of the lower bound. Following the proof of Theorem 3.1, we arrive at

$$\|\widehat{\overline{\mu}}(\cdot;\Delta(\Gamma))-\overline{\mu}^*(\cdot;\Delta(\Gamma))\| \leq \|\widehat{\overline{\mu}}(\cdot;\Delta(\Gamma))-\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(\Gamma))-\tilde{b}(\cdot)\|+\|\tilde{b}(\cdot)\|+\|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(\Gamma))-\overline{\mu}^*(\cdot;\Delta(\Gamma))\|,$$

now for $\tilde{b}(x)=\mathbb{E}_n[\hat{b}(X_i)\mid X_i=x]$ and

$$\hat{b}(x)=\underbrace{\mathbb{E}[\phi_1(Y_i;\hat{\eta})-\phi_1(Y_i;\eta)\mid \mathcal{O}_1, X_i=x]}_{(a)}+ \\ (\overline{\Gamma}-1)\underbrace{\mathbb{E}[\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta})-\phi(\pi_0(X_i)\mu_1(X_i);\eta)\mid \mathcal{O}_1, X_i=x]}_{(b)}.$$

Given Assumption B.1, $\|\widehat{\overline{\mu}}(\cdot;\Delta(\Gamma))-\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(\Gamma))-\tilde{b}(\cdot)\|=o_{\mathbb{P}}(R_{oracle})$ by Lemma B.1. Furthermore, $\hat{b}(x)^2\leq 2(a)^2+2(\bar{\Gamma}-1)^2(b)^2$, where

$$(a)^2\leq\frac{1}{\epsilon^2}\left\{\left(\hat{\pi}_1(x)-\pi_1(x)\right)\left(\hat{\mu}_1(x)-\mu_1(x)\right)\right\}^2$$

by the proof of Theorem 3.1, and

$$(b)^2=\left\{(\pi_0(x)-\hat{\pi}_0(x))\hat{\mu}_1(x)+\frac{\pi_1(x)}{\hat{\pi}_1(x)}\left(\mu_1(x)-\hat{\mu}_1(x)\right)\hat{\pi}_0(x)+\hat{\pi}_0(x)\hat{\mu}_1(x)-\pi_0(x)\mu_1(x)\right\}^2=$$

$$\left\{(\pi_0(x)-\hat{\pi}_0(x))\hat{\mu}_1(x)+\frac{\pi_1(x)}{\hat{\pi}_1(x)}\left(\mu_1(x)-\hat{\mu}_1(x)\right)\hat{\pi}_0(x)+\hat{\pi}_0(x)(\hat{\mu}_1(x)-\mu_1(x))+\mu_1(x)(\hat{\pi}_0(x)-\pi_0(x))\right\}^2=$$

$$\left\{(\pi_0(x)-\hat{\pi}_0(x))(\hat{\mu}_1(x)-\mu_1(x))+\frac{\hat{\pi}_0(x)}{\hat{\pi}_1(x)}(\pi_1(x)-\hat{\pi}_1(x))(\mu_1(x)-\hat{\mu}_1(x))\right\}^2=$$

$$\left\{(\pi_1(x)-\hat{\pi}_1(x))(\mu_1(x)-\hat{\mu}_1(x))+\frac{\hat{\pi}_0(x)}{\hat{\pi}_1(x)}(\pi_1(x)-\hat{\pi}_1(x))(\mu_1(x)-\hat{\mu}_1(x))\right\}^2\leq\frac{1}{\epsilon^2}\left\{(\pi_1(x)-\hat{\pi}_1(x))(\hat{\mu}_1(x)-\mu_1(x))\right\}^2$$

by iterated expectations and the assumption of bounded propensity score. Putting this together then yields

$$\|\widehat{\overline{\mu}}(\cdot;\Delta(\Gamma))-\overline{\mu}^*(\cdot;\Delta(\Gamma))\| \leq \|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(\Gamma))-\overline{\mu}^*(\cdot;\Delta(\Gamma))\|+\epsilon^{-1}\|\tilde{R}(\cdot)\|+o_{\mathbb{P}}(R_{oracle})$$

as desired. □

A.2.3 Proof of Proposition 3.2

Proof. We prove the result for the DR-Learner of the upper bound, and the same argument applies for the DR-Learner of the lower bound. To ease notation, write $\mu_z^{DY}(x)=\mathbb{E}[D_iY_i\mid Z_i=z, X_i=x]$, $\lambda_z(x)=\mathbb{P}(Z_i=x\mid X_i=x)$, and $\widehat{\overline{\mu}}(O_i;\Delta(z))=\phi_z(1-D_i;\hat{\eta})+\phi_z(D_iY_i;\hat{\eta})$. Following the proof of

Theorem 1, we arrive at

$$\|\widehat{\overline{\mu}}(\cdot;\Delta(z))-\overline{\mu}^*(\cdot;\Delta(z))\| \le \|\widehat{\overline{\mu}}(\cdot;\Delta(z))-\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(z))-\tilde{b}(\cdot)\|+\|\tilde{b}(\cdot)\|+\|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(z))-\overline{\mu}^*(\cdot;\Delta(z))\|,$$

where $\tilde{b}(x) = \mathbb{E}_n[\hat{b}(x) \mid X_i = x]$ and

$$\hat{b}(x) = \underbrace{\mathbb{E}[\phi_z(1-D_i;\hat{\eta}) - \phi_z(1-D_i;\eta) \mid X_i = x, \mathcal{O}_1]}_{(a)} + \underbrace{\mathbb{E}[\phi_z(D_iY_i;\hat{\eta}) - \phi_z(D_iY_i;\eta) \mid X_i = x, \mathcal{O}_1]}_{(b)}.$$

Given Assumption B.1, $\|\widehat{\overline{\mu}}(\cdot;\Delta(z)) - \widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(z)) - \tilde{b}(\cdot)\| = o_{\mathbb{P}}(R_{oracle}(z))$ by Lemma B.1. Furthermore, $\hat{b}(x)^2 \le 2(a)^2 + 2(b)^2$, where

$$(a)^2 = \left\{\frac{\lambda_z(x)}{\hat{\lambda}_z(x)}\left(\pi_0(x,z) - \hat{\pi}_0(x,z)\right) + \left(\hat{\pi}_0(x,z) - \pi_0(x,z)\right)\right\}^2 \le \frac{1}{\epsilon^2}\left\{(\lambda_z(x) - \hat{\lambda}_z(x))(\pi_0(x,z) - \hat{\pi}_0(x,z))\right\}^2$$

and

$$(b) = \frac{\lambda_z(x)}{\hat{\lambda}_z(x)}(\mu_z^{DY}(x) - \hat{\mu}_z^{DY}(x)) + (\hat{\mu}_z^{DY}(x) - \mu_z^{DY}(x)) \le \frac{1}{\epsilon^2}\left\{(\lambda_z(x) - \hat{\lambda}_z(x))(\mu_z^{DY}(x) - \hat{\mu}_z^{DY}(x))\right\}^2$$

by iterated expectations and bounded instrument propensity. Putting this together then yields

$$\|\widehat{\overline{\mu}}(\cdot;\Delta(z)) - \overline{\mu}^*(\cdot;\Delta(z))\| \le \|\widehat{\overline{\mu}}_{oracle}(\cdot;\Delta(z)) - \overline{\mu}^*(\cdot;\Delta(z))\| + \epsilon^{-1}\|\tilde{R}_1(\cdot)\| + \epsilon^{-1}\|\tilde{R}_2(\cdot)\| + o_{\mathbb{P}}(R_{oracle}(z))$$

as desired. □

A.3 Section 4: robust recommendations under the outcome sensitivity model

A.3.1 Proof of Lemma 4.1

Proof. At each value $x \in \mathcal{X}$, notice that if $d(x) = 1$, then

$$(-u_{1,1,i}\mu^*(x)+u_{1,0,i}(1-\mu^*(x)))d(x)+(-u_{0,0,i}(1-\mu^*(x))+u_{0,1,i}\mu^*(x))(1-d(x)) = u_{1,0,i}-(u_{1,1,i}+u_{1,0,i})\mu^*(x).$$

This is minimized over $\mu^*(x) \in \mathcal{H}(\mu^*(x);\Delta)$ at $\mu^*(x) = \overline{\mu}^*(x;\Delta)$. If $d(x) = 0$, then

$$(-u_{1,1,i}\mu^*(x)+u_{1,0,i}(1-\mu^*(x)))d(x)+(-u_{0,0,i}(1-\mu^*(x))+u_{0,1,i}\mu^*(x))(1-d(x)) = -u_{0,0,i}+(u_{0,0,i}+u_{0,1,i})\mu^*(x).$$

This is minimized over $\mu^*(x) \in \mathcal{H}(\mu^*(x);\Delta)$ at $\mu^*(x) = \underline{\mu}^*(x;\Delta)$. The result for the lower bound immediately follows. The result for the upper bound follows by an analogous argument. □

A.3.2 Proof of Lemma 4.2

Proof. This follows directly from Lemma 4.1 and the characterization of $\underline{U}(d;\Delta)$ for any $d(\cdot)\colon \mathcal{X} \to \{0,1\}$. Recall

$$\underline{U}(d;\Delta) := \mathbb{E}[(u_{1,0,i} - (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(x))\, d(X_i) + \left(-u_{0,0,i} + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(x)\right)(1 - d(X_i))].$$

Therefore, at any $x \in \mathcal{X}$, it is optimal to set $d^*(x) = 1$ if

$$u_{1,0,i} - (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(x;\Delta) \ge -u_{0,0,i} + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(x;\Delta),$$

or equivalently

$$u_{1,0,i} + u_{0,0,i} \ge (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(x;\Delta) + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(x;\Delta).$$

Analogously, it is optimal to set $d^*(x) = 0$ if

$$u_{1,0,i} + u_{0,0,i} < (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(x;\Delta) + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(x;\Delta).$$

□

A.3.3 Proof of Theorem 4.1

Proof. Recall from Lemma 4.1 that, for any decision rule $d(\cdot)\colon \mathcal{X} \to \{0,1\}$,

$$\underline{U}(d;\Delta) := \mathbb{E}[(u_{1,0,i} - (u_{1,1,i} + u_{1,0,i})\overline{\mu}^*(x))\, d(X_i) + \left(-u_{0,0,i} + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(x)\right)(1 - d(X_i))] =$$

$$\mathbb{E}[-u_{0,0,i} + (u_{0,0,i} + u_{0,1,i})\underline{\mu}^*(X_i;\Delta)] + \mathbb{E}[((u_{1,0,i} + u_{0,0,i}) - \tilde{\mu}^*(x;\Delta))\, d(X_i)].$$

Therefore, we can rewrite regret as

$$R(\hat{d};\Delta) = \underline{U}(d^*;\Delta) - \underline{U}(\hat{d};\Delta) =$$

$$\mathbb{E}[(c(X_i) - \tilde{\mu}^*(X_i;\Delta))\,(d^*(X_i) - \hat{d}(X_i)],$$

where we defined the shorthand notation $c(X_i) = u_{1,0}(X_i) + u_{0,0}(X_i)$. It then follows that

$$R(\hat{d};\Delta) = \int_{x\in\mathcal{X}} (c(x) - \tilde{\mu}^*(x;\Delta))\left(d^*(x;\Delta) - \hat{d}(x;\Delta)\right) dP(x) \le$$

$$\int_{x\in\mathcal{X}} |\tilde{\mu}^*(x;\Delta) - c(x)|\, 1\{d^*(x;\Delta) \ne \hat{d}(x;\Delta)\} dP(x).$$

Furthermore, at any fixed $X_i = x$, $\hat{d}(X_i) \ne d^*(X_i)$ implies that $|\tilde{\mu}^*(x) - \widehat{\tilde{\mu}}(x)| \ge |\tilde{\mu}^*(x) - c(x)|$. Combining this with the previous display implies that

$$R(\hat{d};\Delta) \le \int_{x\in\mathcal{X}} |\tilde{\mu}^*(x) - \widehat{\tilde{\mu}}(x)| dP(x).$$

Substituting in the definition of $\tilde{\mu}^*(x)$ and $\widehat{\tilde{\mu}}(x)$, we have

$$|\tilde{\mu}^*(x) - \widehat{\tilde{\mu}}(x)| =$$

$$|(u_{1,1}(x)+u_{1,0}(x))\overline{\mu}^*(x;\Delta)+(u_{0,0}(x)+u_{0,1}(x))\underline{\mu}^*(x;\Delta)-(u_{1,1}(x)+u_{1,0}(x))\widehat{\overline{\mu}}(x;\Delta)-(u_{0,0}(x)+u_{0,1}(x))\widehat{\underline{\mu}}(x;\Delta)| \le$$

$$|\overline{\mu}^*(x;\Delta) - \widehat{\overline{\mu}}(x;\Delta)| + |\underline{\mu}^*(x;\Delta) - \widehat{\underline{\mu}}(x;\Delta)|,$$

which follows by the triangle inequality and using $u_{0,0}(x), u_{0,1}(x), u_{1,0}(x), u_{1,1}(x)$ are non-negative and sum to one. Substituting back into the bound on $R(\hat{d};\Delta)$ then delivers

$$R(\hat{d};\Delta) \le \int_{x\in\mathcal{X}} |\overline{\mu}^*(x;\Delta) - \widehat{\overline{\mu}}(x;\Delta)| dP(x) + \int_{x\in\mathcal{X}} |\underline{\mu}^*(x;\Delta) - \widehat{\underline{\mu}}(x;\Delta)| dP(x) =$$

$$\|\overline{\mu}^*(x;\Delta) - \widehat{\overline{\mu}}(x;\Delta)\|_1 + \|\underline{\mu}^*(x;\Delta) - \widehat{\underline{\mu}}(x;\Delta)\|_1.$$

Therefore, using the Cauchy-Schwarz inequality $\|\overline{\mu}^*(x;\Delta) - \widehat{\overline{\mu}}(x;\Delta)\|_1^2 \le \|\overline{\mu}^*(x;\Delta) - \widehat{\overline{\mu}}(x;\Delta)\|_2^2$ and $\|\underline{\mu}^*(x;\Delta) - \widehat{\underline{\mu}}(x;\Delta)\|_1^2 \le \|\underline{\mu}^*(x;\Delta) - \widehat{\underline{\mu}}(x;\Delta)\|_2^2$ and the inequality $(a+b)^2 \le 2(a^2+b^2)$,

$$R(\hat{d};\Delta)^2 \le 2\|\overline{\mu}^*(x;\Delta) - \widehat{\overline{\mu}}(x;\Delta)\|_2^2 + 2\|\underline{\mu}^*(x;\Delta) - \widehat{\underline{\mu}}(x;\Delta)\|_2^2.$$

The result then follows by applying Theorem 3.1. □

A.4 Section 5: robust audits under the outcome sensitivity model

A.4.1 Proof of Theorem 5.1

Proof. To prove the first claim, consider our proposed estimator of the upper bound on overall predictive performance $\widehat{\overline{\text{perf}}}(s;\beta,\Delta)$. To ease notation, let

$$\begin{aligned}
\overline{\text{perf}}_i &= \beta_{0,i} + \beta_{1,i}(1-D_i)\left(1\{\beta_{1,i}>0\}\overline{\delta}_i + 1\{\beta_{1,i}\le 0\}\underline{\delta}_i\right) + \beta_{1,i}\phi_1(Y_i;\eta) \\
\widehat{\overline{\text{perf}}}_i &= \beta_{0,i} + \beta_{1,i}(1-D_i)\left(1\{\beta_{1,i}>0\}\overline{\delta}_i + 1\{\beta_{1,i}\le 0\}\underline{\delta}_i\right) + \beta_{1,i}\phi_1(Y_i;\hat{\eta}_{-K_i}).
\end{aligned}$$

Note that we can write $\overline{\text{perf}}(s;\beta,\Delta) = \mathbb{E}[\overline{\text{perf}}_i]$, where we used that $\mathbb{E}[\mu_1(X_i)] = \mathbb{E}[\phi_1(Y_i;\eta)]$ by iterated expectations. Therefore, $|\widehat{\overline{\text{perf}}}(s;\beta,\Delta) - \overline{\text{perf}}(s;\beta,\Delta)|$ equals

$$\left|\mathbb{E}_n[\widehat{\overline{\text{perf}}}_i] - \mathbb{E}[\overline{\text{perf}}_i]\right| \le \left|\underbrace{\left(\mathbb{E}_n[\overline{\text{perf}}_i] - \mathbb{E}[\overline{\text{perf}}_i]\right)}_{(a)}\right| + \left|\underbrace{\mathbb{E}_n\left[\left(\widehat{\overline{\text{perf}}}_i - \overline{\text{perf}}_i\right)\right]}_{(b)}\right|.$$

By Chebyshev's inequality, (a) is $O_{\mathbb{P}}(1/\sqrt{n})$. Next, recall we can rewrite (b) as

$$|\mathbb{E}_n\left[\left(\widehat{\overline{\text{perf}}}_i - \overline{\text{perf}}_i\right)\right]| = |\sum_{k=1}^{K}\mathbb{E}_n[1\{K_i=k\}]\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i]| \le \sum_{k=1}^{K}|\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i]|.$$

We will show that each term in the sum is $O_{\mathbb{P}}(R_{1,n}^k + R_{1,n}^k/\sqrt{n})$. For any k, observe that

$$\left|\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i]\right| \le |\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i] - \mathbb{E}[\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i \mid \mathcal{O}_{-k}]| + |\mathbb{E}[\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i \mid \mathcal{O}_{-k}]|,$$

where $\widehat{\overline{\text{perf}}}_{i,-k} - \overline{\text{perf}}_i = \beta_{i,1}(\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta))$. The first term on the right hand side of the previous display is therefore $O_{\mathbb{P}}(R_{1,n}^k/\sqrt{n})$ by by Lemma B.4 and Lemma B.5. The second term on the right hand side of the previous display is $O_{\mathbb{P}}(R_{1,n}^k)$ by Lemma B.2. Putting this together, we have shown the first claim

$$\left|\widehat{\overline{\text{perf}}}(s;\beta,\Delta) - \overline{\text{perf}}(s;\beta,\Delta)\right| = O_{\mathbb{P}}\left(1/\sqrt{n} + \sum_{k=1}^{K}R_{1,n}^k + \sum_{k=1}^{K}R_{1,n}^k/\sqrt{n}\right).$$

The result for $\widehat{\underline{\text{perf}}}(s;\beta,\Delta)$ follows the same argument. The second claim follows by noticing that the proof of the first claim showed that

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{\text{perf}}}(s;\beta,\Delta)\\ \widehat{\underline{\text{perf}}}(s;\beta,\Delta)\end{pmatrix} - \begin{pmatrix}\overline{\text{perf}}(s;\beta,\Delta)\\ \underline{\text{perf}}(s;\beta,\Delta)\end{pmatrix}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\overline{\text{perf}}_i - \mathbb{E}[\overline{\text{perf}}_i]\\ \underline{\text{perf}}_i - \mathbb{E}[\underline{\text{perf}}_i]\end{pmatrix} + o_{\mathbb{P}}(1)$$

if $R_{1,n} = o_{\mathbb{P}}(1/\sqrt{n})$. By the central limit theorem,

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{\text{perf}}}(s;\beta,\Delta)\\ \widehat{\underline{\text{perf}}}(s;\beta,\Delta)\end{pmatrix} - \begin{pmatrix}\overline{\text{perf}}(s;\beta,\Delta)\\ \underline{\text{perf}}(s;\beta,\Delta)\end{pmatrix}\right) \xrightarrow{d} N\left(0, Cov\left(\begin{pmatrix}\overline{\text{perf}}_i\\ \underline{\text{perf}}_i\end{pmatrix}\right)\right),$$

from which the result follows. □

A.4.2 Proof of Proposition 5.1

Proof. To prove the first claim, consider our proposed estimator of the upper bound $\widehat{\overline{\text{perf}}}(s; \beta, \Delta(\Gamma))$. Now let

$$\overline{\text{perf}}_i^\Gamma = \beta_{0,i} + \beta_{1,i}\phi_1(Y_i;\eta) + \beta_{1,i}\left(1\{\beta_{1,i} > 0\}(\overline{\Gamma} - 1) + 1\{\beta_{1,i} \le 0\}(\underline{\Gamma} - 1)\right)\phi(\pi_0(X_i)\mu_1(X_i);\eta)$$
$$\widehat{\overline{\text{perf}}}_i^\Gamma = \beta_{0,i} + \beta_{1,i}\phi_1(Y_i;\hat{\eta}_{-K_i}) + \beta_{1,i}\left(1\{\beta_{1,i} > 0\}(\overline{\Gamma} - 1) + 1\{\beta_{1,i} \le 0\}(\underline{\Gamma} - 1)\right)\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-K_i}).$$

Observe $|\widehat{\overline{\text{perf}}}(s; \beta, \Delta(\Gamma)) - \overline{\text{perf}}(s; \beta, \Gamma)|$ equals

$$|\mathbb{E}_n[\widehat{\overline{\text{perf}}}_i^\Gamma] - \mathbb{E}[\overline{\text{perf}}_i^\Gamma]| \le |\underbrace{\mathbb{E}_n[\overline{\text{perf}}_i^\Gamma] - \mathbb{E}[\overline{\text{perf}}_i^\Gamma]}_{(a)}| + |\underbrace{\mathbb{E}_n\left[\left(\widehat{\overline{\text{perf}}}_i^\Gamma - \overline{\text{perf}}_i^\Gamma\right)\right]}_{(b)}|.$$

As in the proof of Theorem 5.1, (a) is $O_{\mathbb{P}}(1/\sqrt{n})$. Next, we can further rewrite (b) as

$$|\mathbb{E}_n\left[\left(\widehat{\overline{\text{perf}}}_i^\Gamma - \overline{\text{perf}}_i^\Gamma\right)\right]| = |\sum_{k=1}^K \mathbb{E}_n[1\{K_i = k\}]\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma] \le \sum_{k=1}^K |\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma]|.$$

We will again show that each term in the sum is $O_{\mathbb{P}}(R_{1,n}^k + R_{1,n}^k/\sqrt{n})$. Observe that

$$\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma]| \le$$
$$|\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma] - \mathbb{E}[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma \mid \mathcal{O}_{-k}]| + |\mathbb{E}[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma \mid \mathcal{O}_k]|,$$

where

$$\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma = \beta_{1,i}(\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta)) + \tilde{\beta}_{1,i}\left(\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k}) - \phi(\pi_0(X_i)\mu_1(X_i);\eta)\right)$$

for $\tilde{\beta}_i = \beta_{1,i}\left(1\{\beta_{1,i} > 0\}(\overline{\Gamma} - 1) + 1\{\beta_{1,i} \le 0\}(\underline{\Gamma} - 1)\right)$. So $|\mathbb{E}_n^k[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma] - \mathbb{E}[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma \mid \mathcal{O}_{-k}]|$ is bounded by

$$\underbrace{|\mathbb{E}_n^k[\beta_{1,i}(\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta))] - \mathbb{E}[\beta_{1,i}(\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta)) \mid \mathcal{O}_{-k}]|}_{(c)} +$$

$$\underbrace{|\mathbb{E}_n^k[\tilde{\beta}_{1,i}(\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k}) - \phi(\pi_0(X_i)\mu_1(X_i);\eta))] - \mathbb{E}[\tilde{\beta}_{1,i}(\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k}) - \phi(\pi_0(X_i)\mu_1(X_i);\eta)) \mid \mathcal{O}_{-k}]|}_{(d)},$$

where (c) is $O_{\mathbb{P}}(R_{1,n}^k/\sqrt{n})$ by Lemma B.4 and Lemma B.5, and (d) is also $O_{\mathbb{P}}(R_{1,n}^k/\sqrt{n})$ Lemma B.4 and Lemma B.6. The second term $\mathbb{E}[\widehat{\overline{\text{perf}}}_{i,-k}^\Gamma - \overline{\text{perf}}_i^\Gamma \mid \mathcal{O}_k]|$ is bounded by

$$\underbrace{|\mathbb{E}[\beta_{1,i}(\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta)) \mid \mathcal{O}_{-k}]|}_{(e)} + |\underbrace{\mathbb{E}[\tilde{\beta}_{1,i}(\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k}) - \phi(\pi_0(X_i)\mu_1(X_i);\eta)) \mid \mathcal{O}_{-k}]}_{(f)}|,$$

where (e) is $O_{\mathbb{P}}(R_{1,n}^k)$ by Lemma B.2 and (f) is $O_{\mathbb{P}}(R_{1,n}^k)$ by Lemma B.7. The first result then follows as in the proof of Theorem 5.1. The second result also follows as in the proof of Theorem 5.1, where the asymptotic variance matrix $\Sigma(\Gamma)$ is now defined as $Cov((\overline{\text{perf}}_i^\Gamma, \underline{\text{perf}}_i^\Gamma)')$. □

A.4.3 Proof of Proposition 5.2

Proof. To prove the first claim, begin by our considering our proposed estimator $\widehat{\overline{\text{perf}}}(s;\beta,\Delta(z))$. To ease notation, let $\phi(\overline{\delta}_z(X_i);\eta) = \overline{\phi}(X_i;\eta)$ and $\phi(\underline{\delta}_z(X_i);\eta) = \underline{\phi}(X_i;\eta)$. Further define

$$\overline{\text{perf}}_i^z = \beta_{0,i} + \beta_{1,i}\phi_1(Y_i;\eta) + \beta_{1,i}1\{\beta_{1,i} > 0\}\overline{\phi}(X_i;\eta) + \beta_{1,i}1\{\beta_{1,i} \le 0\}\underline{\phi}(X_i;\eta),$$

$$\widehat{\overline{\text{perf}}}_i^z = \beta_{0,i} + \beta_{1,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \beta_{1,i}1\{\beta_{1,i} > 0\}\overline{\phi}(X_i;\hat{\eta}_{-K_i}) + \beta_{1,i}1\{\beta_{1,i} \le 0\}\underline{\phi}(X_i;\hat{\eta}_{-K_i}).$$

Observe that $|\widehat{\overline{\text{perf}}}(s;\beta,\Delta(z)) - \overline{\text{perf}}(s;\beta,\Delta(z))|$ equals

$$|\mathbb{E}_n[\widehat{\overline{\text{perf}}}_i^z] - \mathbb{E}[\overline{\text{perf}}_i^z]| \le \underbrace{|\mathbb{E}_n[\overline{\text{perf}}_i^z] - \mathbb{E}[\overline{\text{perf}}_i^z]|}_{(a)} + \underbrace{|\mathbb{E}_n[\widehat{\overline{\text{perf}}}_i^z - \overline{\text{perf}}_i^z]|}_{(b)}.$$

The proof then follows the same steps as the proof of Proposition 5.1, except invoking Lemma B.9 and Lemma B.8. The second claim then follows by noticing that the proof of the first claim established that

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{\text{perf}}}(s;\beta,\Delta(z)) \\ \widehat{\underline{\text{perf}}}(s;\beta,\Delta(z))\end{pmatrix} - \begin{pmatrix}\overline{\text{perf}}(s;\beta,\Delta(z)) \\ \underline{\text{perf}}(s;\beta,\Delta(z))\end{pmatrix}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\overline{\text{perf}}_i^z - \mathbb{E}[\overline{\text{perf}}_i^z] \\ \underline{\text{perf}}_i^z - \mathbb{E}[\underline{\text{perf}}_i^z]\end{pmatrix} + o_{\mathbb{P}}(1)$$

if $R_{1,n} = o_{\mathbb{P}}(1/\sqrt{n})$, $R_{2,n} = o_{\mathbb{P}}(1/\sqrt{n})$, and $R_{3,n} = o_{\mathbb{P}}(1/\sqrt{n})$ and applying the central limit theorem. The asymptotic variance matrix is defined as $\Sigma(z) = Cov((\overline{\text{perf}}_i^z, \underline{\text{perf}}_i^z)')$. □

A.4.4 Proof of Lemma 5.1

Proof. We first use the change-of-variables $\delta(X_i) = \underline{\delta}(X_i) + (\overline{\delta}(X_i) - \underline{\delta}(X_i))U_i$ for $U_i \in [0,1]$ to rewrite $\widehat{\text{perf}}_{+}^k(s;\beta,\Delta)$ as

$$\begin{aligned}\widehat{\text{perf}}_{+}^k(s;\beta,\Delta_n) := \max_U & \frac{\mathbb{E}_n^k[\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\underline{\delta}_i + \beta_{0,i}(1-D_i)(\overline{\delta}_i - \underline{\delta}_i)U_i]}{\mathbb{E}_n^k[\phi_1(Y_i;\hat{\eta}) + (1-D_i)\underline{\delta}_i + (1-D_i)(\overline{\delta}_i - \underline{\delta}_i)U_i]} \\ & \text{s.t. } 0 \le U_i \le 1 \text{ for } i = 1,\dots,n_k,\end{aligned}$$

where $U = (U_1,\dots,U_n)'$.

Define $\hat{c}^k = \mathbb{E}_n^k[\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\underline{\delta}_i]$, $\hat{d} = \mathbb{E}_n^k[\phi_1(Y_i;\hat{\eta}) + (1-D_i)\underline{\delta}_i]$, $\hat{\alpha}_i := \beta_{0,i}(1-D_i)(\overline{\delta}_i - \underline{\delta}_i)$, $\hat{\gamma}_i := (1-D_i)(\overline{\delta}_i - \underline{\delta}_i)$. We can further rewrite the estimator as

$$\widehat{\text{perf}}_{+}^k(s;\beta,\Delta_n) = \max_U \frac{\hat{\alpha}'U + \hat{c}^k}{\hat{\gamma}'U + \hat{d}^k} \text{ s.t. } 0 \le U_i \le 1 \text{ for } i = 1,\dots,n_k,$$

where $\hat{\alpha} = (\hat{\alpha}_1,\dots,\hat{\alpha}_n)'$, $\hat{\gamma} = (\hat{\gamma}_1,\dots,\hat{\gamma}_n)'$. Applying the Charnes-Cooper transformation with $\tilde{U} = \frac{U}{\hat{\gamma}'U + \hat{d}^k}$, $\tilde{V} = \frac{1}{\hat{\gamma}'U + \hat{d}^k}$, this linear-fractional program is equivalent to the linear program

$$\begin{aligned}\widehat{\text{perf}}_{+}^k(s;\beta,\Delta_n) = \max_{\tilde{U},\tilde{V}} & \hat{\alpha}'\tilde{U} + \hat{c}^k\tilde{V} \\ & \text{s.t. } 0 \le \tilde{U}_i \le \tilde{V} \text{ for } i = 1,\dots n_k, \\ & 0 \le \tilde{V},\ \hat{\gamma}'\tilde{U} + \tilde{V}\hat{d}^k = 1.\end{aligned}$$

□

A.4.5 Proof of Lemma 5.2

Proof. We first show this result for the fold-specific estimator $\widehat{\text{perf}}^k_+(s;\beta,\Delta)$ by using the proof strategy of Proposition 2 in Kallus and Zhou (2021) By Lemma 5.1, recall

$$\begin{aligned}\widehat{\text{perf}}^k_+(s;\beta,\Delta) = \max_{\tilde{U},\tilde{V}} & \hat{\alpha}'\tilde{U} + \hat{c}^k\tilde{V} \\ \text{s.t. } & 0 \le \tilde{U}_i \le \tilde{V} \text{ for } i = 1,\ldots n_k, \\ & 0 \le \tilde{V},\ \hat{\gamma}'\tilde{U} + \tilde{V}\hat{d}^k = 1.\end{aligned}$$

Next, define the dual program associated with this primal linear program. Let P_i be the dual variables associated with the constraints $\tilde{U}_i \le \tilde{V}$, Q_i be the dual variables associated with the constraints $\tilde{U}_i \ge 0$, and λ be the dual variable associated with the constraint $\gamma'\tilde{U} + \tilde{V}d^k = 1$. The dual linear program is

$$\begin{aligned}\min_{\lambda,P,Q} & \lambda \\ \text{s.t. } & P_i - Q_i + \lambda\hat{\gamma}_i = \hat{\alpha}_i, \\ & -\mathbf{1}'P + \lambda\hat{d}^k \ge \hat{c}^k, \\ & P_i \ge 0, Q_i \ge 0 \text{ for } i = 1,\ldots,n,\end{aligned}$$

where $\mathbf{1}$ is the vector of all ones of appropriate dimension. By re-arranging the first constraint and substituting in the expressions for $\hat{\alpha}_i, \hat{\gamma}_i$, we observe that

$$P_i - Q_i = (\beta_0 - \lambda)(1 - D_i)(\overline{\delta}_i - \underline{\delta}_i).$$

By complementary slackness, at most only one of P_i or Q_i will be non-zero at the optimum, and so combined with the previous display, this implies

$$\begin{aligned}P_i &= \max\{\beta_{0,i} - \lambda, 0\}(1 - D_i)(\overline{\delta}_i - \underline{\delta}_i), \\ Q_i &= \max\{\lambda - \beta_{0,i}, 0\}(1 - D_i)(\overline{\delta}_i - \underline{\delta}_i).\end{aligned}$$

Next, notice that the constraint $-\mathbf{1}'P + \lambda\hat{d}^k \ge \hat{c}^k$ must be tight at the optimum. Plugging in the previous expression for P_i and the expressions for $\hat{c}^k, \hat{d}^k$, this is implies that λ satisfies

$$-\mathbb{E}^k_n[\max\{\beta_{0,i} - \lambda, 0\}(1 - D_i)(\overline{\delta}_i - \underline{\delta}_i)] = \mathbb{E}^k_n[(\beta_{0,i} - \lambda)\left(\phi_1(Y_i;\hat{\eta}_{-k}) + (1 - D_i)\underline{\delta}_i\right)].$$

Finally, we consider three separate cases:

1. Suppose that $\lambda \ge \max_{i\colon K_i = k}\beta_{0,i}$. From the previous display, λ must satisfy

$$0 = \mathbb{E}^k_n[(\beta_{0,i} - \lambda)\left(\phi_1(Y_i;\hat{\eta}_{-k}) + (1 - D_i)\underline{\delta}_i\right)] \implies \lambda = \frac{\mathbb{E}^k_n[\beta_{0,i}\left(\phi_1(Y_i;\hat{\eta}_{-k}) + (1 - D_i)\underline{\delta}_i\right)]}{\mathbb{E}^k_n[\phi_1(Y_i;\hat{\eta}_{-k}) + (1 - D_i)\underline{\delta}_i]}.$$

   At this value for λ, the expressions for P_i, Q_i imply that $P_i = 0$, $Q_i > 0$ for all i. By complementary slackness, this in turn implies that $\tilde{U}_i = 0$, or equivalently $U_i = 0$ for all i.

2. Suppose that $\lambda \le \min_{i\colon K_i = k}\beta_{0,i}$. From the previous display, λ must satisfy

$$-\mathbb{E}^k_n[(\beta_{0,i} - \lambda)(1 - D_i)(\overline{\delta}_i - \underline{\delta}_i)] = \mathbb{E}^k_n[(\beta_{0,i} - \lambda)\left(\phi_1(Y_i;\hat{\eta}_{-k}) + (1 - D_i)\underline{\delta}_i\right)]$$

$$\implies \lambda = \frac{\mathbb{E}^k_n[\beta_{0,i}\left(\mu_{Y|1}(X_i) + (1 - D_i)\overline{\delta}_i\right)]}{\mathbb{E}^k_n[\phi_1(Y_i;\hat{\eta}_{-k}) + (1 - D_i)\overline{\delta}_i]}.$$

At this value for λ, the expressions for P_i, Q_i imply that $P_i > 0, Q_i = 0$ for all i. By complementary slackness, this implies that $\tilde{U}_i = \tilde{V}$, or equivalently $U_i = 1$ for all i.

3. Suppose that $\min_{i\colon K_i=k} \beta_{0,i} < \lambda < \max_{i\colon K_i=k} \beta_{0,i}$. Then, $\beta_{0,(j)} < \lambda \leq \beta_{0,(j+1)}$ for some j where $\beta_{0,(1)}, \ldots, \beta_{0,(n_k)}$ are the order statistics of the sample outcomes. The expressions for P_i, Q_i in turn imply that $Q_i > 0$ only when $\beta_{0,i} \leq \beta_{0,(k)}$ (in which case $U_i = 0$) and $P_i > 0$ only when $\beta_{0,i} \geq \beta_{0,(k+1)}$ (in which case $U_i = 1$).

Therefore, in all three cases, the optimal solution is such that there exists a non-decreasing function $u(\cdot)\colon \mathbb{R} \to [0,1]$ such that $U_i = u(\beta_{0,i})$ attains the upper bound.

We next prove the result for the population bound $\overline{\text{perf}}_+(s;\beta,\Delta)$ via a similar argument. Applying the same change-of-variables, we rewrite the population bound as

$$\overline{\text{perf}}_+(s;\beta,\Delta) := \sup_{U(\cdot)\colon \mathcal{X}\to[0,1]} \frac{\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\underline{\delta}_i + \beta_{0,i}\pi_0(X_i)(\overline{\delta}_i - \underline{\delta}_i)U(X_i)]}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\underline{\delta}_i + \pi_0(X_i)(\overline{\delta}_i - \underline{\delta}_i)U(X_i)]}.$$

Define $c := \mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\underline{\delta}_i]$, $d := \mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\underline{\delta}_i]$, and $\alpha(x) := \beta_0(x;s)\pi_0(x)(\overline{\delta}(x) - \underline{\delta}(x))$, $\gamma(x) := \pi_0(x)(\overline{\delta}(x) - \underline{\delta}(x))$. Letting $\langle f, g\rangle_{P(\cdot)}$ denote the inner product $\mathbb{E}[f(X_i)g(X_i)]$ for functions $f, g\colon \mathcal{X} \to \mathbb{R}$, we can further rewrite the population bound as

$$\overline{\text{perf}}_+(s;\beta,\Delta) := \sup_{U(\cdot)\colon \mathcal{X}\to[0,1]} \frac{c + \langle \alpha, U\rangle_{P(\cdot)}}{d + \langle \gamma, U\rangle_{P(\cdot)}}.$$

Define the change-of-variables $\tilde{U}(\cdot) = \frac{U(\cdot)}{d + \langle\gamma, U\rangle_{P(\cdot)}}$ and $\tilde{V} = \frac{1}{\langle\gamma, U\rangle_{P(\cdot)}}$. The previous linear-fractional optimization is equivalent to

$$\begin{aligned}
\sup_{\tilde{U}(\cdot),\tilde{V}} \ & \langle \alpha, \tilde{U}\rangle_{P(\cdot)} + c\tilde{V} \\
\text{s.t. } & 0 \leq \tilde{U}(x) \leq \tilde{V} \text{ for all } x \in \mathcal{X}, \\
& \langle \gamma, \tilde{U}\rangle_{P(\cdot)} + \tilde{V}d = 1.
\end{aligned}$$

Define the dual associated with this primal program. Let $\tilde{P}(x)$ be the dual function associated with the constraint $\tilde{U}(x) \leq \tilde{V}$, $\tilde{Q}(x)$ be the dual variables associated with the constraints $\tilde{U}(x) \geq 0$, and λ be the dual variable associated with the constraint $\langle\gamma, \tilde{U}\rangle_{P(\cdot)} + \tilde{V}d = 1$. The dual is

$$\begin{aligned}
\inf_{\lambda,\tilde{P}(\cdot),\tilde{Q}(\cdot)} \ & \lambda \\
\text{s.t.} \, & \tilde{P}(x) - \tilde{Q}(x) + \lambda\gamma(x) = \alpha(x) \text{ for all } x \in \mathcal{X} \\
& -\langle \mathbf{1}, \tilde{P}\rangle_{P(\cdot)} + \lambda d \geq c \\
& \tilde{P}(x) \geq 0, \tilde{Q}(x) \geq \ \text{ for all } x \in \mathcal{X}.
\end{aligned}$$

By complementary slackness, at most only one of $\tilde{P}(x)$ or $\tilde{Q}(x)$ can be non-zero at the optimum for all $x \in \mathcal{X}$. Therefore, by re-arranging the first constraint and substituting in for $\alpha(x), \gamma(x)$, we observe

$$\tilde{P}(x) - \tilde{Q}(x) = (\beta_0(x) - \lambda)\pi_0(x)(\overline{\delta}(x) - \underline{\delta}(x)),$$

which in turn implies that

$$\begin{aligned}
\tilde{P}(x) &= \max\{\beta_0(x) - \lambda, 0\}\pi_0(x)(\overline{\delta}(x) - \underline{\delta}(x)), \\
\tilde{Q}(x) &= \max\{\lambda - \beta_0(x), 0\}\pi_0(x)(\overline{\delta}(x) - \underline{\delta}(x)).
\end{aligned}$$

Furthermore, the constraint $\langle \mathbf{1}, \tilde{P} \rangle_{P(\cdot)} + \lambda d \geq c$ must be tight at the optimum. Plugging in the previous expression for $\tilde{P}(\cdot)$, this is implies that λ satisfies

$$-\mathbb{E}[\max\{\beta_0(X_i) - \lambda, 0\}\pi_0(X_i)(\overline{\delta}(X_i) - \underline{\delta}(X_i))] = \mathbb{E}[(\beta_0(X_i) - \lambda)(\mu_1(X_i) + \pi_0(X_i)\underline{\delta}(X_i)].$$

As in the proof for the estimator, we can consider three cases: (i) $\lambda \geq \overline{\beta}_0$, (ii) $\lambda \leq \underline{\beta}_0$ and (iii) $\underline{\beta}_0 < \lambda < \overline{\beta}_0$ for $\underline{\beta}_0 := \inf_{x \in \mathcal{X}} \beta_0(x)$, $\overline{\beta}_0 = \sup_{x \in \mathcal{X}} \beta_0(x)$. In each case, the optimal solution is such that there exists a non-decreasing function $u(\cdot) \colon \mathbb{R} \to [0, 1]$ such that $U(x) = u(\beta_0(x))$ attains the upper bound. □

A.4.6 Proof of Theorem 5.2

Proof. To ease notation, let $\widehat{\text{perf}}^k_+ := \mathbb{E}^k_n[\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i] / \mathbb{E}^k_n[\phi_1(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i]$. To prove this result, first observe that

$$\begin{aligned}
\left\| \widehat{\overline{\text{perf}}}^k_+(s; \beta, \Delta_n) - \overline{\text{perf}}_+(s; \beta, \Delta) \right\| &= \left\| \sup_{\tilde{\delta} \in \Delta^M_n} \widehat{\text{perf}}^k_+(s; \beta, \tilde{\delta}) - \sup_{\tilde{\delta} \in \Delta^M} \text{perf}_+(s; \beta, \tilde{\delta}) \right\| \\
&= \left\| \sup_{\tilde{\delta} \in \Delta^M} \widehat{\text{perf}}^k_+(s; \beta, \tilde{\delta}) - \sup_{\tilde{\delta} \in \Delta^M} \text{perf}+(s; \beta, \tilde{\delta}) \right\| \\
&\leq \sup_{\tilde{\delta} \in \Delta^M} \left\| \widehat{\text{perf}}^k_+(s; \beta, \tilde{\delta}) - \text{perf}_+(s; \beta, \tilde{\delta}) \right\|,
\end{aligned}$$

where the first equality uses Lemma 5.2. Furthermore, for any $\tilde{\delta} \in \Delta^M$, we have that

$$\widehat{\text{perf}}^k_+(s; \beta, \tilde{\delta}) - \text{perf}_+(s; \beta, \tilde{\delta}) =$$

$$\frac{\mathbb{E}^k_n[\beta_{0,i}\phi_{Y|1}(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i]}{\mathbb{E}^k_n[\phi_{Y|1}(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i]} - \frac{\mathbb{E}[\beta_{0,i}\phi_{Y|1}(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i]}{\mathbb{E}[\phi_{Y|1}(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i]} =$$

$$\frac{\mathbb{E}^k_n[(1)]}{\mathbb{E}^k_n[(2)]} - \frac{\mathbb{E}[(3)]}{\mathbb{E}[(4)]} = \mathbb{E}^k_n[(2)]^{-1} \left\{ \mathbb{E}^k_n[(1)] - \mathbb{E}[(3)] - \frac{\mathbb{E}[(3)]}{\mathbb{E}[(4)]}(\mathbb{E}^k_n[(2)] - \mathbb{E}[(4)]) \right\},$$

where

$$\begin{aligned}
\mathbb{E}^k_n[(1)] - \mathbb{E}[(3)] &= \mathbb{E}^k_n[\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i] - \mathbb{E}[\beta_{0,i}\phi_1(Y_i; \eta) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i] \\
&= \left(\mathbb{E}^k_n[\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k})] - \mathbb{E}[\beta_{0,i}\phi_1(Y_i; \eta)] \right) + (\mathbb{E}^k_n - \mathbb{E})[\beta_{0,i}(1 - D_i)\tilde{\delta}_i] \\
\mathbb{E}^k_n[(2)] - \mathbb{E}[(4)] &= \mathbb{E}^k_n[\phi_1(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i] - \mathbb{E}[\phi_1(Y_i; \eta) + (1 - D_i)\tilde{\delta}_i] \\
&= \left(\mathbb{E}^k_n[\phi_1(Y_i; \hat{\eta}_{-k})] - \mathbb{E}[\phi_1(Y_i; \eta)] \right) + (\mathbb{E}^k_n - \mathbb{E})[(1 - D_i)\tilde{\delta}_i].
\end{aligned}$$

Furthermore, observe that

$$\begin{aligned}
&\mathbb{E}^k_n[(2)] = \mathbb{E}^k_n[\phi_1(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i] \geq \mathbb{E}^k_n[\phi_1(Y_i; \hat{\eta}_{-k}) + (1 - D_i)\tilde{\delta}_i] \\
&\mathbb{E}[(3)] = \mathbb{E}[\beta_{0,i}\phi_1(Y_i; \eta) + \beta_{0,i}(1 - D_i)\tilde{\delta}_i] \leq \mathbb{E}[\beta_{0,i}\phi_1(Y_i; \eta) + \beta_{0,i}(1 - D_i)\overline{\delta}_i] \\
&\mathbb{E}[(4)] = \mathbb{E}[\mu_1(X_i) + (1 - D_i)\tilde{\delta}_i] \geq \mathbb{E}[\mu_1(X_i) + (1 - D_i)\underline{\delta}_i].
\end{aligned}$$

Therefore, there exists $C_1 > 0$ such that $\mathbb{E}_n[(2)] > C_1$ for all n under the assumption of bounded nuisance parameter estimates. There also exists constants $C_2 < \infty$, $C_3 > 0$ such that $\mathbb{E}[(3)] < C_2$ and

$\mathbb{E}[(4)] > C_3$. Putting this together, we therefore have

$$\left\|\widehat{\overline{\text{perf}}}^k(s;\beta,\Delta_n) - \overline{\text{perf}}_+(s;\beta,\Delta)\right\| \leq$$

$$C_1\left\|\underbrace{\mathbb{E}_n^k[\beta_{0,i}\phi_1(Y_i;\hat{\eta}_{-k})] - \mathbb{E}[\beta_{0,i}\phi_1(Y_i;\eta)]}_{(a)}\right\| + C_1\underbrace{\sup_{\tilde{\delta}\in\Delta^M}\left\|(\mathbb{E}_n^k - \mathbb{E})[\beta_{0,i}(1-D_i)\tilde{\delta}_i]\right\|}_{(b)} +$$

$$C_1\frac{C_2}{C_3}\left\|\underbrace{\mathbb{E}_n^k[\phi_1(Y_i;\hat{\eta}_{-k})] - \mathbb{E}[\phi_1(Y_i;\eta)]}_{(c)}\right\| + C_1\frac{C_2}{C_3}\underbrace{\sup_{\tilde{\delta}\in\Delta^M}\left\|(\mathbb{E}_n^k - \mathbb{E})[(1-D_i)\tilde{\delta}_i]\right\|}_{(d)}.$$

We analyze each term separately. From the proof of Theorem 5.1, we showed that (a), (c) are $O_{\mathbb{P}}(1/\sqrt{n}+R_{1,n})$. Consider (b), which we may write out as

$$\sup_{\tilde{\delta}\in\Delta^M}\left|n_k^{-1}\sum_{i\colon K_i=k}(1-D_i)\beta_{0,i}\tilde{\delta}(\beta_{0,i}) - \mathbb{E}[(1-D_i)\beta_{0,i}\tilde{\delta}(\beta_{0,i})]\right|.$$

Define $f(a,b)=ab$, and $\mathcal{F}=\{f_{\tilde{\delta}}\}_{\tilde{\delta}\in\Delta^M}$ to be the class of functions $f_{\tilde{\delta}}\colon(d,\beta)\to(1-d)\beta\tilde{\delta}(\beta)$. Observe that f is a contraction in its second argument over $\{0,1\}\times[0,1]$. Observe that we can then rewrite (b) as

$$\sup_{f_{\tilde{\delta}}\in\mathcal{F}}\left|n_k^{-1}\sum_{i\colon K_i=k}f_{\tilde{\delta}}(D_i,\beta_{0,i}) - \mathbb{E}[f_{\tilde{\delta}}(D_i,\beta_{0,i})]\right|.$$

Applying a standard concentration inequality (e.g., Theorem 4.10 in Wainwright (2019)), we observe that, with probability at least $1-\delta$,

$$\sup_{f_{\tilde{\delta}}\in\mathcal{F}}\left|n_k^{-1}\sum_{i\colon K_i=k}f_{\tilde{\delta}}(D_i,\beta_{0,i}) - \mathbb{E}[f_{\tilde{\delta}}(D_i,\beta_{0,i})]\right| \leq R_n(\mathcal{F}) + \sqrt{\frac{2\log(1/\delta)}{n_k}},$$

where $R_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$. Now we relate $R_n(\mathcal{F})$ to $R_n(\Delta^M)$. For any fixed tuples $(d_1,\beta_{0,1}),\dots,(d_{n_k},\beta_{0,n_k})$, observe that

$$\begin{aligned}\mathbb{E}_\epsilon[\sup_{\tilde{\delta}\in\Delta^M}\left|\sum_{i=1}^{n_k}\epsilon_i f(d_i,\tilde{\delta}(\beta_{0,i}))\right|] &= \mathbb{E}_\epsilon[\sup_{\tilde{\delta}\in\Delta^M}\left|\sum_{i=1}^{n_k}\epsilon_i(1-d_i)\tilde{\delta}(\beta_{0,i})\right|] \\ &\leq 2\mathbb{E}_\epsilon[\sup_{\tilde{\delta}\in\Delta^M}\left|\sum_{i=1}^{n_k}\epsilon_i\tilde{\delta}(\beta_{0,i})\right|]\end{aligned}$$

where the inequality applies the Ledoux-Talagrand contraction inequality (e.g., Eq. (5.61) in Wainwright (2019)). Dividing by n and averaging over the tuples yields $R_n(\mathcal{F}) \leq 2R_n(\Delta^M)$. Finally, we can bound the Rademacher complexity of Δ^M using Dudley's entropy integral (e.g., Theorem 5.22 Wainwright (2019)) as

$$R_n(\Delta^M) \leq \frac{C}{\sqrt{n_k}}\int_0^1\sqrt{\log(N(\xi,\Delta^M,\|\cdot\|_{\mathbb{P}_n}))}d\xi \leq \frac{C}{\sqrt{n_k}}\int_0^1\sqrt{\log(N_{[]}(2\xi,\Delta^M,\|\cdot\|_{\mathbb{P}_n}))}d\xi,$$

for some constant C, where $N(\xi, \Delta^M, \|\cdot\|_{\mathbb{P}_n})$ is the covering number and $N_{[]}(2\xi, \Delta^M, \|\cdot\|_{p_n})$ is the bracketing number. But, Theorem 2.7.5 of van der Vaart and Wellner (1996) establishes that the bracketing entropy $\log(N_{[]}(\xi, \Delta^M, \|\cdot\|_{\mathbb{P}_n})$ of the class of monotone non-decreasing functions is bounded by $(1/\xi)\log(1/\xi)$, and so $\int_0^1 \sqrt{\log(N_{[]}(\xi, \Delta^M, \|\cdot\|_{\mathbb{P}_n}))}d\xi = \sqrt{2\pi}$. It follows that, for any $\delta > 0$,

$$\sup_{\tilde{\delta}\in\Delta^M} \left|\mathbb{E}_n^k[(1-D_i)\beta_{0,i}\tilde{\delta}(X_i)] - \mathbb{E}[(1-D_i)\beta_{0,i}\tilde{\delta}(X_i)]\right| \leq \frac{C}{\sqrt{n_k}} + \sqrt{\frac{2\log(1/\delta)}{2n_k}}$$

holds with probability $1-\delta$. We therefore conclude that (b) is $O_{\mathbb{P}}(1/\sqrt{n})$. Similarly, (d) is $O_{\mathbb{P}}(1/\sqrt{n})$ by the same argument. This proves the result for any fold k. The claim in the Theorem follows by averaging over the folds. □

A.4.7 Proof of Proposition 5.3

Proof. Applying the change-of-variables in the Proof of Lemma 5.1, we notice that

$$\widehat{\text{perf}}_+^k(s;\beta,\Delta_n) := \max_{0\leq U\leq 1} \frac{\sum_{i=1}^{n_k}\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\underline{\delta}_i + \beta_{0,i}(1-D_i)(\overline{\delta}_i - \underline{\delta}_i)U_i}{\sum_{i=1}^{n_k}\phi_1(Y_i;\hat{\eta}) + (1-D_i)\underline{\delta}_i + (1-D_i)(\overline{\delta}_i - \underline{\delta}_i)U_i}$$

$$\widehat{\text{perf}}_+^k(s;\beta,\hat{\Delta}_n) := \max_{0\leq U\leq 1} \frac{\sum_{i=1}^{n_k}\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\hat{\underline{\delta}}_i + \beta_{0,i}(1-D_i)(\hat{\overline{\delta}}_i - \hat{\underline{\delta}}_i)U_i}{\sum_{i=1}^{n_k}\phi_1(Y_i;\hat{\eta}) + (1-D_i)\hat{\underline{\delta}}_i + (1-D_i)(\hat{\overline{\delta}}_i - \hat{\underline{\delta}}_i)U_i}.$$

We can therefore rewrite

$$\|\widehat{\text{perf}}_+^k(s;\beta,\hat{\Delta}_n) - \widehat{\text{perf}}_+^k(s;\beta,\Delta_n)\| \leq$$

$$\max_{0\leq U\leq 1}\left\|\frac{\sum_{i=1}^{n}\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\hat{\underline{\delta}}_i + \beta_{0,i}(1-D_i)(\hat{\overline{\delta}}_i - \hat{\underline{\delta}}_i)U_i}{\sum_{i=1}^{n}\phi_1(Y_i;\hat{\eta}) + (1-D_i)\hat{\underline{\delta}}_i + (1-D_i)(\hat{\overline{\delta}}_i - \hat{\underline{\delta}}_i)U_i} - \frac{\sum_{i=1}^{n}\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\underline{\delta}_i + \beta_{0,i}(1-D_i)(\overline{\delta}_i - \underline{\delta}_i)U_i}{\sum_{i=1}^{n}\phi_1(Y_i;\hat{\eta}) + (1-D_i)\underline{\delta}_i + (1-D_i)(\overline{\delta}_i - \underline{\delta}_i)U_i}\right\| =$$

$$\max_{0\leq U\leq 1}\left\|\frac{\mathbb{E}_n^k[(1)]}{\mathbb{E}_n^k[(2)]} - \frac{\mathbb{E}_n^k[(3)]}{\mathbb{E}_n^k[(4)]}\right\| = \max_{0\leq U\leq 1}\mathbb{E}_n[(2)]^{-1}\left\{\underbrace{\left(\mathbb{E}_n^k[(1)] - \mathbb{E}_n^k[(3)]\right)}_{(a)} - \frac{\mathbb{E}_n^k[(3)]}{\mathbb{E}_n^k[(4)]}\underbrace{\left(\mathbb{E}_n^k[(2)] - \mathbb{E}_n^k[(4)]\right)}_{(b)}\right\}.$$

Notice that we can rewrite (a), (b) as

$$(a) = \mathbb{E}_n^k[\beta_{0,i}(1-D_i)(\hat{\underline{\delta}}_i - \underline{\delta}_i)(1-U_i) + \beta_{0,i}(1-D_i)(\hat{\overline{\delta}}_i - \overline{\delta}_i)U_i]$$

$$(b) = \mathbb{E}_n^k[(1-D_i)(\hat{\underline{\delta}}_i - \underline{\delta}_i)(1-U_i) + (1-D_i)(\hat{\overline{\delta}}_i - \overline{\delta}_i)U_i].$$

Furthermore, notice that

$$\mathbb{E}_n^k[(2)] \geq \mathbb{E}_n^k[\phi_1(Y_i;\hat{\eta}) + (1-D_i)\underline{\delta}_i] \text{ for all } n,$$

$$\mathbb{E}_n^k[(3)] \leq \mathbb{E}_n^k[\beta_{0,i}\phi_1(Y_i;\hat{\eta}) + \beta_{0,i}(1-D_i)\overline{\delta}_i] \text{ for all } n,$$

$$\mathbb{E}_n^k[(4)] \leq \mathbb{E}_n^k[\phi_1(Y_i;\hat{\eta}) + (1-D_i)\underline{\delta}_i] \text{ for all } n.$$

So, there exists a constant $0 < C_1$ such that $\mathbb{E}_n[(2)] > C_1$ for all n under the assumption of bounded nuisance parameter estimators and the assumption on the estimated bounds and there exists constants $0 < C_2 < \infty, C_3 > 0$ such that $\mathbb{E}[(3)] < C_2$, $\mathbb{E}[(4)] > C_3$ under the assumption of bounded nuisance parameter estimators. Putting this together implies that

$$\|\widehat{\text{perf}}^k_+(s;\beta,\hat{\Delta}_n) - \widehat{\text{perf}}^k_+(s;\beta,\Delta_n)\| \lesssim$$

$$\max_{0\leq U\leq 1} \|\mathbb{E}_n[\beta_{0,i}(1-D_i)(\hat{\underline{\delta}}_i-\underline{\delta}_i)(1-U_i)+\beta_{0,i}(1-D_i)(\hat{\overline{\delta}}_i-\overline{\delta}_i)U_i]\|+\|\mathbb{E}_n[(1-D_i)(\hat{\underline{\delta}}_i-\underline{\delta}_i)(1-U_i)+(1-D_i)(\hat{\overline{\delta}}_i-\overline{\delta}_i)U_i]\| \leq$$

$$\max_{0\leq U\leq 1} n_k^{-1}\sum_{i=1}^{n_k}\|\beta_{0,i}(1-D_i)\{(\hat{\underline{\delta}}_i-\underline{\delta}_i)(1-U_i)+(\hat{\overline{\delta}}_i-\overline{\delta}_i)U_i\}\|+n_k^{-1}\sum_{i=1}^{n_k}\|(1-D_i)\{(\hat{\underline{\delta}}_i-\underline{\delta}_i)(1-U_i)+(\hat{\overline{\delta}}_i-\overline{\delta}_i)U_i\}\| \leq$$

$$\max_{0\leq U\leq 1} n_k^{-1}\sum_{i=1}^{n_k}\|(\hat{\underline{\delta}}_i-\underline{\delta}_i)(1-U_i)+(\hat{\overline{\delta}}_i-\overline{\delta}_i)U_i\|+n_k^{-1}\sum_{i=1}^{n_k}\|(\hat{\underline{\delta}}_i-\underline{\delta}_i)(1-U_i)+(\hat{\overline{\delta}}_i-\overline{\delta}_i)U_i\}\| \lesssim \mathbb{E}_{n_k}[|\hat{\underline{\delta}}_i-\underline{\delta}_i|]+\mathbb{E}_{n_k}[|\hat{\overline{\delta}}_i-\overline{\delta}_i|].$$

Then, using the inequality $\|v\|_1 \leq \sqrt{n_k}\|v\|_2$ for $v \in \mathbb{R}^{n_k}$, it follows that

$$\|\widehat{\text{perf}}^k_+(s;\beta,\hat{\Delta}_n) - \widehat{\text{perf}}^k_+(s;\beta,\Delta_n)\| \lesssim \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(\hat{\underline{\delta}}_i-\underline{\delta}_i)^2} + \sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(\hat{\overline{\delta}}_i-\overline{\delta}_i)^2}.$$

□

A.4.8 Proof of Corollary 5.1

Proof. By Proposition 5.3, it suffices to show that $\|\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta)\|_{L_2(\mathbb{P}^k_n)} = o_{\mathbb{P}}(1)$ under the stated conditions. Following the proof of Lemma B.5, we observe that

$$\|\phi_1(Y_i;\hat{\eta}_{-k}) - \phi_1(Y_i;\eta)\|_{L_2(\mathbb{P}^k_n)} \leq$$

$$O_{\mathbb{P}}(\|\hat{\pi}_1-\pi_1\|_{L_2(\mathbb{P})}\|Y_i-\mu_1\|_{L_2(\mathbb{P})} + \|\hat{\pi}_1-\pi_1\|_{L_2(\mathbb{P})}\|\mu_1-\hat{\mu}_1\|_{L_2(\mathbb{P})} + \|\mu_1-\hat{\mu}_1\|_{L_2(\mathbb{P})}).$$

The result then follows by the stated rate conditions. □

A.5 Section 8: connections to existing sensitivity analysis models

A.5.1 Proof of Proposition 8.1

Proof. For brevity, we omit the conditioning on X_i throughout the proof. Consider the first claim. Notice that by Bayes' rule, $\frac{\mathbb{P}(D_i=1|Y_i(1),Y_i(0))\mathbb{P}(D_i=0)}{\mathbb{P}(D_i=0|Y_i(1),Y_i(0))\mathbb{P}(D_i=1)} = \frac{\mathbb{P}(Y_i(1),Y_i(0)|D_i=1)}{\mathbb{P}(Y_i(1),Y_i(0)|D_i=0)}$. The MSM therefore implies bounds

$$\underline{\Lambda} \leq \frac{\mathbb{P}(Y_i(1),Y_i(0)\mid D_i=1)}{\mathbb{P}(Y_i(1),Y_i(0)\mid D_i=0)} \leq \overline{\Lambda},$$

which can be equivalently written as

$$\overline{\Lambda}^{-1}\mathbb{P}(Y_i(1),Y_i(0)\mid D_i=1) \leq \mathbb{P}(Y_i(1),Y_i(0)\mid D_i=0) \leq \underline{\Lambda}^{-1}\mathbb{P}(Y_i(1),Y_i(0)\mid D_i=1).$$

Since $\mathbb{P}(Y_i(1)=1\mid D_i=0) = \mathbb{P}(Y_i(0)=0,Y_i(1)=1\mid D_i=0) + \mathbb{P}(Y_i(0)=1,Y_i(1)=1\mid D_i=0)$, it then follows that

$$\overline{\Lambda}^{-1}\mathbb{P}(Y_i(1)=1\mid D_i=1) \leq \mathbb{P}(Y_i(1)=1\mid D_i=0) \leq \underline{\Lambda}^{-1}\mathbb{P}(Y_i(1)=1\mid D_i=1).$$

Adding and subtracting $\mathbb{P}(Y_i(1)=1\mid D_i=1)$ then delivers the first claim.

Consider the second claim. The MOSM under nonparametric outcome regression bounds implies that

$$\overline{\Gamma}^{-1} \leq \frac{\mathbb{P}(Y_i(1)=1\mid D_i=1)}{\mathbb{P}(Y_i(1)=1\mid D_i=0)} \leq \underline{\Gamma}^{-1}.$$

But, by Bayes' rule, $\frac{\mathbb{P}(Y_i(1)=1|D_i=1)}{\mathbb{P}(Y_i(1)=1|D_i=0)} = \frac{\mathbb{P}(D_i=1|Y_i(1)=1)\mathbb{P}(D_i=0)}{\mathbb{P}(D_i=0|Y_i(1)=1)\mathbb{P}(D_i=1)}$, and so the MOSM implies the bounds

$$\overline{\Gamma}^{-1} \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = 1)\mathbb{P}(D_i = 0)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = 1)\mathbb{P}(D_i = 1)} \leq \underline{\Gamma}^{-1}.$$

Further, the MOSM implies

$$\frac{\underline{\Gamma} - 1}{\underline{\Gamma}\mathbb{P}(Y_i(1) = 0 \mid D_i = 0)} + \frac{1}{\underline{\Gamma}} \leq \frac{\mathbb{P}(Y_i(1) = 0 \mid D_i = 1)}{\mathbb{P}(Y_i(1) = 0 \mid D_i = 0)} \leq \frac{\overline{\Gamma} - 1}{\overline{\Gamma}P(Y_i(1) \mid D_i = 0)} + \frac{1}{\overline{\Gamma}}.$$

Applying the analogous identity $\frac{\mathbb{P}(Y_i(1)=0|D_i=1)}{\mathbb{P}(Y_i(1)=0|D_i=0)} = \frac{\mathbb{P}(D_i=1|Y_i(1)=0)\mathbb{P}(D_i=0)}{\mathbb{P}(D_i=0|Y_i(1)=0)\mathbb{P}(D_i=1)}$ then delivers

$$\frac{\underline{\Gamma} - 1}{\underline{\Gamma}\mathbb{P}(Y_i(1) = 0 \mid D_i = 0)} + \frac{1}{\underline{\Gamma}} \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = 0)\mathbb{P}(D_i = 0)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = 0)\mathbb{P}(D_i = 1)} \leq \frac{\overline{\Gamma} - 1}{\overline{\Gamma}P(Y_i(1) \mid D_i = 0)} + \frac{1}{\overline{\Gamma}}.$$

But since the MOSM also implies that $1-\overline{\Gamma}\mu_1(x) \leq \mathbb{P}(Y_i(0) = 1 \mid D_i = 1) \leq 1-\underline{\Gamma}\mu_1(x)$, we can plug-in to deliver the final bounds

$$\frac{\underline{\Gamma} - 1}{\underline{\Gamma}(1 - \underline{\Gamma}\mu_1(x))} + \frac{1}{\underline{\Gamma}} \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = 0)\mathbb{P}(D_i = 0)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = 0)\mathbb{P}(D_i = 1)} \leq \frac{\overline{\Gamma} - 1}{\overline{\Gamma}(1 - \overline{\Gamma}\mu_1(x))} + \frac{1}{\overline{\Gamma}}$$

This completes the proof of the second claim. □

A.5.2 Proof of Proposition 8.2

Proof. For brevity, we omit the conditioning on X_i throughout the proof. To show the first claim, as a first step, apply Bayes' rule and observe that

$$\frac{\mathbb{P}(Y_i(1), Y_i(0) \mid D_i = 1)}{\mathbb{P}(Y_i(1), Y_i(0) \mid D_i = 0)} = \frac{\mathbb{P}(D_i = 1 \mid Y_i(1), Y_i(0))\mathbb{P}(D_i = 0)}{\mathbb{P}(D_i = 0 \mid Y_i(1), Y_i(0))\mathbb{P}(D_i = 1)}.$$

Then, further notice that

$$\frac{\mathbb{P}(D_i = 0)}{\mathbb{P}(D_i = 1)} = \frac{\sum_{(y_0,y_1)\in\{0,1\}^2} P(D_i = 0 \mid Y_i(0) = y_0, Y_i(1) = y_1)P(Y_i(0) = y_0, Y_i(1) = y_1)}{\sum_{(y_0,y_1)\in\{0,1\}^2} P(D_i = 1 \mid Y_i(0) = y_0, Y_i(1) = y_1)P(Y_i(0) = y_0, Y_i(1) = y_1)}$$

Letting $(y_0^*, y_1^*) = \arg\max_{(y_0,y_1)\in\{0,1\}^2} \frac{P(D_i=0|Y_i(0)=y_0,Y_i(1)=y_1)}{P(D_i=1|Y_i(0)=y_0,Y_i(1)=y_1)}$, the quasi-linearity of the ratio function implies that

$$\frac{\sum_{(y_0,y_1)\in\{0,1\}^2} P(D_i = 0 \mid Y_i(0) = y_0, Y_i(1) = y_1)P(Y_i(0) = y_0, Y_i(1) = y_1)}{\sum_{(y_0,y_1)\in\{0,1\}^2} P(D_i = 1 \mid Y_i(0) = y_0, Y_i(1) = y_1)P(Y_i(0) = y_0, Y_i(1) = y_1)} \leq \frac{P(D_i = 0 \mid Y_i(0) = y_0^*, Y_i(1) = y_1^*)}{P(D_i = 1 \mid Y_i(0) = y_0^*, Y_i(1) = y_1^*)}.$$

This then implies that, for any $(y_0, y_1) \in \{0, 1\}^2$,

$$\frac{\mathbb{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid D_i = 1)}{\mathbb{P}(Y_i(1) = y_1, Y_i(0) = y_0 \mid D_i = 0)} \leq \frac{\mathbb{P}(D_i = 1 \mid Y_i(1) = y_1, Y_i(0) = y_0)}{\mathbb{P}(D_i = 0 \mid Y_i(1) = y_1, Y_i(0) = y_0)} \frac{P(D_i = 0 \mid Y_i(0) = y_0^*, Y_i(1) = y_1^*)}{P(D_i = 1 \mid Y_i(0) = y_0^*, Y_i(1) = y_1^*)} \leq \Gamma,$$

where the last inequality is implied by Rosenbaum's sensitivity analysis model (23). From this, we follow the same argument as the proof of Proposition 8.1 to show that $\overline{\delta}(x) = (\Gamma - 1)\mu_1(x)$. The proof for the lower bound follows an analogous argument.

The second claim is an immediate consequence of claim (ii) in Proposition 8.1. □

B Auxiliary lemmas

B.1 An oracle inequality for pseudo-outcome regression

In this section, we provide a model-free oracle inequality on the $L_2(\mathbb{P})$-error of nonparametric regression with estimated pseudo-outcomes. This generalizes the analysis of pseudo-outcome regressions provided in Kennedy (2022*b*). In the main text, we apply this oracle inequality to the DR-Learners.

We state an $L_2(\mathbb{P})$-stability condition required on the second-stage regression estimator, extending the pointwise stability condition in Kennedy (2022*b*).

Assumption B.1. Suppose $\mathcal{O}_{train} = (V_{01}, \ldots, V_{0n})$ and $\mathcal{O}_{test} = (V_1, \ldots, V_n)$ are independent train and test sets with covariate $X_i \subseteq V_i$. Let (i) $\hat{f}(w) := \hat{f}(w; \mathcal{O}_{train})$ be an estimate of a function $f(w)$ using the training data $\mathcal{O}_{train}$; (ii) $\hat{b}(x) = \mathbb{E}[\hat{f}(V_i) - f(V_i) \mid X_i = x, \mathcal{O}_{train}]$ be the conditional bias of the estimator $\hat{f}$; and (iii) $\hat{\mathbb{E}}_n[V_i \mid X_i = x]$ be a generic regression estimator that regresses outcomes $(V_1, \ldots, V_n)$ on covariates $(X_1, \ldots, X_n)$ in the test sample $\mathcal{O}_{test}$.

The regression estimator $\hat{E}_n[\cdot]$ is $L_2(\mathbb{P})$-*stable* with respect to a distance metric $d(\cdot, \cdot)$ if

$$\frac{\int \left[\widehat{\mathbb{E}}_n\{\widehat{f}(V_i) \mid X_i = x\} - \widehat{\mathbb{E}}_n\{f(V_i) \mid X_i = x\} - \widehat{\mathbb{E}}_n\{\widehat{b}(X_i) \mid X_i = x\}\right]^2 d\mathbb{P}(x)}{\mathbb{E}\left(\int \left[\widehat{\mathbb{E}}_n\{f(V_i) \mid X_i = x\} - \mathbb{E}\{f(V_i) \mid X_i = x\}\right]^2 d\mathbb{P}(x)\right)} \xrightarrow{p} 0 \tag{24}$$

whenever $d(\widehat{f}, f) \xrightarrow{p} 0$.

The $L_2(\mathbb{P})$-stability condition on the second-stage pseudo-outcome regression estimator is quite mild in practice. We next show that the $L_2(\mathbb{P})$-stability condition is satisfied by a variety of generic linear smoothers such as linear regression, series regression, nearest neighbor matching, random forest model and several others. This extends Theorem 1 of Kennedy (2022*b*), which shows that linear smoothers satisfy an analogous pointwise stability condition.

Proposition B.1. *Linear smoothers of the form* $\widehat{\mathbb{E}}_n\{\widehat{f}(V_i) \mid X_i = x\} = \sum_i w_i(x; X^n)\widehat{f}(V_i)$ *are* $L_2(\mathbb{P})$-*stable with respect to distance*

$$d(\widehat{f}, f) = \|\widehat{f} - f\|_{w^2} \equiv \sum_{i=1}^{n} \left\{ \frac{\|w_i(\cdot; X^n)\|^2}{\sum_j \|w_j(\cdot; X^n)\|^2} \right\} \int \left\{\widehat{f}(v) - f(v)\right\}^2 \, d\mathbb{P}(v \mid X_i),$$

whenever $1/\|\sigma\|_{w^2} = O_{\mathbb{P}}(1)$ *for* $\sigma(x)^2 = Var\{f(V_i) \mid X_i = x\}$.

Proof. The proof follows an analogous argument as Theorem 1 of Kennedy (2022*b*). Letting $T_n(x) = \widehat{m}(x) - \widetilde{m}(x) - \widehat{\mathbb{E}}_n\{\widehat{b}(X) \mid X = x\}$ denote the numerator of the left-hand side of (24), and $R_n^2 = \mathbb{E}[\|\widetilde{m} - m\|]^2$ denote the oracle error, we will show that

$$\|T_n\| = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|_{w^2}}{\|\sigma\|_{w^2}} R_n\right)$$

which yields the result when $1/\|\sigma\|_{w^2} = O_{\mathbb{P}}(1)$.

First, note that for linear smoothers

$$T_n(x) = \widehat{\mathbb{E}}_n\{\widehat{f}(V_i) - f(V_i) - \widehat{b}(X_i) \mid X_i = x\} = \sum_{i=1}^{n} w_i(x; X^n)\left\{\widehat{f}(V_i) - f(V_i) - \widehat{b}(X_i)\right\}$$

and this term has mean zero since

$$\mathbb{E}\left\{\widehat{f}(V_i) - f(V_i) - \widehat{b}(X_i) \mid \mathcal{O}_{train}, X^n\right\} = \mathbb{E}\left\{\widehat{f}(V_i) - f(V_i) - \widehat{b}(X_i) \mid \mathcal{O}_{train}, X_i\right\} = 0$$

by definition of $\widehat{b}$ and iterated expectation. Therefore,

$$\begin{aligned}\mathbb{E}(T_n(x)^2 \mid \mathcal{O}_{train}, X^n) &= Var\left[\sum_{i=1}^{n} w_i(x;X^n)\Big\{\widehat{f}(V_i) - f(V_i) - \widehat{b}(X_i)\Big\} \;\Big|\; \mathcal{O}_{train}, X^n\right] \\ &= \sum_{i=1}^{n} w_i(x;X^n)^2 \; Var\Big\{\widehat{f}(V_i) - f(V_i) \mid \mathcal{O}_{train}, X_i\Big\} \end{aligned} \tag{25}$$

where the second line follows since $\widehat{f}(V_i) - f(V_i)$ are independent given the training data. Thus

$$\begin{aligned}\mathbb{E}\left(\|T_n\|^2 \;\Big|\; \mathcal{O}_{train}, X^n\right) &= \int \sum_{i=1}^{n} w_i(x;X^n)^2 \; Var\Big\{\widehat{f}(V_i) - f(V_i) \mid \mathcal{O}_{train}, X_i\Big\} \; d\mathbb{P}(x) \\ &= \sum_{i=1}^{n} \|w_i(\cdot;X^n)\|^2 Var\Big\{\widehat{f}(V_i) - f(V_i) \mid \mathcal{O}_{train}, X_i\Big\} \\ &\leq \sum_{i=1}^{n} \|w_i(\cdot;X^n)\|^2 \int \Big\{\widehat{f}(v) - f(v)\Big\}^2 \; d\mathbb{P}(v \mid X_i) \\ &= \|\widehat{f} - f\|_{w^2} \sum_{j} \|w_j(\cdot;X^n)\|^2 \end{aligned}$$

where the third line follows since $Var(\widehat{f} - f \mid \mathcal{O}_{train}, X_i) \leq \mathbb{E}\{(\widehat{f} - f)^2 \mid \mathcal{O}_{train}, X_i\}$, and the fourth by definition of $\|\cdot\|_{w^2}$.

Further note that R_n^2 equals

$$\begin{aligned}\mathbb{E}[\|\widetilde{m} - m\|]^2 &= \mathbb{E}\left(\int \left[\sum_{i=1}^{n} w_i(x;X^n)\Big\{f(V_i) - m(X_i)\Big\} + \sum_{i=1}^{n} w_i(x;X^n)m(X_i) - m(x)\right]^2 d\mathbb{P}(x)\right) \\ &= \mathbb{E}\left(\int \left[\sum_{i=1}^{n} w_i(x;X^n)\Big\{f(V_i) - m(X_i)\Big\}\right]^2 d\mathbb{P}(x)\right) + \mathbb{E}\left[\int \left\{\sum_{i=1}^{n} w_i(x;X^n)m(X_i) - m(x)\right\}^2 d\mathbb{P}(x)\right] \\ &= \mathbb{E}\left\{\int \sum_{i=1}^{n} w_i(x;X^n)^2 \; \sigma(X_i)^2 \; d\mathbb{P}(x)\right\} + \mathbb{E}\left[\int \left\{\sum_{i=1}^{n} w_i(x;X^n)m(X_i) - m(x)\right\}^2 d\mathbb{P}(x)\right] \\ &\geq \mathbb{E}\sum_{i=1}^{n} \|w_i(\cdot;X^n)\|^2 \sigma(X_i)^2 = \mathbb{E}\left\{\|\sigma\|_{w^2}^2 \sum_{j} \|w_j(\cdot;X^n)\|^2\right\} \end{aligned} \tag{26}$$

where the second and third lines follow from iterated expectation and independence of the samples, and the fourth by definition of $\|\cdot\|_{w^2}$ (and since the integrated squared bias term from the previous line is non-negative).

Therefore

$$\begin{aligned}
\mathbb{P}\left\{\frac{\|\sigma\|_{w^2}\|T_n\|}{\|\widehat{f}-f\|_{w^2}R_n} \geq t\right\} &= \mathbb{E}\left[\mathbb{P}\left\{\frac{\|\sigma\|_{w^2}\|T_n\|}{\|\widehat{f}-f\|_{w^2}R_n} \geq t \;\Big|\; \mathcal{O}_{train}, X^n\right\}\right] \\
&\leq \left(\frac{1}{t^2R_n^2}\right)\mathbb{E}\left\{\|\sigma\|_{w^2}^2\mathbb{E}\left(\frac{\|T_n\|^2}{\|\widehat{f}-f\|_{w^2}^2} \;\Big|\; \mathcal{O}_{train}, X^n\right)\right\} \\
&\leq \left(\frac{1}{t^2R_n^2}\right)\mathbb{E}\left\{\|\sigma\|_{w^2}^2\sum_{i=1}^n \|w_i(\cdot; X^n)\|^2\right\} \leq \frac{1}{t^2}
\end{aligned}$$

where the second line follows by Markov's inequality, the third from the bound in (25) and iterated expectation, and the last from the bound in (26). The result follows since we can always pick $t^2 = 1/\epsilon$ to ensure the above probability is no more than any ϵ. ☐

We next show that the $L_2(\mathbb{P})$-stability condition and the consistency of $\hat{f}$ yields an inequality on the $L_2(\mathbb{P})$-convergence of a feasible pseudo-outcome regression relative to an oracle estimator that regresses the true unknown function $f(V_i)$ on X_i.

Lemma B.1. *Under the same setup from Assumptions B.1, define (i) $m(x) = \mathbb{E}[f(V_i) \mid X_i = x]$ the conditional expectation of $f(V_i)$ given X_i; (ii) $\hat{m}(x) := \hat{\mathbb{E}}_n[\hat{f}(V_i) \mid X_i = x]$ the regression of $\hat{f}(V_i)$ on X_i in the test samples; (iii) $\tilde{m}(x) := \hat{\mathbb{E}}_n[f(V_i) \mid X_i = x]$ the oracle regression of $f(V_i)$ on X_i in the test samples. Furthermore, let $\tilde{b}(x) := \hat{\mathbb{E}}_n[b(V_i) \mid X_i = x]$ be the $\hat{\mathbb{E}}_n$-smoothed bias and $R_n^2 = E[\|\widetilde{m} - m\|]^2$ be the oracle L_2-error. If*

i. the regression estimator $\hat{\mathbb{E}}_n[\cdot]$ is $L_2(\mathbb{P})$-stable with respect to distance metric $d(\cdot,\cdot)$;

ii. $d(\hat{f}, f) \xrightarrow{p} 0$,

then

$$\|\hat{m} - \tilde{m}\| = \|\tilde{b}(\cdot)\| + o_{\mathbb{P}}(R_n).$$

If further $\|\tilde{b}\| = o_{\mathbb{P}}\left(\sqrt{\mathbb{E}\|\widetilde{m} - m\|^2}\right)$, then $\widehat{m}$ is oracle efficient in the L_2-norm, i.e., asymptotically equivalent to the oracle estimator $\widetilde{m}$ in the sense that

$$\frac{\|\widehat{m} - \widetilde{m}\|}{\sqrt{\mathbb{E}\|\widetilde{m} - m\|^2}} \xrightarrow{p} 0$$

and

$$\|\widehat{m} - m\| = \|\tilde{m} - m\| + o_{\mathbb{P}}(R_n).$$

Proof. Note that

$$\begin{aligned}
\|\widehat{m} - \widetilde{m}\| &\leq \|\widehat{m} - \widetilde{m} - \widetilde{b}\| + \|\widetilde{b}\| \\
&= \|\widetilde{b}\| + o_{\mathbb{P}}\left(\sqrt{\mathbb{E}\|\widetilde{m} - m\|^2}\right)
\end{aligned}$$

where the first line follows by the triangle inequality, and the second by $L_2(\mathbb{P})$-stability and $d(\cdot,\cdot)$-consistency of $\widehat{f}$. ☐

This generalizes Proposition 1 of Kennedy (2022*b*), which shows that a pointwise stability condtion and consistency of $\hat{f}$ implies an oracle inequality on the pointwise convergence of a feasible pseudo-outcome regression. In Section 3, we apply Lemma B.1 to analyze the convergence of our proposed DR-Learners for the target regression bounds under the MOSM.

B.2 Influence function-based estimators

In this section, we state and prove several auxiliary lemmas that are used in the proofs of the main results for analyzing the behavior of our influence function-based estimators of the predictive performance bounds (Section 5).

Lemma B.2. *Let $\beta(\cdot)$ be some function of X_i such that $\|\beta(\cdot)\| \leq M$ for some $M < \infty$ and define the remainder $R^k_{1,n} = \|\hat{\mu}_{1,-k}(\cdot) - \mu_1(\cdot)\|\|\hat{\pi}_{1,-k}(\cdot) - \pi_1(\cdot)\|$. Assume that there exists $\epsilon > 0$ s.t. $\mathbb{P}(\hat{\pi}_{1,-k}(X_i) \geq \epsilon) = 1$. Then,*

$$\mathbb{E}[\beta(X_i)\phi_1(Y_i; \hat{\eta}_{-k}) - \beta(X_i)\phi_1(Y_i; \eta) \mid \mathcal{O}_{-k}] = O_{\mathbb{P}}(R^k_{1,n}).$$

Proof. We follow the proof of Lemma 3 in Mishler, Kennedy and Chouldechova (2021). Suppressing the dependence on $\mathcal{O}_{-k}$ to ease notation, we observe that

$$\mathbb{E}[\beta(X_i)\phi_1(Y_i; \hat{\eta}_{-k}) - \beta(X_i)\phi_1(Y_i; \eta)] =$$

$$\mathbb{E}\left[\beta(X_i)\left(\frac{D_i}{\hat{\pi}_1(X_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{D}{\pi_1(X_i)}(Y_i - \mu_1(X_i)) + (\hat{\mu}_1(X_i) - \mu_1(X_i))\right)\right] \overset{(1)}{=}$$

$$\mathbb{E}\left[\beta(X_i)\left(\frac{\pi_1(X_i)}{\hat{\pi}_1(X_i)}(\mu_1(X_i) - \hat{\mu}_1(X_i)) + (\hat{\mu}_1(X_i) - \mu_1(X_i)\right)\right] =$$

$$\mathbb{E}\left[\beta(X_i)\frac{(\hat{\mu}_1(X_i) - \mu_1(X_i))(\hat{\pi}_1(X_i) - \pi_1(X_i))}{\hat{\pi}_1(X_i)}\right] \overset{(2)}{\leq}$$

$$\frac{1}{\epsilon}\mathbb{E}[\beta(X_i)(\hat{\mu}_1(X_i) - \mu_1(X_i))(\hat{\pi}_1(X_i) - \pi_1(X_i))],$$

where (1) follows by iterated expectations and (2) by the assumption of a bounded propensity score estimator. The result follows by applying the Cauchy-Schwarz inequality and using $\|\beta(\cdot)\| \leq M$ to conclude that $\|\mathbb{E}[\beta(X_i)\phi_1(Y_i; \hat{\eta}) - \beta(X_i)\phi_1(Y_i; \eta)]\| = O_{\mathbb{P}}(R^k_{1,n})$. □

Lemma B.3 (Lemma 2 in Kennedy, Balakrishnan and G'Sell (2020))**.** *Let $\hat{\phi}(X_i)$ be a function estimated from a sample $O_i \sim P(\cdot)$ i.i.d. for $i = 1, \ldots, N$ and let $\mathbb{E}_n[\cdot]$ denote the empirical average over another independent sample $O_j \sim P(\cdot)$ i.i.d. for $j = N+1, \ldots, n$. Then,*

$$\mathbb{E}_n[\hat{\phi}(X_i) - \phi(X_i)] - \mathbb{E}[\hat{\phi}(X_i) - \phi(X_i)] = O_P\left(\frac{\|\hat{\phi}(\cdot) - \phi(\cdot)\|}{\sqrt{n}}\right).$$

Lemma B.4. *Let $\beta(\cdot)$ be some function of X_i such that $\|\beta(\cdot)\| \leq M$ for some $M < \infty$. Let $\hat{\phi}(O_i)$ be a function estimated from a sample $O_i \sim P(\cdot)$ i.i.d. for $i = 1, \ldots, N$ and let $\mathbb{E}_n[\cdot]$ denote the empirical average over another independent sample $O_j \sim P(\cdot)$ i.i.d. for $j = N+1, \ldots, n$. Then,*

$$\mathbb{E}_n[\beta(X_i)\hat{\phi}(O_i) - \beta(X_i)\phi(O_i)] - \mathbb{E}[\beta(X_i)\hat{\phi}(O_i) - \beta(X_i)\phi(O_i)] = O_P\left(\frac{\|\hat{\phi}(\cdot) - \phi(\cdot)\|}{\sqrt{n}}\right).$$

Proof. The proof follows the same argument as the proof of Lemma 2 in Kennedy, Balakrishnan and G'Sell (2020). Observe that, conditional on the estimation sample $\mathcal{O}^{est} = \{O_i\}_{i=1}^N$, $\mathbb{E}\{\mathbb{E}_n[\beta(X_i)(\hat{\phi}(O_i) - \phi(O_i))] \mid \mathcal{O}^{est}\} = \mathbb{E}[\beta(X_i)(\hat{\phi}(O_i) - \phi(O_i)) \mid \mathcal{O}^{est}] = \mathbb{E}[\beta(X_i)(\hat{\phi}(O_i) - \phi(O_i))]$. Next, observe that the conditional variance is $V\left\{(\mathbb{E}_n - \mathbb{E})[\beta(X_i)(\hat{\phi}(O_i) - \phi(O_i))] \mid \mathcal{O}^{est}\right\} = V\left\{\mathbb{E}_n[\beta(X_i)(\hat{\phi}(O_i) - \phi(O_i))] \mid \mathcal{O}^{est}\right\} = n^{-1}V(\beta(X_i)(\hat{\phi}(O_i) - \phi(O_i)) \mid \mathcal{O}^{est}) \leq M\|\hat{\phi}(\cdot) - \phi(\cdot)\|/n$. The result then follows by applying Chebyshev's inequality. □

Lemma B.5 (Convergence of plug-in influence function estimator $\phi_1(Y_i;\hat{\eta})$). *Define the remainder* $\|\hat{\mu}_{1,-k}(\cdot)-\mu_1(\cdot)\|\|\hat{\pi}_{1,-k}(\cdot)-\pi_1(\cdot)\| = R^k_{1,n}$. *Assume (i) there exists* $\delta > 0$ *such that* $\mathbb{P}(\pi_1(X_i)\geq\delta)=1$; *(ii) there exists* $\epsilon > 0$ *such that* $\mathbb{P}(\hat{\pi}_{1-k}(X_i)\geq\epsilon)=1$; *and (iii)* $\|\hat{\mu}_{1,-k}(\cdot)-\mu_1(\cdot)\| = o_P(1)$ *and* $\|\hat{\pi}_{1-k}(\cdot)-\pi_1(\cdot)\| = o_P(1)$. *Then,*

$$\|\phi_1(\cdot;\hat{\eta}_{-k})-\phi_1(\cdot;\eta)\| = O_P(R^k_{1,n}).$$

Proof. This result follows directly from the stated conditions after some algebra. Suppressing dependence on $-k$ to ease notation, observe that we can rewrite

$$\|\phi_1(\cdot;\hat{\eta})-\phi_1(\cdot;\eta)\| =$$

$$\left\|\frac{D_i}{\hat{\pi}_1(X_i)}(Y_i-\hat{\mu}_1(X_i))-\frac{D_i}{\pi_1(X_i)}(Y_i-\mu_1(X_i))+(\mu_1(X_i)-\hat{\mu}_1(X_i))\right\| \overset{(1)}{=}$$

$$\left\|\frac{D_i}{\pi_1(X_i)}\frac{\pi_1(X_i)-\hat{\pi}_1(X_i)}{\hat{\pi}_1(X_i)}(Y_i-\hat{\mu}_1(X_i))-\frac{D_i}{\pi_1(X_i)}(\hat{\mu}_1(X_i)-\mu_1(X_i))+(\hat{\mu}_1(X_i)-\mu_1(X_i))\right\| \overset{(2)}{\leq}$$

$$\left\|\frac{D_i}{\pi_1(X_i)}\frac{\pi_1(X_i)-\hat{\pi}_1(X_i)}{\hat{\pi}_1(X_i)}(Y_i-\mu_1(X_i))\right\|+\left\|\frac{D_i}{\pi_1(X_i)}\frac{\pi_1(X_i)-\hat{\pi}_1(X_i)}{\hat{\pi}_1(X_i)}(\mu_1(X_i)-\hat{\mu}_1(X_i))\right\|+$$

$$\left\|\frac{D_i}{\pi_1(X_i)}(\hat{\mu}_1(X_i)-\mu_1(X_i))\right\|+\|(\hat{\mu}_1(X_i)-\mu_1(X_i))\| \overset{(3)}{\leq}$$

$$\frac{\|D_i\|}{\delta}\frac{\|\pi_1-\hat{\pi}_1\|}{\epsilon}\|Y_i-\mu_1(X_i)\|+\frac{\|D_i\|}{\delta}\frac{\|\pi_1-\hat{\pi}_1\|}{\epsilon}\|\hat{\mu}_1-\mu_1\|+$$

$$\frac{\|D_i\|}{\delta}\|\hat{\mu}_1-\mu_1\|+\|\hat{\mu}_1-\mu_1\| \overset{(4)}{=} o_P(1)+O_P(R^k_{1,n})+o_P(1)+o_P(1)=O_P(R^k_{1,n})$$

where (1) follows by adding and subtracting $\frac{D_i}{\pi_1(X_i)}(Y_i-\hat{\mu}_1(X_i))$, (2) follows by adding and subtracting $\frac{D_i}{\pi_1(X_i)}\frac{\pi_1(X_i)-\hat{\pi}_1(X_i)}{\hat{\pi}_1(X_i)}\mu_1(X_i)$ and applying the triangle inequality, (3) applies the assumption of strict overlap and bounded propensity score estimator, and (4) follows by application of the stated rate conditions. ∎

Lemma B.6 (Convergence of plug-in influence function estimator $\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta})$). *Define the remainder* $\|\hat{\mu}_{1,-k}(\cdot)-\mu_1(\cdot)\|\|\hat{\pi}_{1,-k}(\cdot)-\pi_1(\cdot)\| = R^k_{1,n}$. *Assume that (i) there exists* $\delta > 0$ *such that* $\mathbb{P}(\pi_1(X_i)\geq\delta)=1$; *(ii) there exists* $\epsilon>0$ *such that* $\mathbb{P}(\hat{\pi}_{1,-k}(X_i)\geq\epsilon)=1$; *and (iii)* $\|\hat{\mu}_{1,-k}(\cdot)-\mu_1(\cdot)\| = o_P(1)$ *and* $\|\hat{\pi}_{1,-k}(\cdot)-\pi_1(\cdot)\| = o_P(1)$. *Then,*

$$\|\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k})-\phi(\pi_0(X_i)\mu_1(X_i);\eta)\| = O_{\mathbb{P}}(R^k_{1,n})$$

Proof. This result follows directly from the stated conditions after some simple algebra. For ease of notation, we omit the dependence on X_i and $-k$. Observe that we can rewrite

$$\|\phi(\pi_0(X_i)\mu_1(Y_i);\hat{\eta})-\phi(\pi_0(X_i)\mu_1(Y_i);\hat{\eta})\| =$$

$$\|((1-D_i)-\hat{\pi}_0)\hat{\mu}_1+\frac{D_i}{\hat{\pi}_1}(Y_i-\hat{\mu}_1)\hat{\pi}_0+\hat{\pi}_0\hat{\mu}_1-((1-D_i)-\pi_0)\mu_1-\frac{D_i}{\pi_1}(Y_i-\mu_1)\pi_0-\pi_0\mu_1\| \overset{(1)}{=}$$

$$\|((1-D_i)-\hat{\pi}_0)\hat{\mu}_1+\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(Y_i-\hat{\mu}_1)+\frac{D_i}{\pi_1}\pi_0(\mu_1-\hat{\mu}_1)+\hat{\pi}_0\hat{\mu}_1-\pi_0\mu_1\| \overset{(2)}{\leq}$$

$$\|\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(Y_i-\mu_1)+\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)+\frac{D_i}{\pi_1}\pi_0(\mu_1-\hat{\mu}_1)+((1-D_i)-\hat{\pi}_0)\hat{\mu}_1+\hat{\pi}_0\hat{\mu}_1-\pi_0\mu_1\| \overset{(3)}{=}$$

$$\|\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(Y_i-\mu_1)+\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)+\frac{D_i}{\pi_1}\pi_0(\mu_1-\hat{\mu}_1)+((1-D_i)-\hat{\pi}_0)(\hat{\mu}_1-\mu_1)+(\pi_0-\hat{\pi}_0)\mu_1+\hat{\pi}_0\hat{\mu}_1-\pi_0\mu_1\| \overset{(4)}{\leq}$$

$$\|\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(Y_i-\mu_1)+\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)+\frac{D_i}{\pi_1}\pi_0(\mu_1-\hat{\mu}_1)+((1-D_i)-\pi_0)(\hat{\mu}_1-\mu_1)+(\pi_0-\hat{\pi}_0)(\hat{\mu}_1-\mu_1)+(\pi_0-\hat{\pi}_0)\mu_1\|+$$

$$+\|\hat{\pi}_0\hat{\mu}_1-\pi_0\mu_1\| \overset{(5)}{=}$$

$$\|\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(Y_i-\mu_1)+\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)+\frac{D_i}{\pi_1}\pi_0(\mu_1-\hat{\mu}_1)+((1-D_i)-\pi_0)(\hat{\mu}_1-\mu_1)+(\pi_0-\hat{\pi}_0)(\hat{\mu}_1-\mu_1)+(\pi_0-\hat{\pi}_0)\mu_1\|+$$

$$\|\hat{\pi}_0(\hat{\mu}_1-\mu_1)-\mu_1(\pi_0-\hat{\pi}_0)\|$$

where (1) follows by adding/subtracting $\frac{D_i}{\pi_1}(Y_i-\hat{\mu}_1)$, (2) follows by adding/subtracting $\frac{D_i}{\pi_1}\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}\mu_1$, (3) follows by adding/subtracting $\mu_1((1-D_i)-\hat{\pi}_0)$, (4) follows by adding/subtracting $\pi_0(\hat{\mu}_1-\mu_1)$ and applying the triangle inequality once, and (5) follows by adding/subtracting $\hat{\pi}_0\mu_1$. We then again apply the triangle inequality and use the assumptions of strict overlap and bounded propensity score estimator to arrive at

$$\leq \frac{1}{\epsilon\delta}\|\hat{\pi}_1-\pi_1\|\|Y_i-\mu_1\|+\frac{1}{\epsilon\delta}\|\hat{\pi}_1-\pi_1\|\|\mu_1-\hat{\mu}_1\|+\frac{1-\delta}{\delta}\|\mu_1-\hat{\mu}_1\|+\|(1-D_i)-\pi_0\|\|\hat{\mu}_1-\mu_1\|+$$

$$\|\hat{\pi}_1-\pi_1\|\|\hat{\mu}_1-\mu_1\|+\|\hat{\pi}_1-\pi_1\|\|\mu_1\|+(1-\epsilon)\|\hat{\mu}_1-\mu_1\|+\|\mu_1\|\|\hat{\pi}_1-\pi_1\|.$$

The result then follows by applying the stated rate conditions. □

Lemma B.7. *Let $\beta(\cdot)$ be some function of X_i such that $\|\beta(\cdot)\|\leq M$ for some $M<\infty$ and define the remainder $R^k_{1,n}=\|\hat{\mu}_{1,-k}(\cdot)-\mu_1(\cdot)\|\|\hat{\pi}_{1,-k}(\cdot)-\pi_1(\cdot)\|$. Assume that there exists $\epsilon>0$ s.t. $\mathbb{P}(\hat{\pi}_{1,-k}(X_i)\geq\epsilon)=1$. Then,*

$$\mathbb{E}[\beta(X_i)\,(\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta}_{-k})-\phi(\pi_0(X_i)\mu_1(X_i);\eta))\mid\mathcal{O}_{-k}]=O_{\mathbb{P}}(R^k_{1,n})$$

Proof. For ease of notation, we omit the dependence on X_i and $-k$. The proof follows an analogous argument to Lemma B.2. Observe that

$$\mathbb{E}[\beta(X_i)\,(\phi(\pi_0(X_i)\mu_1(X_i);\hat{\eta})-\phi(\pi_0(X_i,\mu_1(X_i);\eta))]=$$

$$\mathbb{E}[\beta(X_i)\left\{((1-D_i)-\hat{\pi}_0)\hat{\mu}_1+\frac{D_i}{\hat{\pi}_1}(Y_i-\hat{\mu}_1)\hat{\pi}_0+\hat{\pi}_0\hat{\mu}_1-((1-D_i)-\pi_0)\mu_1-\frac{D_i}{\pi_1}(Y_i-\mu_1)\pi_0-\pi_0\mu_1\right\}] \overset{(1)}{=}$$

$$\mathbb{E}[\beta(X_i)\left\{(\pi_0-\hat{\pi}_0)\hat{\mu}_1+\frac{\pi_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)\hat{\pi}_0\right\}+\hat{\pi}_0\hat{\mu}_1-\pi_0\mu_1\}] \overset{(2)}{=}$$

$$\mathbb{E}[\beta(X_i)\left\{(\pi_0-\hat{\pi}_0)\hat{\mu}_1+\frac{\pi_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)\hat{\pi}_0+\hat{\pi}_0(\hat{\mu}_1-\mu_1)+\mu_1(\hat{\pi}_0-\pi_0)\right\}]=$$

$$\mathbb{E}[\beta(X_i)\left\{(\hat{\mu}_1-\mu_1)(\pi_0-\hat{\pi}_0)+\frac{\pi_1-\hat{\pi}_1}{\hat{\pi}_1}(\mu_1-\hat{\mu}_1)\hat{\pi}_0\right\}]$$

where (1) applies iterated expectations, (2) adds/subtracts $\hat{\pi}_0\mu_1$, and the final equality re-arranges. The result then follows by applying the assumption of bounded propensity score estimator and applying the Cauchy-Schwarz inequality. □

Lemma B.8 (Convergence of plug-in influence function estimators for instrumental variable bounds)**.** *Suppose $O_i=(X_i,Z_i,D_i,Y_i)\sim P(\cdot)$ i.i.d. for $i=1,\ldots,n$, where $Z_i\in\mathcal{Z}$ has finite support and satisfies $(Y_i(0),Y_i(1))\perp Z_i\mid X_i$. Define the remainder terms $R^k_{2,n}=\|\hat{\mathbb{E}}_{-k}[D_iY_i\mid X_i,Z_i=z]-\mathbb{E}[D_iY_i\mid X_i,Z_i=x]\|\|\hat{\mathbb{P}}_{-k}(Z_i=z\mid X_i)-\mathbb{P}(Z_i=z\mid X_i)\|$ and $R^k_{3,n}=\|\hat{\pi}_{0,-k}(z,\cdot)-\pi_0(z,\cdot)\|\|\hat{\mathbb{P}}_{-k}(Z_i=z\mid X_i)-\mathbb{P}(Z_i=z\mid X_i)\|$. Assume that (i) there exists $\delta>0$ such that $\mathbb{P}\{\mathbb{P}(Z_i=z\mid X_i)\geq\delta\}=1$; (ii) there exists $\epsilon>0$*

such that $\mathbb{P}\{\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) \geq \epsilon\} = 1$; *(iii)* $\|\hat{\mathbb{E}}_{-k}[D_iY_i \mid X_i, Z_i = z] - \mathbb{E}[D_iY_i \mid X_i, Z_i = z]\| = o_p(1)$, $\|\hat{\pi}_{-k}(\cdot, z) - \pi(\cdot, z)\| = o_P(1)$ *and* $\|\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) - \mathbb{P}(Z_i = z \mid X_i)\| = o_P(1)$. *Then,*

$$\|\phi_z(D_iY_i; \hat{\eta}_{-k})] - \phi_z(D_iY_i; \eta)\| = O_P(R^k_{2,n}),$$
$$\|\phi_z(1 - D_i; \hat{\eta}_{-k}) - \phi_z(1 - D_i; \eta)\| = O_P(R^k_{3,n}).$$

Proof. The proof of this result is analogous to Lemma B.5. To ease notation, we write $\mu_z^{DY}(x) = \mathbb{E}[D_iY_i \mid Z_i = z, X_i = x]$ and $\lambda_z(x) = \mathbb{P}(Z_i = x \mid X_i = x)$ and suppress the dependence on $-k$. We prove the result for $\phi_z(D_iY_i; \eta)$, and the result for $\phi_z(1 - D_i; \hat{\eta})$ follows the same argument. Observe that we can rewrite

$$\|\phi_z(D_iY_i; \hat{\eta}) - \phi_z(D_iY_i; \eta)\| =$$

$$\left\|\frac{1\{Z_i = z\}}{\hat{\lambda}_z(X_i)}(Y_iD_i - \hat{\mu}_z^{DY}(X_i)) - \frac{1\{Z_i = z\}}{\lambda_z(X_i)}(Y_iD_i - \mu_z^{DY}(X_i)) + (\hat{\mu}_z^{DY}(X_i) - \mu_z^{DY}(X_i))\right\| \overset{(1)}{=}$$

$$\left|\frac{1\{Z_i = z\}}{\lambda_z(X_i)}\frac{\lambda_z(X_i) - \hat{\lambda}_z(X_i)}{\hat{\lambda}_z(X_i)}(Y_iD_i - \hat{\mu}_z^{DY}(X_i)) - \frac{1\{Z_i = z\}}{\lambda_z(X_i)}(\hat{\mu}_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i)) + (\hat{\mu}_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i))\right| \overset{(2)}{\leq}$$

$$\left\|\frac{1\{Z_i = z\}}{\lambda_z(X_i)}\frac{\lambda_z(X_i) - \hat{\lambda}_z(X_i)}{\hat{\lambda}_z(X_i)}(D_iY_i - \mu_z^{DY}(X_i))\right\| + \left\|\frac{1\{Z_i = z\}}{\lambda_z(X_i)}\frac{\lambda_z(X_i) - \hat{\lambda}_z(X_i)}{\hat{\lambda}_z(X_i)}(\mu_z^{DY}(X_i) - \hat{\mu}^{DY}(X_i))\right\| +$$

$$\left\|\frac{1\{Z_i = z\}}{\lambda_z(X_i)}(\hat{\mu}_z^{DY}(X_i) - \mu_z^{DY}(X_i)\right\| + \|\hat{\mu}_z^{DY}(X_i) - \mu_z^{DY}(X_i)\| \overset{(3)}{\leq}$$

$$\frac{1}{\delta\epsilon}\|\lambda_z(X_i) - \hat{\lambda}_z(X_i)\|\|D_iY_i - \mu_z^{DY}(X_i)\| + \frac{1}{\delta\epsilon}\|\lambda_z(X_i) - \hat{\lambda}_z(X_i)\|\|\mu_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i)\| +$$

$$\frac{1}{\delta}\|\mu_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i)\| + \|\mu_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i)\|$$

where (1) follows by adding and subtracting $\frac{1\{Z_i=z\}}{\lambda_z(x)}(Y - \hat{\mu}_z^{DY}(x))$, (2) follows by adding and subtracting $\frac{1\{Z_i=z\}}{\lambda_z(X_i)}\frac{\lambda_z(X_i)-\hat{\lambda}_z(X_i)}{\hat{\lambda}_z(X_i)}\mu_z^{DY}(X_i)$ and applying the triangle inequality, (3) applies the Cauchy-Shwarz inequality and the assumptions of strict instrument overlap and bounded instrument propensity estimator. The result then follows from the stated rate conditions. □

Lemma B.9. *Let* $\beta(\cdot)$ *be some function of* X_i *such that* $\|\beta(\cdot)\| \leq M$ *for some* $M < \infty$ *and define the remainder* $R^k_{2,n} = \|\hat{\mathbb{E}}_{-k}[D_iY_i \mid X_i, Z_i = z] - \mathbb{E}[D_iY_i \mid X_i, Z_i = z]\|\|\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) - \mathbb{P}(Z_i = z \mid X_i)\|$ *and* $R^k_{3,n} = \|\hat{\pi}_{0,-k}(\cdot, z) - \pi_0(\cdot, z)\|\|\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) - \mathbb{P}(Z_i = z \mid X_i)\|$. *Assume that there exists* $\epsilon > 0$ *such that* $\mathbb{P}\{\hat{\mathbb{P}}_{-k}(Z_i = z \mid X_i) \geq \epsilon) = 1$. *Then,*

$$\mathbb{E}[\beta(X_i)\phi_z(D_iY_i; \hat{\eta}_{-k}) - \beta(X_i)\phi_z(D_iY_i; \eta) \mid \mathcal{O}_{-k}] = O_{\mathbb{P}}(R^k_{2,n})$$
$$\mathbb{E}[\beta(X_i)\phi_z(1 - D_i; \hat{\eta}_{-k}) - \beta(X_i)\phi_z(1 - D_i; \eta) \mid \mathcal{O}_{-k}] = O_{\mathbb{P}}(R^k_{3,n}).$$

Proof. The proof follows a similar argument as the proof of Lemma B.2. To ease notation, we write $\mu_z^{DY}(x) = \mathbb{E}[D_iY_i \mid Z_i = z, X_i = x]$ and $\lambda_z(x) = \mathbb{P}(Z_i = x \mid X_i = x)$ and suppress the dependence on $-k$. We prove the result for $\phi_z(D_iY_i; \eta)$, and the result for $\phi_z(1 - D_i; \hat{\eta})$ follows the same argument. Observe that

$$\mathbb{E}[\beta(X_i)\phi_z(D_iY_i; \hat{\eta}) - \beta(X_i)\phi_z(D_iY_i; \eta)] =$$

$$\mathbb{E}[\beta(X_i)\left(\frac{1\{Z_i = z\}}{\hat{\lambda}_z(X_i)}(D_iY_i - \hat{\mu}_z^{DY}(X_i)) - \frac{1\{Z_i = z\}}{\lambda_z(X_i)}(D_iY_i - \mu_z^{DY}(X_i)) + (\hat{\mu}_z^{DY}(X_i) - \mu_z^{DY}(X_i))\right)] \overset{(1)}{=}$$

$$\mathbb{E}[\beta(X_i)\left(\frac{\lambda_z(X_i)}{\hat{\lambda}_z(X_i)}(\mu_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i)) + (\hat{\mu}_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i))\right)] =$$

$$\mathbb{E}[\beta(X_i)\frac{(\hat{\mu}_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i))(\lambda_z(X_i) - \hat{\lambda}_z(X_i))}{\hat{\pi}_1(X_i)}] \overset{(2)}{\leq}$$

$$\frac{1}{\epsilon}\mathbb{E}[\beta(X_i)(\hat{\mu}_z^{DY}(X_i) - \hat{\mu}_z^{DY}(X_i))(\lambda_z(X_i) - \hat{\lambda}_z(X_i))]$$

where (1) follows by iterated expectations and (2) by the assumption of bounded instrument propensity estimator. The result then follows by applying the Cauchy-Schwarz inequality and using $\|\beta(\cdot)\| \leq M$. □

C Additional theoretical results

In this section, we state and prove various additional theoretical results that are discussed briefly in the main text.

C.1 Variance estimation for bounds on overall predictive disparities

We now develop a consistent estimator of the asymptotic covariance matrix of our estimators of the overall predictive performance bounds. Recall from the statement and proof of Theorem 5.1, if $R_{1.,n} = o_{\mathbb{P}}(1/\sqrt{n})$, then

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\overline{\text{perf}}}(s;\beta,\Delta)\\ \widehat{\underline{\text{perf}}}(s;\beta,\Delta)\end{pmatrix} - \begin{pmatrix}\overline{\text{perf}}(s;\beta,\Delta)\\ \underline{\text{perf}}(s;\beta,\Delta)\end{pmatrix}\right) \xrightarrow{d} N\left(0,\Sigma\right),$$

where $\Sigma = Cov\left((\overline{\text{perf}}_i, \underline{\text{perf}}_i)'\right)$ for $\overline{\text{perf}}_i = \beta_{0,i} + \beta_{1,i}(1-D_i)\left(1\{\beta_{1,i} > 0\}\overline{\delta}_i + 1\{\beta_{1,i} \leq 0\}\underline{\delta}_i\right) + \beta_{1,i}\phi_1(Y_i;\eta)$ and $\underline{\text{perf}}_i = \beta_{0,i} + \beta_{1,i}(1-D_i)\left(1\{\beta_{1,i} > 0\}\underline{\delta}_i + 1\{\beta_{1,i} \leq 0\}\overline{\delta}_i\right) + \beta_{1,i}\phi_1(Y_i;\eta)$ and $\mathbb{E}[\overline{\text{perf}}_i] = \overline{\text{perf}}(s;\beta,\Delta)$, $\mathbb{E}[\underline{\text{perf}}_i] = \underline{\text{perf}}(s;\beta,\Delta)$.

Consider the following estimator of the asymptotic covariance matrix

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\widehat{\overline{\text{perf}}}(O_i;\hat{\eta}_{-K_i}) - \widehat{\overline{\text{perf}}}(s;\beta,\Delta)\\ \widehat{\underline{\text{perf}}}(O_i;\hat{\eta}_{-K_i}) - \widehat{\underline{\text{perf}}}(s;\beta,\Delta)\end{pmatrix}\begin{pmatrix}\widehat{\overline{\text{perf}}}(O_i;\hat{\eta}_{-K_i}) - \widehat{\overline{\text{perf}}}(s;\beta,\Delta)\\ \widehat{\underline{\text{perf}}}(O_i;\hat{\eta}_{-K_i}) - \widehat{\underline{\text{perf}}}(s;\beta,\Delta)\end{pmatrix}'.$$

To show that $\widehat{\Sigma} \xrightarrow{p} \Sigma$, it suffices to show convergence in probability for each entry. We prove this directly by establishing the following Lemma, which extends Lemma 1 in Dorn, Guo and Kallus (2021).

Lemma C.1. *Let ϕ_1, ϕ_2 be any two square integrable functions. Let $\hat{\phi}_{1,n} = \left(\hat{\phi}_1(O_i), \ldots, \hat{\phi}_1(O_n)\right)$, $\hat{\phi}_{2,n} = \left(\hat{\phi}_2(O_i), \ldots, \hat{\phi}_2(O_n)\right)$ be random vectors satisfying*

$$\|\hat{\phi}_{1,n} - \phi_{1,n}\|_{L_2(\mathbb{P}_n)} := \sqrt{n^{-1}\sum_{i=1}^{n}(\hat{\phi}_1(O_i) - \phi_1(O_i))^2} = o_{\mathbb{P}}(1),$$

$$\|\hat{\phi}_{2,n} - \phi_{2,n}\|_{L_2(\mathbb{P}_n)} := \sqrt{n^{-1}\sum_{i=1}^{n}(\hat{\phi}_2(O_i) - \phi_2(O_i))^2} = o_{\mathbb{P}}(1),$$

where $\phi_{1,n} = (\phi_1(O_i), \ldots, \phi_1(O_n))$ and $\phi_{2,n} = (\phi_2(O_i), \ldots, \phi_2(O_n))$. Define $\mathbb{P}_n$ to be the empirical distribution. Then, the second moments of $\mathbb{P}_n$ converge in probability to the respective second moments of $(\phi_1(O_i), \phi_2(O_i)) \sim P$

Proof. Let $\hat{\phi}_{i,1} = \hat{\phi}_i(O_i)$ and define $\phi_{i,1}, \hat{\phi}_{i,1}, \phi_{i,2}$ analogously. To prove this result, we first show that

$\mathbb{E}_n[\hat{\phi}_{1,i}^2] = \mathbb{E}[\phi_{1,i}^2] + o_{\mathbb{P}}(1)$ since the same argument applies for $\phi_{2,i}$. Observe that

$$n^{-1}\sum_{i=1}^{n}\hat{\phi}_{1,i} - \mathbb{E}[\phi_{i,1}^2] = n^{-1}\sum_{i=1}^{n}(\hat{\phi}_{i,1}^2 - \phi_{i,1}^2) + (\mathbb{E}_n - \mathbb{E})[\phi_{i,1}^2],$$

where $(\mathbb{E}_n - \mathbb{E})[\phi_{i,1}^2] = o_{\mathbb{P}}(1)$. Furthermore, we can rewrite the first term as

$$n^{-1}\sum_{i=1}^{n}(\hat{\phi}_{i,1}^2 - \phi_{i,1}^2) = n^{-1}\sum_{i=1}^{n}\left(\hat{\phi}_{i,1} - \phi_{i,1}\right)\left(\hat{\phi}_{i,1} + \phi_{i,1}\right) =$$

$$n^{-1}\sum_{i=1}^{n}\left(\hat{\phi}_{i,1} - \phi_{i,1}\right)\left(\hat{\phi}_{i,1} - \phi_{i,1} + 2\phi_{i,1}\right) \leq \|\hat{\phi}_{i,n} - \phi_{i,1}\|\left(\|\hat{\phi}_{i,1} - \phi_{i,1}\| + 2\|\phi_{i,1}\|\right) = o_{\mathbb{P}}(1),$$

where the last inequality applies the Cauchy-Schwarz inequality and triangle inequality. We next show that $\mathbb{E}_n[\hat{\phi}_{i,1}\hat{\phi}_{i,2}] = \mathbb{E}[\phi_{i,1}\phi_{i,2}] + o_{\mathbb{P}}(1)$. Observe that

$$n^{-1}\sum_{i=1}^{n}\hat{\phi}_{i,1}\hat{\phi}_{i,2} - \mathbb{E}[\phi_{i,1}\phi_{i,2}] = n^{-1}\sum_{i=1}^{n}\left(\hat{\phi}_{i,1}\hat{\phi}_{i,2} - \phi_{i,1}\phi_{i,2}\right) + (\mathbb{E}_n - \mathbb{E})[\phi_{i,1}\phi_{i,2}],$$

where $(\mathbb{E}_n - \mathbb{E})[\phi_{i,1}\phi_{i,2}] = o_{\mathbb{P}}(1)$. We can further rewrite the first term as

$$n^{-1}\sum_{i=1}^{n}\left(\hat{\phi}_{i,1}\hat{\phi}_{i,2} - \phi_{i,1}\phi_{i,2}\right) = n^{-1}\sum_{i=1}^{n}\left(\hat{\phi}_{i,1}(\hat{\phi}_{i,2} - \phi_{i,2}) + \phi_{i,2}(\hat{\phi}_{i,1} - \phi_{i,1})\right) =$$

$$n^{-1}\sum_{i=1}^{n}\phi_{i,1}(\hat{\phi}_{i,2} - \phi_{i,2}) + n^{-1}\sum_{i=1}^{n}(\hat{\phi}_{i,1} - \phi_{i,1})(\hat{\phi}_{i,2} - \phi_{i,2}) + n^{-1}\sum_{i=1}^{n}\phi_{i,2}(\hat{\phi}_{i,1} - \phi_{i,1}) \leq$$

$$\|\phi_{1,n}\|\|\hat{\phi}_{2,n} - \phi_{2,n}\| + \|\hat{\phi}_{1,n} - \phi_{1,n}\|\|\hat{\phi}_{2,n} - \phi_{2,n}\| + \|\phi_{2,n}\|\|\hat{\phi}_{1,n} - \phi_{1,n}\| = o_{\mathbb{P}}(1),$$

where the last inequality applies Cauchy-Schwarz inequality. □

We show that the conditions of Lemma C.1 are satisfied for $\widehat{\overline{\text{perf}}}(O_i;\hat{\eta}_{-K_i})$ and $\widehat{\underline{\text{perf}}}(O_i;\hat{\eta}_{-K_i})$. The convergence of probability of the sample estimator $\hat{\Sigma}$ then follows immediately by the continuous mapping theorem since we already established the convergence of the first moments in Theorem 5.1 provided we show that $\|\widehat{\overline{\text{perf}}}_n^k - \overline{\text{perf}}_n^k\|^2$, and $\|\widehat{\underline{\text{perf}}}_n^k - \underline{\text{perf}}_n^k\|^2$ are $o_{\mathbb{P}}(1)$ for each fold k.

Lemma C.2. *Under the same assumptions as Theorem 5.1, for each fold k,*

$$\|\widehat{\overline{perf}}_i - \overline{perf}_i\|_{L_2(\mathbb{P}_n^k)} = o_{\mathbb{P}}(1)$$
$$\|\widehat{\underline{perf}}_i - \underline{perf}_i\|_{L_2(\mathbb{P}_n^k)} = o_{\mathbb{P}}(1)$$

conditionally on $\mathcal{O}_{-k}$.

Proof. We prove the result for $\widehat{\overline{\text{perf}}}_n$ since the analogous argument applies for $\widehat{\underline{\text{perf}}}_n$. Following the proof of Lemma B.5, we observe that

$$\|\widehat{\overline{\text{perf}}}_i - \overline{\text{perf}}_i\|_{L_2(\mathbb{P}_n^k)} \leq \frac{\|D_i\|_{L_2(\mathbb{P}_n^k)}}{\delta}\frac{\|\pi_1 - \hat{\pi}_1\|_{L_2(\mathbb{P}_n^k)}}{\epsilon}\|Y_i - \mu_1(X_i)\|_{L_2(\mathbb{P}_n^k)} +$$

$$\frac{\|D_i\|_{L_2(\mathbb{P}_n^k)}}{\delta}\frac{\|\pi_1 - \hat{\pi}_1\|_{L_2(\mathbb{P}_n^k)}}{\epsilon}\|\hat{\mu}_1 - \mu_1\|_{L_2(\mathbb{P}_n^k)} + \frac{\|D_i\|_{L_2(\mathbb{P}_n^k)}}{\delta}\|\hat{\mu}_1 - \mu_1\|_{L_2(\mathbb{P}_n^k)} + \|\hat{\mu}_1 - \mu_1\|_{L_2(\mathbb{P}_n^k)} =$$

$$O_p(\|\pi_1 - \hat{\pi}_1\|_{L_2(\mathbb{P})}\|Y_i - \mu_1(X_i)\|_{L_2(\mathbb{P})} + \|\pi_1 - \hat{\pi}_1\|_{L_2(\mathbb{P})}\|\hat{\mu}_1 - \mu_1\|_{L_2(\mathbb{P})} + \|\hat{\mu}_1 - \mu_1\|_{L_2(\mathbb{P})}).$$

where the last line applies Markov's Inequality. The result is then immediate. □

By a straightforward extension, we can develop consistent estimators of the asymptotic covariance matrix under nonparametric outcome bounds and instrumental variable bounds as well.

C.2 Bounding predictive disparities under the MOSM

As mentioned in the main text, we can further bound the predictive disparities of a given risk assessment under the MOSM. Recall that now $X_i = (G_i, \bar{X}_i)'$. First consider the overall predictive disparity, and observe that, for $g \in \{0, 1\}$,

$$\text{perf}_g(s; \beta) = \mathbb{E}[\beta_{0,i} + \beta_{1,i}\mu_1(X_i) + \beta_{1,i}\pi_0(X_i)\delta(X_i) \mid G_i = g] =$$

$$P(G_i = g)^{-1}\mathbb{E}[\beta_{0,i}1\{G_i = g\} + \beta_{1,i}1\{G_i = g\}\mu_1(X_i) + \beta_{1,i}1\{G_i = g\}\pi_0(X_i)\delta(X_i)].$$

where $\alpha_g = P(G_i = g)$. Therefore $\text{disp}(s; \beta)$ can be equivalently written as

$$\alpha_1^{-1}\mathbb{E}[\beta_{0,i}G_i + \beta_{1,i}G_i\mu_1(X_i) + \beta_{1,i}G_i\pi_0(X_i)\delta(X_i)] - \alpha_0^{-1}\mathbb{E}[\beta_{0,i}(1-G_i) + \beta_{1,i}(1-G_i)\mu_1(X_i) + \beta_{1,i}(1-G_i)\pi_0(X_i)\delta(X_i)].$$

Since this is a linear function δ, we can immediately obtain sharp bounds. In contrast, for the positive-class predictive disparity, we provide non-sharp bounds since the positive-class predictive-disparity can only be expressed as the difference of two linear-fractional functions in $\delta(\cdot)$.

Lemma C.3. *Define $\mathcal{H}(disp(s; \beta); \Delta)$ to be the set of all overall predictive disparities that are consistent with the MOSM. To ease notation, let $\beta^g_{0,i} = \beta_{0,i}/P(G_i = g)$, $\beta^g_{1,i} = \beta_{1,i}/P(G_i = g)$ for $g \in \{0, 1\}$, and $\tilde{\beta}_{0,i} = \beta^1_{0,i} - \beta^0_{0,i}$, $\tilde{\beta}_{1,i} = \beta^1_{1,i} - \beta^0_{1,i}$. Under Assumption 2.1,*

$$\mathcal{H}(disp(s; \beta); \Delta) = [\underline{disp}(s; \beta, \Delta), \overline{disp}(s; \beta, \Delta)],$$

where

$$\overline{disp}(s; \beta, \Delta) := \mathbb{E}[\tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\mu_1(X_i) + \tilde{\beta}_{1,i}\pi_0(X_i)(\overline{\nu}_i\overline{\delta}_i + \underline{\nu}_i\underline{\delta}_i)]$$
$$\underline{disp}(s; \beta, \Delta) := \mathbb{E}[\tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\mu_1(X_i) + \tilde{\beta}_{1,i}\pi_0(X_i)(\overline{\nu}_i\underline{\delta}_i + \underline{\nu}_i\overline{\delta}_i)].$$

for $\overline{\nu}_i = G_i1\{\beta_{1,i} \geq 0\} + (1 - G_i)1\{\beta_{1,i} \leq 0\}$ and $\underline{\nu}_i = G_i1\{\beta_{1,i} < 0\} + (1 - G_i)1\{\beta_{i,1} > 0\}$.

Proof. To prove the result, we notice that $\text{disp}(s; \beta)$ can be rewritten as

$$\mathbb{E}[\tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\mu_1(X_i) + \tilde{\beta}_{1,i}\pi_0(X_i)\delta(X_i)]$$

using the definitions of $\tilde{\beta}_{0,i}, \tilde{\beta}_{1,i}$. Following the same logic as Lemma 2.1 in the main text, it then follows immediately that $\mathcal{H}(\text{disp}(s; \beta); \Delta)$ equals the closed interval

$$[\tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\mu_1(X_i) + \tilde{\beta}_{1,i}\pi_0(X_i)\left(1\{\tilde{\beta}_{1,i} \geq 0\}\underline{\delta}_i + 1\{\tilde{\beta}_{1,i} < 0\}\overline{\delta}_i\right),$$

$$\tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\mu_1(X_i) + \tilde{\beta}_{1,i}\pi_0(X_i)\left(1\{\tilde{\beta}_{1,i} \geq 0\}\overline{\delta}_i + 1\{\tilde{\beta}_{1,i} < 0\}\underline{\delta}_i\right)].$$

The result then follows by noticing that

$$1\{\tilde{\beta}_{1,i} \geq 0\} = 1\{(G_i - p)\beta_{i,1} \geq 0\} = G_i1\{\beta_{1,i} \geq 0\} + (1 - G_i)1\{\beta_{1,i} \leq 0\}.$$

$$1\{\tilde{\beta}_{1,i} < 0\} = 1\{(G_i - p)\beta_{i,1} < 0\} = G_i1\{\beta_{1,i} < 0\} + (1 - G_i)1\{\beta_{i,1} > 0\}.$$

□

Lemma C.4. *Define $\mathcal{H}(disp_+(s;\beta);\Delta)$ to be the set of all positive-class disparities that are consistent with the MOSM. Under Assumption 2.1,*

$$\mathcal{H}(disp_+(s;\beta);\Delta) \subseteq [\underline{disp}_+(s;\beta,\Delta), \overline{disp}_+(s;\beta,\Delta)],$$

where $\overline{disp}_+(s;\beta,\Delta) = \overline{perf}_{+,1}(s;\beta) - \underline{perf}_{+,0}(s,\beta)$, $\underline{disp}_+(s;\beta,\Delta) = \underline{perf}_{+,1}(s;\beta) - \overline{perf}_{+,0}(s,\beta)$ *for,* $g \in \{0,1\}$,

$$\overline{perf}_{+,g}(s;\beta) = \sup_{\delta\in\Delta} \frac{\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i) \mid G_i = g]}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i) \mid G_i = g]},$$
$$\underline{perf}_{+,g}(s;\beta) = \inf_{\delta\in\Delta} \frac{\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i) \mid G_i = g]}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i) \mid G_i = g]}.$$

Proof. Observe that, for $g \in \{0,1\}$,

$$\text{perf}_{+,g}(s;\beta,\delta) = \frac{\mathbb{E}[\beta_{0,i}\mu_1(X_i) + \beta_{0,i}\pi_0(X_i)\delta(X_i) \mid G_i = g]}{\mathbb{E}[\mu_1(X_i) + \pi_0(X_i)\delta(X_i) \mid G_i = g]},$$

and so the positive-class predictive disparity $\text{disp}_+(s;\beta)$ can be written as $\text{disp}_+(s;\beta,\delta) = \text{perf}_{+,1}(s;\beta,\delta) - \text{perf}_{+,0}(s;\beta,\delta)$. The result then follows since

$$\sup_{\delta\in\Delta} \text{disp}_+(s;\beta,\delta) \le \sup_{\delta\in\Delta} \text{perf}_{+,1}(s;\beta,\delta) - \inf_{\delta\in\Delta} \text{perf}_{+,0}(s;\beta,\delta)$$

and

$$\inf_{\delta\in\Delta} \text{disp}_+(s;\beta,\delta) \ge \inf_{\delta\in\Delta} \text{perf}_{+,1}(s;\beta,\delta) - \sup_{\delta\in\Delta} \text{perf}_{+,0}(s;\beta,\delta).$$

□

C.3 Estimating bounds on overall predictive disparities

We construct estimators for the bounds on the overall predictive disparities under the MOSM, $\underline{\text{disp}}(s;\beta,\Delta)$ and $\overline{\text{disp}}(s;\beta,\Delta)$. We develop the estimators assuming that $\mathbb{P}(G_i = 1)$ is known, but they can be easily extended to the case where this is estimated. We also develop the estimators assuming that the bounding functions $\underline{\delta}(\cdot), \overline{\delta}(\cdot)$ in the MOSM are known. The extensions to the cases of estimated nonparametric outcome regression bounds and estimated instrumental variable bounds are straightforward in light of the results in Section 5.1.

We make use of K-fold cross-fitting. For each fold k, we construct estimators of the nuisance functions $\hat{\eta}{-k} = (\hat{\pi}_{1,-k}, \hat{\mu}_{1,-k})$ using only the sample of observations $\mathcal{O}_{-k}$ not in the k-th fold. For each observation in the k-th fold, we construct

$$\overline{\text{disp}}(O_i;\hat{\eta}_{-k}) := \tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \tilde{\beta}_{1,i}(1-D_i)(\overline{\nu}_i\overline{\delta} + \underline{\nu}_i\underline{\delta}_i), \tag{27}$$
$$\overline{\text{disp}}(O_i;\hat{\eta}_{-k}) := \tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\phi_1(Y_i;\hat{\eta}_{-k}) + \tilde{\beta}_{1,i}(1-D_i)(\overline{\nu}_i\underline{\delta} + \underline{\nu}_i\overline{\delta}_i). \tag{28}$$
$$\tag{29}$$

We then estimate the upper bound on overall predictive disparities under the MOSM by taking the average across all units in the historical data $\widehat{\overline{\text{disp}}}(s;\beta;\Delta) := \mathbb{E}_n[\overline{\text{disp}}(O_i;\hat{\eta}_{-K_i})]$ and $\widehat{\underline{\text{disp}}}(s;\beta,\Delta) := \mathbb{E}_n[\underline{\text{disp}}(O_i;\hat{\eta}_{-K_i})]$. Algorithm 4 summarizes our proposed estimators for the overall predictive disparity bounds under the MOSM and their associated standard errors.

By the same argument as the proof of Theorem 5.1, we can derive the rate of convergence of our proposed estimators and provide conditions under which they are jointly asymptotically normal.

Algorithm 4: Pseudo-algorithm for overall predictive disparity bounds estimators.

Input: Data $\mathcal{O} = \{(O_i)\}_{i=1}^n$ where $O_i = (X_i, D_i, Y_i)$, number of folds K.

for $k = 1, \ldots, K$ **do**

  Estimate $\hat{\eta}_{-k} = (\hat{\pi}_{1,-k}, \hat{\mu}_{1,-k})$.

  Set $\overline{\text{disp}}(O_i; \hat{\eta}_{-K(i)})$ and $\underline{\text{disp}}(O_i; \hat{\eta}_{-K(i)})$ for all $i \in \mathcal{O}_k$.

; Set $\widehat{\overline{disp}}(s;\beta,\Delta) = \mathbb{E}_n[\overline{\text{disp}}(O_i; \hat{\eta}_{-K(i)})]$, $\widehat{\underline{\text{disp}}}(s;\beta,\Delta) = \mathbb{E}_n[\underline{\text{disp}}(O_i; \hat{\eta}_{-K(i)})]$;

Set $\hat{\sigma}_{i,11} = (\overline{\text{disp}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\overline{\text{disp}}}(s;\beta,\Delta))^2$,

$\hat{\sigma}_{i,12} = (\overline{\text{disp}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\overline{\text{disp}}}(s;\beta,\Delta))(\underline{\text{disp}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\underline{\text{disp}}}(s;\beta,\Delta))$, and

$\hat{\sigma}_{i,22} = (\underline{\text{disp}}(O_i; \hat{\eta}_{-K(i)}) - \widehat{\underline{\text{disp}}}(s;\beta,\Delta))^2$;

Output: Estimates $\widehat{\overline{\text{disp}}}(s;\beta,\Delta) = \mathbb{E}_n[\overline{\text{disp}}(O_i; \hat{\eta}_{-K(i)})]$, $\widehat{\underline{\text{disp}}}(s;\beta,\Delta) = \mathbb{E}_n[\underline{\text{disp}}(O_i; \hat{\eta}_{-K(i)})]$.

Output: Estimated covariance matrix $n^{-1} \sum_{i=1}^n \begin{pmatrix} \hat{\sigma}_{i,11} & \hat{\sigma}_{i,12} \\ \hat{\sigma}_{i,12} & \hat{\sigma}_{i,22} \end{pmatrix}$

Proposition C.1. *Under the same assumptions as Theorem 5.1,*

$$|\widehat{\overline{perf}}(s;\beta,\Delta) - \overline{perf}(s;\beta,\Delta)| = O_{\mathbb{P}}(1/\sqrt{n} + \sum_{k=1}^K R_{1,n}^k)$$

$$|\widehat{\underline{perf}}(s;\beta,\Delta) - \underline{perf}(s;\beta,\Delta)| = O_{\mathbb{P}}(1/\sqrt{n} + \sum_{k=1}^K R_{1,n}^k).$$

If further $R_{1,n}^k = o_{\mathbb{P}}(1/\sqrt{n})$ for all folds k, then

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\overline{disp}}(s;\beta,\Delta) \\ \widehat{\underline{disp}}(s;\beta,\Delta) \end{pmatrix} - \begin{pmatrix} \widehat{\overline{disp}}(s;\beta,\Delta) \\ \widehat{\underline{disp}}(s;\beta,\Delta) \end{pmatrix} \right) \xrightarrow{N} (0, \Sigma)$$

for covariance matrix $\Sigma = Cov\left((\overline{disp}_i, \underline{disp}_i) \right)$ where $\overline{disp}_i = \tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\phi_1(Y_i;\eta) + \tilde{\beta}_{1,i}(1-D_i)(\overline{\nu}_i\overline{\delta} + \underline{\nu}_i\underline{\delta}_i)$ and $\underline{disp}_i = \tilde{\beta}_{0,i} + \tilde{\beta}_{1,i}\phi_1(Y_i;\eta) + \tilde{\beta}_{1,i}(1-D_i)(\overline{\nu}_i\underline{\delta} + \underline{\nu}_i\overline{\delta}_i)$.

As in the main text, we can analogously extend our estimators for the bounds on overall predictive disparities under the MOSM to the case with estimated bounding functions (e.g., nonparametric outcome regression bounds and instrumental variable bounds). Since this merely involves replacing plugging in an estimator for the appropriate uncentered efficient influence function into the estimator defined above, we skip providing the details.

C.4 Estimating bounds on positive-class predictive disparities

We now construct estimators for the bounds on positive-class predictive disparities under the MOSM, $\underline{\text{disp}}_+(s;\beta,\Delta)$ and $\overline{\text{disp}}_+(s;\beta,\Delta)$. To do so, we develop our estimator for the group-specific positive-class performance bounds $\overline{\text{perf}}_{+,g}(s;\beta,\Delta)$ and $\underline{\text{perf}}_{+,g}(s;\beta,\Delta)$.

We once again make use of K-fold cross-fitting. For each fold $k = 1, \ldots, K$, we construct estimators of the nuisance functions $\hat{\eta}_{-k}$ using only the sample of observations $\mathcal{O}_{-k}$. We then construct a fold-specific estimate of the upper bound for group g by solving

$$\widehat{\overline{\text{perf}}}^k_{+,g}(s;\beta,\Delta_n) := \max_{\tilde{\delta} \in \Delta_n} \frac{\mathbb{E}_n^k[1\{G_i = g\} \left(\beta_{0,i}\phi_1(Y_i; \hat{\eta}_{-k}) + \beta_{0,i}(1-D_i)\tilde{\delta}_i \right)]}{\mathbb{E}_n^k[1\{G_i = g\} \left(\phi_1(Y_i; \hat{\eta}_{-k}) + (1-D_i)\tilde{\delta}_i \right)]}. \tag{30}$$

The estimator then averages the fold-specific estimates $\widehat{\overline{\text{perf}}}_{+,g}(s;\beta,\Delta_n) = K^{-1}\sum_{k=1}^{K}\widehat{\overline{\text{perf}}}^{k}_{+,g}(s;\beta,\Delta_n)$, and $\widehat{\underline{\text{perf}}}_{+,g}(s;\beta,\Delta)$ is defined analogously. We then estimate the bounds on positive-class predictive disparities by

$$\widehat{\overline{\text{disp}}}_{+}(s;\beta,\Delta) = \widehat{\overline{\text{perf}}}_{+,1}(s;\beta,\Delta) - \widehat{\underline{\text{perf}}}_{+,1}(s;\beta,\Delta),$$
$$\widehat{\underline{\text{disp}}}_{+}(s;\beta,\Delta) = \widehat{\underline{\text{perf}}}_{+,1}(s;\beta,\Delta) - \widehat{\overline{\text{perf}}}_{+,1}(s;\beta,\Delta).$$

To analyze the rate of convergence of $\widehat{\overline{\text{disp}}}_{+}(s;\beta,\Delta))$, we first notice that

$$\|\widehat{\overline{\text{perf}}}_{+,g}(s;\beta,\Delta) - \overline{\text{perf}}_{+,g}(s;\beta,\Delta)\| = O_{\mathbb{P}}(1/\sqrt{n} + \sum_{k=1}^{K} R^{K}_{1,n})$$

by the same argument as the proof of Theorem 5.2. The following result is then an immediate consequence.

Proposition C.2. *Under the same assumptions as Theorem 5.2,*

$$\|\widehat{\overline{disp}}(s;\beta,\Delta) - \overline{disp}(s;\beta,\Delta)\| = O_{\mathbb{P}}(1/\sqrt{n} + \sum_{k=1}^{K} R^{K}_{1,n}),$$
$$\|\widehat{\underline{disp}}(s;\beta,\Delta) - \underline{disp}(s;\beta,\Delta)\| = O_{\mathbb{P}}(1/\sqrt{n} + \sum_{k=1}^{K} R^{K}_{1,n}).$$

D Additional Monte Carlo simulations and empirical results

In this section, we report additional Monte Carlo simulations that examine the performance of our proposed estimators for robust audits.

D.1 Monte Carlo simulations: bounds on true positive rate and false positive rate

Under the same simulation design as described in Section 6.2 of the main text, we now report the performance of our estimators for the bounds on the true positive rate $\text{perf}_{+}(s;\beta_{tpr})$ and the false positive rate $\text{perf}_{-}(s;\beta_{fpr})$

We first audit the true positive rate of the risk score for a fixed choice $\underline{\Gamma} = 2/3$, $\overline{\Gamma} = 3/2$, and evaluate how well our proposed estimators recover the true bounds $[\underline{\text{perf}}_{+}(s;\beta_{tpr},\Delta(\Gamma)), \overline{\text{perf}}_{+}(s;\beta_{tpr},\Gamma)]$ and $[\underline{\text{perf}}_{-}(s;\beta_{fpr},\Delta(\Gamma)), \overline{\text{perf}}_{-}(s;\beta_{fpr},\Gamma)]$. Across 1,000 simulated evaluation datasets of varying size $n \in \{500, 1000, 1500\}$, we calculate the estimates $[\widehat{\underline{\text{perf}}}_{+}(s;\beta_{tpr},\Delta(\Gamma)), \widehat{\overline{\text{perf}}}_{+}(s;\beta_{tpr},\Delta(\Gamma))]$ and $[\widehat{\underline{\text{perf}}}_{-}(s;\beta_{fpr},\Delta(\Gamma)), \widehat{\overline{\text{perf}}}_{-}(s;\beta_{fpr},\Delta(\Gamma))]$. The estimators are constructed using single split of the evaluation data, and we estimate the first-stage nuisance functions $\eta = (\pi_1(X_i), \mu_1(X_i))$ using random forests. Across simulations, we report the average bias of our estimators for the bounds on the true positive rate in Table A1 and the false positive rate in Table A2. As the size of the evaluation data grows larger, the average bias of our estimators of the bounds quickly decline in magnitude, illustrating Corollary 5.1.

We next illustrate how our proposed estimators can be used to conduct sensitivity analyses on the overall performance of the risk score under alternative assumptions on the strength of unmeasured confounding. We now set $\underline{\Gamma} = 1/\widetilde{\Gamma}$, $\overline{\Gamma} = \widetilde{\Gamma}$ for $\widetilde{\Gamma} \geq 1$, and report results varying $\widetilde{\Gamma} \in \{1, \ldots, 2.5\}$. For each choice of $\widetilde{\Gamma}$, we again simulate 1,000 evaluation datasets of size $n = 1000$ and calculate estimates $[\widehat{\underline{\text{perf}}}_{+}(s;\beta_{tpr},\Delta(\Gamma)), \widehat{\overline{\text{perf}}}_{+}(s;\beta_{tpr},\Delta(\Gamma))]$ and $[\widehat{\underline{\text{perf}}}_{-}(s;\beta_{fpr},\Delta(\Gamma)), \widehat{\overline{\text{perf}}}_{-}(s;\beta_{fpr},\Delta(\Gamma))]$. Figure A1 plots the distribution of our estimators across simulations (red, box plots) against the true bounds (black,

n	$\widehat{\overline{\text{perf}}}_+(s;\beta_{tpr},\Delta(\Gamma))$	SD of $\widehat{\overline{\text{perf}}}_+(s;\beta_{tpr},\Delta(\Gamma))$	Bias
500	0.788	0.024	-0.056
1000	0.815	0.018	-0.030
1500	0.829	0.014	-0.015

(a) Upper bound on true positive rate

n	$\widehat{\overline{\text{perf}}}_+(s;\beta_{tpr},\Delta(\Gamma))$	SD of $\widehat{\overline{\text{perf}}}_+(s;\beta_{tpr},\Delta(\Gamma))$	Bias
500	0.233	0.020	0.022
1000	0.205	0.014	-0.005
1500	0.190	0.012	-0.020

(b) Lower bound on true positive rate

Table A1: Bias properties of estimators for the bounds on the true positive rate of a risk score $s(\cdot)$ with nonparametric outcome bounds

Notes: This table summarizes the average bias of our estimators of the bounds on the true positive rate $\widehat{\underline{\text{perf}}}_+(s;\beta_{tpr})$, $\widehat{\overline{\text{perf}}}_+(s;\beta_{tpr})$, and the standard deviation of our estimators across simulations. We report these results for $n \in \{500, 1000, 1500\}$. The positive class performance estimators are constructed using a single sample split, and the nuisance functions are estimated using random forests. The results are computed over 1,000 simulations. See Section 6.2 for further details on the simulation design.

Model	n	$\widehat{\overline{\text{perf}}}_-(s;\beta_{fpr},\Delta(\Gamma))$	SD of $\widehat{\overline{\text{perf}}}_-(s;\beta_{fpr},\Delta(\Gamma))$	Bias
logistic	500	0.781	0.026	-0.054
	1000	0.810	0.019	-0.025
	1500	0.825	0.016	-0.009
random forest	500	0.772	0.025	-0.063
	1000	0.802	0.019	-0.033
	1500	0.818	0.014	-0.017

(a) Upper bound on false positive rate

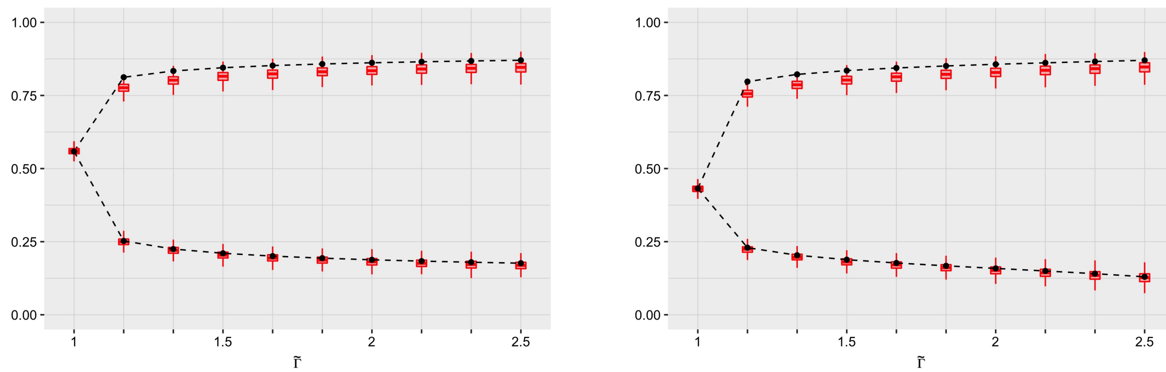
Model	n	$\widehat{\overline{\text{perf}}}_-(s;\beta_{fpr},\Delta(\Gamma))$	SD of $\widehat{\overline{\text{perf}}}_-(s;\beta_{fpr},\Delta(\Gamma))$	Bias
logistic	500	0.237	0.024	0.048
	1000	0.212	0.015	0.024
	1500	0.198	0.012	0.009
random forest	500	0.204	0.020	0.015
	1000	0.182	0.014	-0.006
	1500	0.170	0.012	-0.018

(b) Lower bound on false positive rate

Table A2: Bias properties of estimators for the bounds on the false positive rate of a risk score $s(\cdot)$ with nonparametric outcome bounds

Notes: This table summarizes the average bias of our estimators of the bounds on the fals positive rate $\widehat{\underline{\text{perf}}}_-(s;\beta_{fpr})$, $\widehat{\overline{\text{perf}}}_-(s;\beta_{fpr})$, and the standard deviation of our estimators across simulations. We report these results for $n \in \{500, 1000, 1500\}$. The positive class performance estimators are constructed using a single sample split, and the nuisance functions are estimated using random forests. The results are computed over 1,000 simulations. See Section 6.2 for further details on the simulation design.

dashed line) as the magnitude of unmeasured confounding varies.



(a) True positive rate

(b) False positive rate

Figure A1: Estimated bounds on the true positive rate and false positive rate of a risk score $s(\cdot)$ under the MOSM as $\underline{\Gamma}, \overline{\Gamma}$ varies.

Notes: This figure illustrates box-plots (red) for the distribution of estimators of the bounds on the true positive rate (Panel A) and the false positive rate (Panel B) as $\underline{\Gamma} = 1/\tilde{\Gamma}$, $\overline{\Gamma} = \tilde{\Gamma}$ varies. The dashed black lines show the true upper and lower bounds for each value of $\tilde{\Gamma}$. The positive class performance estimators are constructed using a single split, and the nuisance functions are estimated using random forests. We report these results for $n = 1000$. The results are computed over 1,000 simulations. See Section 6.2 for further details on the simulation design.

D.2 Additional tables for the consumer lending empirical illustration

The table below provides detailed descriptions of the variable names in right panel of Table 3 in the main text.

Variable name	Detailed description
Total net income	Total net income for all applicants on the personal loan application
Occupation type	Industry code of 1st applicant's occupation.
Mos in current employment	Number of months 1st applicant has held current job.
Max delinquency in 12 mos	Maximum delinquency over last 12 months (home loan, personal loan or credit card).
Exposure to loan amount	Exposure to requested personal loan amount.
Existing personal loan balance	Existing personal loan balance of applicants.
Current days in debt	Current number of days in debt of all applicants.
Credit bureau score	External credit score.
Accommodation status	Type of accommodation applicant currently occupies (e.g., owned, rented, etc).
# of credit card apps in 12 mos (all applicants)	Number of credit card applications submitted by all applications in last 12 months.
# of credit card apps in 12 mos (1st applicant)	Number of credit card applications submitted by 1st applicant in last 12 months.
# of check acct payment reversals in 6 mos (all applicants)	Number of checking account payment reversals by all applicants in last 6 months.
# of check acct payment reversals in 6 mos (1st applicants)	Number of checking account payment reversals by first applicant in last 6 months.

Table A3: Detailed description of variable names in right panel of Table 3.