Policy Learning with α-Expected Welfare*

Yanqin Fan fany88@uw.edu
Yuan Qi ayqi@uw.edu
Gaoqian Xu gx8@uw.edu

October 6, 2025

Abstract

This paper proposes an optimal policy that targets the average welfare of the worst-off α-fraction of the post-treatment outcome distribution. We refer to this policy as the α-Expected Welfare Maximization (α-EWM) rule, where $\alpha \in (0, 1]$ denotes the size of the subpopulation of interest. The α-EWM rule interpolates between the expected welfare ($\alpha = 1$) and the Rawlsian welfare ($\alpha \to 0$). For $\alpha \in (0, 1)$, an α-EWM rule can be interpreted as a distributionally robust EWM rule that allows the target population to have a different distribution than the study population. Using the dual formulation of our α-expected welfare function, we propose a debiased estimator for the optimal policy and establish its asymptotic upper regret bounds. In addition, we develop asymptotically valid inference for the optimal welfare based on the proposed debiased estimator. We examine the finite sample performance of the debiased estimator and inference via both real and synthetic data.

JEL codes: C10, C14, C31, C54

Keywords: Average Value at Risk, Optimal Welfare Inference, Regret Bounds, Targeted Policy, Treatment Effects

*We thank Gregory M. Duncan, Sukjin Han, Toru Kitagawa, Alex Luedtke, Hyeonseok Park, Jing Tao, and participants of seminars at University of California San Diego, University of California Irvine, Institute for Advanced Economic Research at Dongbei University of Finance and Economics, Optimal Transport and Distributional Robustness in Banff, and INFORMS Annual Meeting in Seattle for feedback on an earlier version of this paper titled: "Targeted Policy Learning."

1 Introduction

1.1 Motivation

Targeted/personalized policy rules assign treatments to individuals based on their observable characteristics. Learning treatment assignment policies that benefit the relevant population in a desirable way often require careful consideration. The fact that treatment effects tend to vary with individual observable characteristics prompts policy makers to design policies that determine treatment statuses based on individual characteristics. Examples include deciding which patients should receive medical treatment, assigning unemployed workers to training programs, and selecting which students to offer financial aid. Using experimental or observational data from a sample that represents the relevant population, the optimal utilitarian policy maximizes the sum of individual welfare in the sample. The empirical welfare maximization approach in Kitagawa and Tetenov (2018) provides a solution in this regard.

As noted in Kitagawa and Tetenov (2021), maximizing the utilitarian social welfare criterion overlooks distributional impacts. This motivates Kitagawa and Tetenov (2021) to introduce an equality-minded rank-dependent social welfare function that places greater emphasis on individuals with lower-ranked outcomes. *When the policy class is restricted due to considerations such as implementability, cost, and interpretability, maximizing the utilitarian social welfare may even hurt those who are disadvantaged in the population.* For example, if welfare is measured as the (negative) mean blood sugar level of individuals at risk of diabetes and the treatment is a new medication, a utilitarian policy may prescribe the medication to most individuals because it can substantially benefit the low-risk individuals, who form the majority of the sample, but high-risk individuals who receive the medication may be hurt and end up in even worse situations. Similarly, if welfare is evaluated by the average post-training income, a utilitarian policy is more inclined to select individuals who are high school graduates and have experienced relatively short periods of unemployment to participate in the training program, while overlooking those with lower educational attainment or longer unemployment durations who might also benefit substantially from the training; see Section 5.1 in Athey and Wager (2021).

Taking a group-agnostic and risk-averse point of view, this paper proposes to learn an optimal policy that favors individuals on the lower tail of the outcome distribution. Specifically, for any $\alpha \in (0, 1)$, we introduce the α-expected welfare function as the expected outcome among the worst-affected $(\alpha \times 100)\%$ of the population, i.e., a lower-tail conditional average. We study non-randomized binary policies which maximize the α-expected welfare and refer to such policies as *α-expected welfare maximization (α-EWM)* policies. The choice of α is problem-specific and should be based on domain knowledge. A smaller α means that the policy is tailored for the more disadvantaged, whereas a larger α generates a policy that considers a broader less-advantaged subpopulation but those who are most disadvantaged receive less

attention. From a philosophical standpoint, when α is small, our α-EWM objective aligns with John Rawls' *difference principle*, which aims to maximize the welfare of the least-advantaged group to maintain social stability and fairness (Rawls, 2001). Indeed, the α-expected welfare converges to the essential infimum of the outcome random variable as α approaches zero. We note that the definition of the α-expected welfare function also applies to $\alpha = 1$, in which case it reduces to the utilitarian welfare underlying the empirical welfare maximization studied in Kitagawa and Tetenov (2018) and Athey and Wager (2021). We refer to such policies as 1-EWM throughout the rest of this paper.

To further motivate our α-EWM for $\alpha \in (0, 1)$, we provide a simple numerical comparison with the 1-EWM criterion from Kitagawa and Tetenov (2018), the equality-minded welfare criterion from Kitagawa and Tetenov (2021), and quantile maximization from Wang et al. (2018). Section 2.2 discusses the relationship between our α-EWM and these criteria in more detail. We use a simple data generating process (DGP) similar to the motivating example in Wang et al. (2018):

$$Y = 20 + 3A + X - 5AX + (1 + A + 2AX)\,\epsilon, \tag{1.1}$$

where the covariate $X \sim \text{Unif}[0, 1]$, the binary treatment $A \sim \text{Bernoulli}(0.5)$, and $\epsilon \sim N(0, 1)$. We assume that the propensity score $e_o(\cdot) = 0.5$ is known, and the policy class is defined as $\Pi_c = \mathbb{1}\{X \leq c\}$ for the policy parameter $c \in [0, 1]$.

We create a superpopulation of size one million. Since we can generate Y_i for both $A_i = 0$ and $A_i = 1$, we have full knowledge of the true outcome distribution induced by any c. For comparison, we select values of c that maximize the following: the 0.1-expected welfare, the standard Gini social welfare, the 0.1-outcome quantile, and the mean outcome. These correspond to the 0.1-EWM, equality-minded, 0.1-quantile-optimal, and 1-EWM policies, respectively. Figure 1 displays the probability densities of the post-treatment outcomes induced by these policies. Under this DGP, there is a gradual tightening of the post-treatment outcome distribution as we move from the 1-EWM policy to the equality-minded policy, then to the 0.1-quantile-optimal policy, and finally to the 0.1-EWM policy. The 0.1-EWM policy produces the most concentrated outcome distribution, with the thinnest tails on both the left and right compared to the other policies. This suggests that the 0.1-EWM policy not only mitigates the risk of extremely poor outcomes but also avoids disproportionately large gains, resulting in a more equitable distribution centered around the median.

1.2 Main Contributions

This paper makes several contributions to the literature on policy learning. First, under the assumption of unconfoundedness,[1] we show that the α-expected welfare function is identified and propose a debiased estimator. Our debiased estimator utilizes cross-fitted nuisance

[1]The assumption of unconfoundedness is not essential and can be replaced with any assumption that identifies the conditional marginal distributions of the potential outcomes.

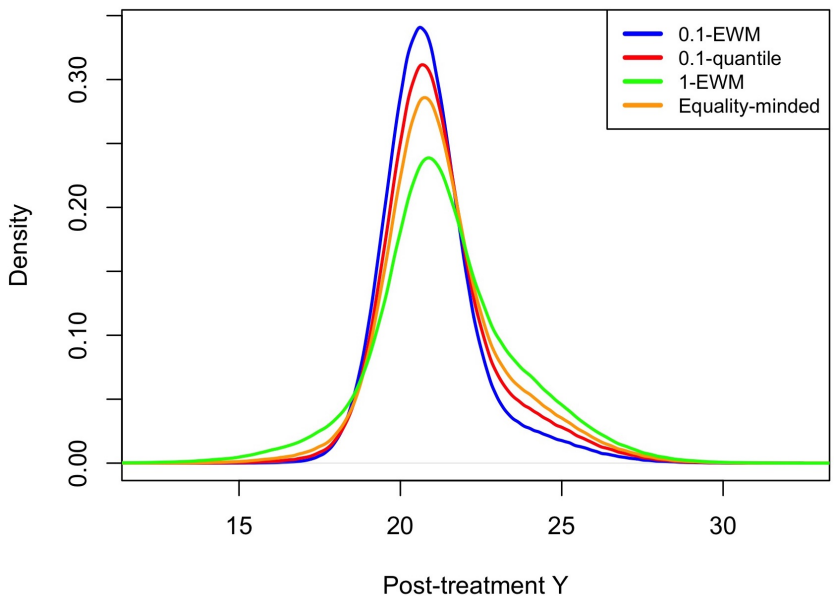

Figure 1: Distributions of post-treatment outcomes induced by the optimal policies under different welfare criteria.

estimators and the orthogonal moment function based on the dual form of the α-expected welfare function. Optimizing the α-expected welfare poses noticeable challenges compared with 1-EWM. Adopting a group-agnostic perspective, the worst-off subpopulation being targeted changes dynamically with different policies. Consequently, estimating the α-expected welfare requires the estimation of the α-quantile of the welfare, which serves as a "cutoff" for computing the tail average (see Section 2 for details).

Second, we establish theoretical guarantees of our α-EWM for any $\alpha \in (0, 1)$ by deriving asymptotic upper regret bounds with an explicit expression for the constant. This complements similar regret bounds for 1-EWM in Kitagawa and Tetenov (2018) and Athey and Wager (2021).

Third, we develop asymptotically valid inference for the optimal α-expected welfare. When the optimal policy is unique, Wald-type inference is asymptotically valid. When the optimal policy is not unique, we develop inference by applying the generalized delta method for Hadamard directionally differentiable functionals; see, e.g., Belloni et al. (2017); Fang and Santos (2019); Hong and Li (2018).

Fourth, we demonstrate that more comprehensive policy evaluations can be performed by consistently estimating the welfare of the worst-off $(\alpha \times 100)\%$ of the population for any $\alpha \in (0, 1)$ and policy. Put differently, even if a policy does not specifically target the worst-affected $(\alpha \times 100)\%$, we can still assess its performance at α to gain insights into the associated trade-offs. We illustrate our α-EWM method using experimental data from the National Job Training Partnership Act (JTPA) Study, as analyzed by Bloom et al. (1997). We find that targeting smaller subpopulations—such as the bottom 25% or 30% of the outcome distri-

bution—leads to more robust welfare performance across a range of welfare objectives. In contrast, targeting broader groups (e.g., the bottom 80%) can result in substantial welfare losses for the bottom 25%, indicating that policies aimed at broader groups may come at the expense of welfare among the most disadvantaged.

Lastly, we conduct simulation studies based on synthetic JTPA data generated using Wasserstein Generative Adversarial Networks (WGANs) developed by Athey et al. (2024), to evaluate the performance of our estimator and compare policy outcomes. In the WGAN-JTPA setup, both the 0.25-EWM and equality-minded policies enhance the welfare of lower-ranked individuals while reducing that of higher-ranked individuals relative to the 1-EWM policy, with the 0.25-EWM policy placing much greater emphasis on these adjustments. Additional simulation studies based on stylized DGPs from Athey and Wager (2021) are provided in Appendix I.3. Across all simulation setups, the debiased estimator and Wald inference perform satisfactorily for all α values considered.

The rest of the paper is organized as follows. Section 1.3 provides an overview of the related literature. Section 2 introduces our model preliminaries, including the α-expected welfare measure and its identification under the selection-on-observables assumption. We point out relations and differences between four welfare measures: the 1-expected welfare, equality-minded welfare, quantile welfare, and our α-expected welfare. Section 3 reviews the dual form of the α-expected welfare function and presents its debiased estimator, and Section 4 establishes an asymptotic upper regret bound for our debiased optimal policy. Section 5 constructs asymptotically valid inference for the optimal α-expected welfare. Section 6 presents numerical results, including an empirical application based on experimental data from the JTPA Study and a simulation study using WGAN-generated JTPA data. Section 7 concludes. Technical proofs are relegated to a series of appendices.

1.3 Related Literature

Our work builds on existing literature on policy learning from experimental and observational data, as well as statistical inference for the mean outcome under the optimal policy. In the following, we provide a brief discussion of related work.

Mean-optimal Policy Learning Existing research on policy learning in economics and statistics has mainly focused on the mean-optimal policy under unconfoundedness (Qian and Murphy, 2011; Zhao et al., 2012; Zhang et al., 2012; Bhattacharya and Dupas, 2012; Luedtke and van der Laan, 2016; Kallus, 2018; Luedtke and Chambaz, 2020; Athey and Wager, 2021). Most work on policy learning focus on establishing theoretical guarantees by deriving regret bounds. The seminal paper by Kitagawa and Tetenov (2018) explores mean-optimal policy learning from experimental data in a nonparametric framework. When propensity scores are known and the policy class denoted as Π has a finite VC dimension, they employ inverse propensity weighting to estimate the welfare function, achieving $n^{-1/2}$-rate regret bounds,

where n is the sample size. Athey and Wager (2021) extend this setup to observational studies where propensity scores are unknown and the policy class Π_n may vary with n. They estimate the objective function using doubly robust scores, a method that is shown to be efficient in the sense of Newey (1994). The resulting policies achieve regret bounds of the order $\sqrt{\text{VC}(\Pi_n)/n}$. Notably, their regret bound depends on the convergence rate of nuisance parameter estimation and the semiparametric efficient variance for evaluating an optimal policy. Finally, under mild conditions, Luedtke and Chambaz (2020) show that the regret can decay faster than $n^{-1/2}$ for a fixed data distribution.

Several studies have examined statistical inference for the mean-optimal welfare associated with the first-best policies. For instance, Luedtke and van der Laan (2016) propose an online one-step estimator that is $\sqrt{n}$-consistent for the optimal value function, where the estimated policy and value function are recursively updated using new observations. Similarly, Shi et al. (2020) conduct inference for the optimal welfare via subsample aggregating and cross-validation. In contrast, Rai (2018) study inference for the optimal mean welfare under a restricted policy class. The author utilizes bootstrap and numerical delta methods in e.g., Fang and Santos (2019) and Hong and Li (2018), to approximate the estimator's limiting distribution. We apply the same set of tools to develop inference for the optimal α-expected welfare associated with a pre-specified policy class when the optimal policy may not be unique.

Fairness and Robustness of Policy Learning. In many real-world scenarios, alternative objective functions beyond the mean outcome may be more appropriate. Some studies design objective functions with fairness considerations. Besides Kitagawa and Tetenov (2021) and Wang et al. (2018), other studies focus on distributional robustness or external validity in decision-making by adopting robust objective functions (Cui and Han, 2023; Qi et al., 2023; Adjaho and Christensen, 2022; Fan et al., 2023; Lei et al., 2023). The optimal policy under a robust objective function can be interpreted as the policy that maximizes the "worst-case" scenario of individualized outcomes when the underlying distribution is perturbed within an uncertainty set. Fang et al. (2023),Viviano and Bradic (2024), and Kim and Zubizarreta (2023) propose to maximize the average welfare subject to some fairness constraints.

The paper most closely related to ours is Qi et al. (2023), which adopts the average value-at-risk (AVaR) welfare criterion to develop robust individualized decision rules. The AVaR criterion is the same as our α-expected welfare criterion, and Qi et al. (2023) is motivated by the distributional robust representation of AVaR, see Eq. (2.3). Apart from differences in motivation, the main results in Qi et al. (2023) and our paper also differ. First, Qi et al. (2023) focus on experimental data with a known propensity score, allowing direct estimation of the objective function. Instead, we consider observational studies with unknown propensity scores and estimate our objective function using doubly robust scores and cross-fitting. Second, we consider a general policy class Π_n with a VC-dimension $\text{VC}(\Pi_n)$ that may be changing with n. In contrast, Qi et al. (2023) consider a more restrictive policy class within a reproducing kernel Hilbert space, which excludes many machine learning algorithms, such

as decision trees and neural networks, from being used to learn the optimal policy. Third, applied to the class of policies in Qi et al. (2023), our regret bound is sharper than theirs. Fourth, we develop inference for the optimal welfare in experimental and observational setups. Computationally, Qi et al. (2023) propose a non-convex optimization algorithm based on a surrogate function that smooths the binary policy function for the use of difference-of-convex optimization, whereas our optimization is done by derivative-free methods.

We close this section by summarizing the notation used in this paper. We use $O, o, O_P, o_P, \asymp, \gtrsim, \lesssim$ in the following sense: $a_n = O(b_n)$ if $|a_n| \leq Cb_n$ for n large enough; $a_n = o(b_n)$ if $a_n/b_n \to 0$; $X_n = O_P(b_n)$, if for any $\delta > 0$, there exist $M, N > 0$, such that $\mathbb{P}[|X_n| \geq Mb_n] \leq \delta$ for any $n > N$; $X_n = o_P(b_n)$, if $\mathbb{P}[|X_n| \geq \epsilon b_n] \to 0$ for any $\epsilon > 0$; $a_n \asymp b_n$ if there exist $k_1, k_2 > 0$ and n_0, such that for all $n > n_0, k_1 a_n \leq b_n \leq k_2 a_n$ if $\lim a_n/b_n = \infty$; $a_n \gtrsim b_n$ if $b_n = O(a_n)$; $a_n \lesssim b_n$ if $a_n = O(b_n)$. Furthermore, we write $f(n) = \widetilde{O}(g(n))$ if there is a function h that grows poly-logarithmically such that $f(n) \leq h(g(n))g(n)$. The notation $f(n) = \Omega(g(n))$ means that there is a universal constant $c_o > 0$ such that $f(n) \geq c_o g(n)$ uniformly in n. We use the shorthand $[n] = \{1, \ldots, n\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The abbreviation i.i.d. stands for *independent and identically distributed*. In the sequel, let c_o denote a generic positive constant, whose value may vary from line to line.

2 α-Expected Welfare Function and Optimal Policy

Suppose that we have a random sample $(X_i, Y_i, A_i)_{i=1}^n$, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ denotes the observable characteristics of individual i (continuous or discrete), $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ represents the outcome of individual i (or utility / welfare), and $A_i \in \{0, 1\}$ denotes the treatment status of individual i, for $i \in [n]$. Without loss of generality, larger values of Y_i are assumed to be preferable. To simplify notation, we define $Z_i := (X_i, Y_i, A_i) \in \mathcal{Z}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \times \{0, 1\}$. Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes that would have been observed if $A_i = 0$ and $A_i = 1$, respectively. Then $Y_i = A_i Y_i(1) + (1 - A_i)Y_i(0)$ is the realized outcome under the Stable Unit Treatment Value Assumption (Rubin, 1978, 1990).

Throughout the rest of this paper, we assume that $\mathbb{E}|Y_i(0)| < \infty$ and $\mathbb{E}|Y_i(1)| < \infty$. We denote by P the distribution of $Z_i \equiv (X_i, Y_i, A_i)$, and by $\mathbb{E}_P$ and Var_P the expectation and variance under P, respectively.

2.1 α-Expected Welfare Function and Identification

We study non-randomized binary policy/rule $\pi : \mathcal{X} \to \{0, 1\}$. Let Π_o denote the policy class that contains all Borel measurable functions from $\mathcal{X}$ to $\{0, 1\}$. For any policy $\pi \in \Pi_o$, let $Y_i(\pi) := Y_i(\pi(X_i))$, the outcome of individual i when π is implemented. Further, let $F_\pi(y)$, $y \in \mathcal{Y}$ denote the distribution function of $Y_i(\pi)$ and $F_\pi^{-1}(\alpha) = \inf\{y \in \mathbb{R} : F_\pi(y) \geq \alpha\}$ denote the quantile function of $Y_i(\pi)$.

As discussed by Kitagawa and Tetenov (2018) and Athey and Wager (2021), practitioners

may adopt a pre-specified policy class $\Pi \subseteq \Pi_o$ that incorporates constraints relevant to the problem context, such as budgetary limitations, specific functional forms, fairness considerations, and other pertinent factors.

Definition 2.1 (α-Expected Welfare and Optimal Policy)**.** Given a policy class Π chosen by the policymaker, we define *the α-expected welfare of* $Y_i(\pi)$ as the expected welfare of the worst-off subpopulation of size $\alpha \in (0, 1]$, i.e.,

$$\mathbb{W}_\alpha(\pi) := \frac{1}{\alpha} \int_0^\alpha F_\pi^{-1}(t)dt \quad \text{for } \pi \in \Pi. \tag{2.1}$$

An *α-expected welfare maximization (α-EWM) policy* is defined as

$$\pi_\alpha^* \in \operatorname{argmax}_{\pi \in \Pi} \mathbb{W}_\alpha(\pi).$$

As discussed in Section 1, $\lim_{\alpha \to 0} \mathbb{W}_\alpha(\pi) = \operatorname{ess\,inf} Y_i(\pi)$ and $\mathbb{W}_1(\pi) = \mathbb{E}[Y_i(\pi)]$. Our welfare function $\mathbb{W}_\alpha(\pi)$ therefore flexibly interpolates between the expected welfare and infimum welfare of the target population by varying $\alpha \in (0, 1]$, where $\alpha = 1$ gives the expected welfare of the target population adopted in Kitagawa and Tetenov (2018) and Athey and Wager (2021).

Remark 2.1. (i) Our welfare function $\mathbb{W}_\alpha(\pi)$ is identical to Expected Shortfall, a commonly used coherent risk measure in finance and risk management. When the distribution function of $Y_i(\pi)$ is continuous at $F_\pi^{-1}(\alpha)$, $\mathbb{W}_\alpha(\pi)$ is also the same as Conditional Value at Risk (CVaR), defined as $\mathrm{CVaR}_\alpha(\pi) := \mathbb{E}\left[Y_i(\pi) \mid Y_i(\pi) \leq F_\pi^{-1}(\alpha)\right]$, see Rockafellar et al. (2000); Shapiro et al. (2021).

(ii) $\mathbb{W}_\alpha(\pi)$ is also closely related to the generalized Lorenz function, a popular tool for measuring and comparing inequality, see Greselin and Zitikis (2018) and Shorrocks (1983). Specifically, let $L_\alpha^{\mathrm{gen}}(Y_i(\pi))$ denote the *generalized* (unnormalized) Lorenz function at level α: $L_\alpha^{\mathrm{gen}}(Y_i(\pi)) := \int_0^\alpha F_\pi^{-1}(t)dt$. Then $\mathbb{W}_\alpha(\pi) = \frac{1}{\alpha} L_\alpha^{\mathrm{gen}}(Y_i(\pi))$.

To identify $\mathbb{W}_\alpha(\pi)$ as defined in Eq. (2.1), we note that

$$Y_i(\pi) = \pi(X_i)Y_i(1) + [1 - \pi(X_i)]Y_i(0).$$

The conditional (given $X_i = x$) and unconditional distribution functions of $Y_i(\pi)$ are

$$F_\pi(y|x) = \pi(x)F_1(y|x) + (1 - \pi(x))F_0(y|x) \quad \text{and} \quad F_\pi(y) = \int_{\mathcal{X}} F_\pi(y|x)\mathrm{d}P_X(x), \tag{2.2}$$

where $F_1(y|x)$ and $F_0(y|x)$ are the conditional distribution functions of $Y_i(1)$ and $Y_i(0)$ given $X_i = x$, respectively.

Eq. (2.1) and Eq. (2.2) imply that $\mathbb{W}_\alpha(\pi)$ is a function of the policy $\pi(\cdot)$ and the conditional distribution functions $F_1(\cdot|\cdot)$ and $F_0(\cdot|\cdot)$. Consequently, for any $\pi \in \Pi_o$, $\mathbb{W}_\alpha(\pi)$ is identified as long as $F_1(\cdot|\cdot)$ and $F_0(\cdot|\cdot)$ are identified. Any assumption that ensures the identification

of $F_1(\cdot|\cdot)$ and $F_0(\cdot|\cdot)$ is sufficient to identify $\mathbb{W}_\alpha(\pi)$. In the rest of this paper, we adopt the selection-on-observables assumption, which includes unconfoundedness and common support, as detailed in Assumption 2.1.

Assumption 2.1. (1) *Unconfoundedness:* $(Y_i(0), Y_i(1)) \perp\!\!\!\perp A_i \mid X_i$.

(2) *Strong overlap*: Let $e_o(x) := \mathbb{P}[A_i = 1 \mid X_i = x]$ denote the propensity score. There is a constant $\kappa \in \left(0, \frac{1}{2}\right)$ such that $e_o(x) \in [\kappa, 1-\kappa]$ for all $x \in \mathcal{X}$.

Assumption 2.1 (1) states that the potential outcomes are independent of the treatments after conditioning on the observed covariates. Heuristically, it requires that all confounders that affect both treatments and potential outcomes simultaneously be observed. For identification, Assumption 2.1 (2) can be relaxed to the weaker condition that $e_o(x) \in (0,1)$ for all $x \in \mathcal{X}$, but the regret bounds and inference developed in later sections of this paper rely on it.

Under Assumption 2.1, the distribution functions $F_a(\cdot|x)$ for all $x \in \mathcal{X}$ are point-identified:

$$F_a(y|x) := \mathbb{P}[Y_i(a) \le y | X_i = x] = \mathbb{P}[Y_i \le y | X_i = x, A_i = a].$$

Consequently, $\mathbb{W}_\alpha(\pi)$ is identified for any $\pi \in \Pi_o$.

2.2 Relations with Other Welfare Maximization Criteria

In this subsection, we compare our α-expected welfare $\mathbb{W}_\alpha(\pi)$, defined for $\alpha \in (0,1)$, with three welfare functions commonly used in the literature: the expected welfare, the equality-minded welfare, and the quantile welfare functions.

2.2.1 1-Expected Welfare Maximization

1-EWM in Kitagawa and Tetenov (2018) and Athey and Wager (2021) take the mean outcome $\mathbb{E}[Y_i(\pi)]$, which equals $\mathbb{W}_1(\pi)$, as the population welfare function, *assuming that the distribution of* $Y_i(\pi)$ *in the target population is the same as that in the study population.*

For $\alpha \in (0,1)$*, our* α*-expected welfare* $\mathbb{W}_\alpha(\pi)$ *represents a distributionally robust version of the* 1*-expected welfare function.* To see this, consider the uncertainty set centered at probability distribution F_π of the outcome under policy $\pi : \mathcal{X} \to \{0,1\}$:

$$\begin{aligned}\mathcal{U}_\alpha(F_\pi) &= \left\{ Q : D_\infty(Q \| F_\pi) \le \log \frac{1}{\alpha} \right\} \\ &= \{ Q : \exists\ P \in \mathcal{P}(\mathcal{Y}), t \in [\alpha, 1] \text{ s.t. } F_\pi = tQ + (1-t)P \}.\end{aligned}$$

where $D_\infty(Q\|F_\pi) = \operatorname{ess\,sup} \log \frac{dQ}{dF_\pi}$. From Rockafellar et al. (2002) and Duchi et al. (2023), it follows that

$$\mathbb{W}_\alpha(\pi) = \inf_{Q \in \mathcal{U}_\alpha(F_\pi)} \mathbb{E}_{Z \sim Q}[Z]. \tag{2.3}$$

The uncertainty set $\mathcal{U}_\alpha(F_\pi)$ is the risk envelope capturing the distributional uncertainty of $Y_i(\pi)$ in the target population, comprising distributions with minority subpopulations of at least size α. We can therefore interpret π^*_α as the distributionally robust policy that maximizes the average welfare under the worst-case perturbation of the study population in $\mathcal{U}_\alpha(F_\pi)$. As α decreases, the uncertainty set expands, making the α-expected welfare function more robust to potential distributional shifts in $Y_i(\pi)$ within the target population.

2.2.2 Equality-Minded Welfare Maximization

Since the 1-EWM may worsen inequality, Kitagawa and Tetenov (2021) propose equality-minded policies by maximizing rank-dependent social welfare functions (SWFs), which assign greater weights to lower-ranked individuals. Given a decreasing function $\Lambda : [0,1] \to [0,1]$ with $\Lambda(0) = 1$ and $\Lambda(1) = 0$, the equality-minded welfare under policy π is defined as

$$W_\Lambda(F_\pi) := \int_0^\infty \Lambda\left(F_\pi(y)\right) \mathrm{d}y = \int_0^1 F_\pi^{-1}\left(t\right) \omega(t) \mathrm{d}t, \tag{2.4}$$

where $\omega(t) := -\frac{\mathrm{d}}{\mathrm{d}t}\Lambda(t)$ is the associated weight function. When Λ is strictly convex, the associated SWF, W_Λ, upholds the Pigou-Dalton Principle of Transfers, as rank-preserving transfers from higher-ranked individuals to lower-ranked individuals are preferred under the welfare W_Λ. The function Λ, chosen by practitioners, captures the degree of inequality aversion through its level of complexity. An important class of rank-dependent SWFs is the extended Gini SWFs, where $\Lambda(t) = \Lambda_k(t) = (1-t)^{k-1}$ for some $k \geq 2$, and the weight function is $\omega(t) = \omega_k(t) = (k-1)(1-t)^{k-2}$. The expected welfare and the standard Gini SWF correspond to $k = 2$ and $k = 3$, respectively.

Equality-minded SWFs can, in fact, be expressed in terms of our α-expected welfare $\mathbb{W}_\alpha(\pi)$. For example, when $k > 2$, the extended Gini SWF can be written as a weighted average of $\mathbb{W}_\alpha(\pi)$:

$$W_\Lambda(F_\pi) = (k-2) \int_0^1 \mathbb{W}_\alpha(\pi) \alpha (1-\alpha)^{k-3} \mathrm{d}\alpha. \tag{2.5}$$

Although our α-expected welfare can be written as

$$\mathbb{W}_\alpha(\pi) = \frac{1}{\alpha} \int_0^\alpha F_\pi^{-1}(t) \mathrm{d}t = \int_0^1 F_\pi^{-1}\left(t\right) \sigma(t) \mathrm{d}t, \tag{2.6}$$

where $\Lambda(t) = (1 - t/\alpha)\, \mathbb{1}\{0 \leq t \leq \alpha\}$ and $\sigma(t) = \frac{1}{\alpha}\mathbb{1}\{0 \leq t \leq \alpha\}$, it does not satisfy the Pigou-Dalton Principle of Transfers, as $\Lambda(t)$ is not strictly convex. This principle is satisfied only if the rank-preserving transfer happens across the probability level α, i.e., from an individual ranked above α to an individual ranked below α. Transfers on the same side do not affect $\mathrm{W}_\alpha(\pi)$ since all the individuals involved have the same weight.

2.2.3 Quantile Welfare Maximization

To prioritize the lower tail of population welfare over the (weighted) expected welfare, Wang et al. (2018) propose a quantile-optimal policy, defined as

$$\text{argmax}_{\pi\in\Pi} \text{VaR}_\alpha(Y_i(\pi)) = F_\pi^{-1}(\alpha),$$

where $\alpha \in (0,1)$ is the quantile level of interest. For the class of linear policies with a fixed number of covariates Π, Wang et al. (2018) establish the cube root asymptotics for the estimator of the parameter that defines the optimal linear policy.

Compared with quantile welfare $F_\pi^{-1}(\alpha)$ that overlooks the welfare of the population with outcomes below it, our α-expected welfare function $\mathbb{W}_\alpha(\pi)$ integrates $F_\pi^{-1}(t)$ over the range $[0,\alpha]$, thereby accounting for welfare levels below the α-quantile and providing a more comprehensive assessment of the lower tail of the welfare distribution.

3 Debiased Estimation and Practical Implementation

The α-expected welfare function $\mathbb{W}_\alpha(\pi)$ has a convenient dual representation, which we will use to construct a debiased estimator of $\mathbb{W}_\alpha(\pi)$.

Let $(u)_- := \min(u,0)$ and $(u)_+ := \max(u,0)$. Further, let $\theta = (\pi,\eta)$ and

$$\mathbb{V}_\alpha(\theta) = \frac{1}{\alpha}\mathbb{E}\left[(Y_i(\pi)-\eta)_-\right] + \eta.$$

Lemma 3.1 (Dual Representation of $\mathbb{W}_\alpha(\pi)$)**.** For any $\alpha \in (0,1]$ and $\pi \in \Pi$,

$$\mathbb{W}_\alpha(\pi) = \sup_{\eta\in\mathbb{R}} \mathbb{V}_\alpha(\pi,\eta).$$

Furthermore, for $\alpha \in (0,1)$, the supremum is attained on the interval $[t^\star, t^{\star\star}]$, where $t^\star = \sup\{y \in \mathbb{R} : F_\pi(y) \le \alpha\}$ and $t^{\star\star} = F_\pi^{-1}(\alpha)$. When $\alpha = 1$, if the support of $Y_i(\pi)$ is bounded, then the supremum is attained on $\left[F_\pi^{-1}(1), \infty\right)$. Otherwise, the supremum is unattainable and $\sup_{\eta\in\mathbb{R}} \mathbb{V}_1(\pi,\eta) = \lim_{\eta\to\infty} \mathbb{V}_1(\pi,\eta)$.

Let

$$\mu_a(x,\eta) := \mathbb{E}\left[(Y_i(a)-\eta)_- \,|X_i = x\right] \text{ for } a \in \{0,1\},$$

and $\tau(x,\eta) := \mu_1(x,\eta) - \mu_0(x,\eta)$ for any $x \in \mathcal{X}$ and $\eta \in \mathbb{R}$. Under Assumption 2.1, $\tau(x,\eta)$ is identified for any given η.

Theorem 3.1. Under Assumption 2.1, for any $0 < \alpha \le 1$ and any $\theta = (\pi,\eta) \in \Pi_o \times \mathbb{R}$, it

holds that

$$\begin{aligned}\mathbb{V}_\alpha(\theta) &= \frac{1}{\alpha}\left\{\mathbb{E}\left[\pi(X_i)\mu_1(X_i,\eta)\right] + \mathbb{E}\left[(1-\pi(X_i))\,\mu_0(X_i,\eta)\right]\right\} + \eta, \\ &= \frac{1}{\alpha}\left\{\mathbb{E}\left[\pi(X_i)\tau(X_i,\eta)\right] + \mathbb{E}\left[\mu_0(X_i,\eta)\right]\right\} + \eta, \\ &= \frac{1}{\alpha}\mathbb{E}\left[w(X_i,A_i,\pi)(Y_i-\eta)_-\right] + \eta,\end{aligned} \tag{3.1}$$

where the function $w : \mathcal{X} \times \{0,1\} \times \Pi_o \to [0,\infty)$ is defined as

$$w(x,a,\pi) := \frac{a\pi(x)}{e_o(x)} + \frac{(1-a)\,(1-\pi(x))}{1-e_o(x)}.$$

Remark 3.1. (i) When $0 < \alpha < 1$, the feasible set in the dual representation of $\mathbb{W}_\alpha(\pi)$ in Lemma 3.1 can be restricted to a compact set. Since $|Y_i(\pi)| \leq |Y_i(0)| + |Y_i(1)|$ for all $\pi \in \Pi_o$, the α-quantile of $|Y_i(\pi)|$ is no greater than the α-quantile of $|Y_i(0)| + |Y_i(1)|$, while the α-quantile of $-|Y_i(\pi)|$ is no less than the α-quantile of $-|Y_i(0)| - |Y_i(1)|$. Therefore, the solution to $\sup_{\eta\in\mathbb{R}} \mathbb{V}(\pi,\eta)$ is $\mathrm{VaR}_\alpha(Y_i(\pi))$, which satisfies the bounds

$$-\mathrm{VaR}_{1-\alpha}\left(|Y_i(0)| + |Y_i(1)|\right) \leq \mathrm{VaR}_\alpha(Y_i(\pi)) \leq \mathrm{VaR}_\alpha\left(|Y_i(0)| + |Y_i(1)|\right).$$

Thus, we can express $\mathbb{W}_\alpha(\pi)$ as $\sup_{\eta\in\mathcal{B}_Y} \mathbb{V}_\alpha(\pi,\eta)$ for some compact set $\mathcal{B}_Y \subset \mathbb{R}$.

(ii) When Y_i has a bounded support, the claim in (i) holds for $\alpha = 1$ as well.

As noted in the previous sections, the 1-expected welfare function $\mathbb{W}_1(\pi)$ is the same as the expected welfare $\mathbb{E}[Y_i(\pi)]$ in Kitagawa and Tetenov (2018) and Athey and Wager (2021). In the rest of this paper, we focus on estimation and asymptotic theory for an α-EWM rule when $\alpha \in (0,1)$.

Theorem 3.1 suggests two plug-in methods for estimating $\mathbb{V}_\alpha(\theta)$ or the welfare function $\mathbb{W}_\alpha(\pi)$: IPW and outcome equation estimation. It is known that the IPW estimator is sensitive to the estimator of the propensity score and may suffer from severe bias. The outcome equation estimator may be sensitive to the estimators of τ (μ_1 and μ_0). This motivates the debiased estimator proposed in this section.

Theorem 3.1 implies that under Assumption 2.1, the function $\mathbb{V}_\alpha(\theta)$ is identified for any fixed $\theta = (\pi,\eta)$. Following Robins et al. (1994) and Robins et al. (1995), we build our doubly robust score for $\mathbb{V}_\alpha(\theta)$ by introducing the augmentation term. Given any $\theta = (\eta,\pi)$, and for any function $\check{e} : \mathcal{X} \to (0,1)$ and $\check{\mu}_a : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ with $a \in \{0,1\}$, define

$$\begin{aligned}g_\theta(z;\check{\mu},\check{e}) = &\frac{1}{\alpha}\left[(1-\pi(x))\,\check{\mu}_0(x,\eta) + \pi(x)\check{\mu}_1(x,\eta)\right] + \eta \\ &+ \frac{1}{\alpha}\left[\frac{(1-\pi(x))(1-a)}{1-\check{e}\,(x)}\left((y-\eta)_- - \check{\mu}_0(x,\eta)\right)\right] \\ &+ \frac{1}{\alpha}\left[\frac{\pi(x)a}{\check{e}(x)}\left((y-\eta)_- - \check{\mu}_1(x,\eta)\right)\right],\end{aligned} \tag{3.2}$$

where $\check{\mu} = (\check{\mu}_0, \check{\mu}_1)$ and the augmentation term is defined as the sum of the last two components

in (3.2). The augmentation term has mean zero and the Neyman orthogonality condition holds:

$$\partial_\mu \mathbb{E}_P[g_\theta(Z_i;\mu_o,e_o)][\check{\mu}-\mu_o]=0, \quad \text{and} \quad \partial_e \mathbb{E}_P[g_\theta(Z_i;\mu_o,e_o)][\check{e}-e_o]=0,$$

where $\mu_o=(\mu_0,\mu_1)$. To simplify notation, we let $g_\theta(\cdot)=g_\theta(\cdot;\mu_o,e_o)$, where the function $g_\theta(\cdot)$ indexed by θ is referred to as the (doubly robust) score function for estimating $\mathbb{V}_\alpha(\theta)$. It is clear that for any given θ, the function $g_\theta-\mathbb{E}_P[g_\theta(Z_i)]$ is the efficient influence function for $\mathbb{V}_\alpha(\theta)$; see Luedtke and van der Laan (2016); Kennedy (2016) for more detailed discussion.

Building upon Chernozhukov et al. (2018) and Chernozhukov et al. (2022), we construct our doubly robust score $\widehat{g}_\theta(Z_i)$ for $\mathbb{V}_\alpha(\theta)$ based on K-fold cross-fitting, a sample-splitting method used to validate asymptotic properties and leverage high-level conditions concerning the predictive accuracy of nuisance estimation methods.

We describe the estimation steps below, see Algorithm 1 in Appendix H for details.

(a) Randomly partition the sample into K folds $\cup_{k=1}^K \mathcal{I}_k$ such that $|\mathcal{I}_k|=n/K$.

(b) For each k, define $\mathcal{I}_k^c=[n]\setminus\mathcal{I}_k$. Fit estimators for the nuisance parameters $e_o(\cdot)$ and $\mu_a(\cdot,\cdot)$ for $a\in\{0,1\}$ using the observations in the remaining $K-1$ folds, specifically, $(Z_i)_{i\in\mathcal{I}_k^c}$. Denote these estimators as $\widehat{e}^{(-k)}(\cdot)$ and $\widehat{\mu}_a^{(-k)}(\cdot,\cdot)$.

(c) The doubly robust score is

$$\begin{aligned}\widehat{g}_\theta(Z_i;\widehat{\mu}^{-k(i)},\widehat{e}^{-k(i)}) &= \frac{1}{\alpha}\left[(1-\pi(X_i))\,\widehat{\mu}_0^{-k(i)}(X_i,\eta)+\pi(X_i)\,\widehat{\mu}_1^{-k(i)}(X_i,\eta)\right]+\eta \\ &+\frac{1}{\alpha}\left[\frac{(1-\pi(X_i))\,(1-A_i)}{1-\widehat{e}^{-k(i)}(X_i)}\left[(Y_i-\eta)_- -\widehat{\mu}_0^{-k(i)}(X_i,\eta)\right]\right] \\ &+\frac{1}{\alpha}\left[\frac{\pi(X_i)A_i}{\widehat{e}^{(-k(i))}(X_i)}\left[(Y_i-\eta)_- -\widehat{\mu}_1^{-k(i)}(X_i,\eta)\right]\right],\end{aligned} \tag{3.3}$$

where $k(i)$ is the index in $[n]$ such that $i\in\mathcal{I}_k$.

(d) For each $\theta=(\pi,\eta)$, $\mathbb{V}(\theta)$ and $\mathbb{W}_\alpha(\pi)$ can be estimated by

$$\widehat{\mathbb{V}}_n(\theta)=\frac{1}{n}\sum_{i=1}^n \widehat{g}_\theta(Z_i) \quad \text{and} \quad \widehat{\mathbb{W}}_n(\pi)=\sup_{\eta\in\mathcal{B}_Y}\widehat{\mathbb{V}}_n(\pi,\eta),$$

where $\mathcal{B}_Y$ is introduced in Remark 3.1.[2]

(e) The debiased estimator $\widehat{\theta}_n=(\widehat{\pi}_n,\widehat{\eta}_n)$ is the maximizer of $\widehat{\mathbb{V}}_n(\theta)$.

Remark 3.2. Since Lemma 3.1 implies that $\mathbb{W}_\alpha(\pi)=\mathbb{V}_\alpha(\pi,F_\pi^{-1}(\alpha))$, a debiased estimator of π^* can also be constructed from the expression $\mathbb{V}_\alpha(\pi,F_\pi^{-1}(\alpha))$. Noting that $F_\pi^{-1}(\alpha)$ is an

[2]If the support of Y_i is bounded, i.e., Assumption 5.1 (1) holds, then one can take $\mathcal{B}_Y$ as the support of Y_i. In our numerical work, we took $\mathcal{B}_Y$ as the closed interval with lower and upper bounds as the minimum and maximum statistics of Y_i respectively.

optimal solution to $\sup_{\eta\in\mathbb{R}} \mathbb{V}_\alpha(\theta)$, the orthogonal moment function based on $\mathbb{V}_\alpha(\pi, F_\pi^{-1}(\alpha))$ is equal to

$$g_{(\pi,F_\pi^{-1}(\alpha))}(z;\check{\mu},\check{e}) - \mathbb{V}_\alpha(\pi, F_\pi^{-1}(\alpha)).$$

A cross-fitting estimator can be constructed from $g_{(\pi,F_\pi^{-1}(\alpha))}(z;\check{\mu},\check{e})$, but it requires an estimator of the quantile function $F_\pi^{-1}(\alpha)$. We leave a detailed comparison between these two estimators in future work.

4 Asymptotic Upper Regret Bounds

In this section, we establish asymptotic regret bounds on the debiased α-EWM policy proposed in Section 3 for any fixed $\alpha \in (0,1)$. They complement similar regret bounds for the 1-EWM and equality-minded policies established in Kitagawa and Tetenov (2018), Athey and Wager (2021) and Kitagawa and Tetenov (2021).

4.1 Policy Class and Examples

For each n, let Π_n denote the class of candidate policies and $\Theta_n = \Pi_n \times \mathcal{B}_Y$, where $\mathcal{B}_Y \subset \mathbb{R}$ is a compact set introduced in Remark 3.1. For brevity, we write $\mathbb{V}(\theta) \equiv \mathbb{V}_\alpha(\theta)$, omitting the subscript α.

The following assumption restricts the complexity of the policy class Π_n.

Assumption 4.1. There exists a constant $b_o > 0$ such that the VC-dimension of Π_n is bounded as $\mathrm{VC}(\Pi_n) \leq n^{b_o}$ for all $n \in \mathbb{N}^+$.

A policy is a classifier that assigns the covariate X_i to a binary treatment status. Any machine learning classification model can serve as a candidate policy class. In the following, we list three examples of policy classes and their VC dimensions.

Example 1 (Linear Rules)**.** The linear policy class can be characterized by

$$\Pi_n = \left\{\mathbb{1}\{x'\beta > 0\} : \beta \in B\right\}, \tag{4.1}$$

where B is compact subset of $\mathbb{R}^{p_n}$, where the dimension of the covariates p_n is allowed to grow with the sample size n. Although the eligibility score is linear in β, it can include intercepts, interaction terms, higher-order terms, and other transformations of the original covariate X_i. The VC-dimension of Π_n is $p_n + 1$.

Example 2 (Decision Trees)**.** A decision tree is a predictor $\pi : \mathcal{X} \subset \mathbb{R}^p \to \{0,1\}$ that recursively partitions the feature space $\mathcal{X}$ into a set of rectangles and assigns a label to each resulting partition. Following Bertsimas and Dunn (2017) and Zhou et al. (2023), we define a decision tree recursively. A decision tree of depth L consists of L levels, with the first $L-1$ levels containing branch nodes and the final L-th level comprising exclusively of leaf nodes. For any branch node, we choose the split-point b and the variable $x(j)$ that is a single

component of x . If $x(j) < b$, the path taken is towards the left; if not, the decision leads to the right branch. Each path will end with a leaf node that is assigned a unique label. Zhou et al. (2023) show that the VC-dimension of the class Π of decision trees of depth-L over $\mathbb{R}$ is $\mathrm{VC}(\Pi) = \widetilde{\mathcal{O}}(2^L \log p)$.

Example 3 (ReLU Neural Networks)**.** Deep neural networks have achieved significant success in complex classification tasks, especially in image and speech recognition. Formally, a neural network is defined by an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, structured as a directed acyclic graph, alongside a set of parameters that include a weight for each edge within the graph and a bias for each node. Common activation functions include the sigmoid, $\sigma(x) = 1/(1 + e^{-x})$, and the Rectified Linear Unit (ReLU), $\sigma(x) = \max(0, x)$. Each edge represents a connection that transmits the output from one neuron to the input of another. This input is calculated as a weighted sum of the outputs from all connected neurons, allowing the network to capture complex relationships and patterns in the data.

Let W denote the total number of parameters (weights and biases), U the total number of computation units (nodes), and L the length of the longest path in the network graph. Let Π denote the policy class of deep ReLU networks characterized by W weights and L layers. Bartlett et al. (2019) establish that $\mathrm{VC}(\Pi) = O(WL \log(W))$ and $\mathrm{VC}(\Pi) = \Omega\left(WL \log(W/L)\right)$.

4.2 Assumptions on Nuisance Estimators and a Preliminary Lemma

In this section, we establish a fundamental lemma showing that the estimation error of the nuisance parameters can be ignored when $\mathbb{V}(\cdot)$ is estimated using the doubly robust score with cross-fitting. Before presenting the lemma, we introduce additional assumptions.

Let $\mathbb{V}_n(\theta) = \mathbb{P}_n g_\theta$, where $g_\theta(z) := g_\theta(z; \mu_o, e_o)$ is defined in (3.2). We assume that the nuisance parameter estimators $\widehat{\mu}_a(\cdot, \cdot)$ and $\widehat{e}(\cdot)$ converge to their true values at sufficiently fast rates.

Assumption 4.2. (1) $\sup_{(x,\eta)\in\mathcal{X}\times\mathcal{B}_Y} |\widehat{\mu}_a(x, \eta) - \mu_a(x, \eta)| = o_P(1)$ for $a \in \{0, 1\}$, and $\sup_{x\in\mathcal{X}} |\widehat{e}(x) - e_o(x)| = o_P(1)$.

(2) Suppose there are $\zeta_\mu > 0$ and $\zeta_e > 0$ such that

$$\sup_{\eta\in\mathcal{B}_Y} \left[\mathbb{E}\,|\widehat{\mu}_a(X_i, \eta) - \mu_a(X_i, \eta)|^2\right]^{1/2} = O(n^{-\zeta_\mu}),$$
$$\left[\mathbb{E}\,|\widehat{e}(X_i) - e_o(X_i)|^2\right]^{1/2} = O(n^{-\zeta_e}).$$

(3) $\mathrm{VC}(\Pi_n) = o\left(n^{2\zeta_\mu \wedge 2\zeta_e}\right)$.

Remark 4.1. (i) The regression function $\mu_a(x, \eta)$ can readily be estimated by regressing $\{(Y_i - \eta)_- : A_i = a\}$ on $\{X_i : A_i = a\}$. The uniformity in η does not severely impact

the uniform convergence of the estimator. For example, given a bandwidth $b_n = o(1)$, a higher-order kernel can be employed to estimate $\mu_a(x,\eta)$ as follows:

$$\widehat{\mu}_a(x,\eta) = \frac{\sum_{i:A_i=a}^n (Y_i - \eta)_- K\left(\frac{x-X_i}{b_n}\right)}{\sum_{i:A_i=a}^n K\left(\frac{x-X_i}{b_n}\right)}.$$

Under Assumption 2.1 and certain regularity conditions on the kernel function $K(\cdot)$, if $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$ and the function $\mu(\cdot,\cdot)$ belongs to the Hölder space $\mathcal{C}^s(\mathcal{X} \times \mathcal{B}_Y)$ with smoothness parameter s, it follows that

$$\sup_{x\in\mathcal{X},\eta\in\mathcal{B}_Y} |\widehat{\mu}_a(x,\eta) - \mu(x,\eta)| = O\left(b_n^s\right) + O_P\left(\sqrt{\frac{(p+1)\log b_n}{n b_n^{p+1}}}\right).$$

With a careful choice of bandwidth, the optimal convergence rates—both uniform and in L^2, can be achieved and are given by $(\log n/n)^{s/(2s+p)}$; see Giné and Guillou (2002); Giné and Nickl (2021). The sieve-based approach can also be applied in this context; see Chen and Christensen (2015); Belloni et al. (2015); Ai and Chen (2003); Blundell et al. (2007). Furthermore, machine learning techniques can be employed to estimate nuisance parameters. The L^2-convergence rates for nonparametric regression using deep neural networks have been extensively studied; see Farrell et al. (2021); Kohler and Langer (2021); Schmidt-Hieber (2020).

(ii) Alternatively, for $a \in \{0,1\}$ and $x \in \mathcal{X}$, it holds that

$$\mu_a(x,\eta) = \mathbb{E}\left[(Y_i(a) - \eta)_- \,|X_i = x\right] = \int_{-\infty}^{\eta} y \mathrm{d}F_a(y|x) - \eta F_a(\eta|x). \tag{4.2}$$

This suggests a plug-in estimator based on an estimator of $F_a(y|x)$.

We conclude this subsection by demonstrating that $\widehat{\mathbb{V}}_n(\theta)$ is a good approximation to $\mathbb{V}_n(\theta) = \mathbb{P}_n g_\theta$ with convergence rate faster than $n^{-1/2}$. Consequently, we can ignore the nuisance parameter estimation errors in subsequent asymptotic analysis.

Lemma 4.1. Suppose Assumption 2.1, Assumption 4.1 and Assumption 4.2 hold. If $b_o/2 < \zeta_e \wedge \zeta_\mu$, then

$$\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}\left|\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}_n(\theta)\right|\right] = O(n^{-1/2}).$$

4.3 Asymptotic Upper Regret Bound

In this subsection, we study the regret upper bound of implementing $\widehat{\pi}_n$ under the following assumption.

Assumption 4.3. $Y_i(a)$ is $L^2(P)$-bounded, i.e., $\mathbb{E}_P\left[|Y_i(a)|^2\right] < \infty$ for $a \in \{0,1\}$.

For any policy class Π_n, which may depend on n, the regret of deploying a policy $\pi \in \Pi_n$

relative to the best policy in Π_n, is defined as

$$\mathrm{Reg}(\pi, \Pi_n) = \max_{\pi' \in \Pi_n} \mathbb{W}_\alpha(\pi') - \mathbb{W}_\alpha(\pi).$$

When Π_n is clearly understood from the context, we write $\mathrm{Reg}(\pi) = \mathrm{Reg}(\pi, \Pi_n)$ for notational simplicity. Our primary result regarding the asymptotic regret of our α-EWM policy incorporates the following two key quantities:

$$\Xi := \sup_{\eta \in \mathcal{B}_Y} \mathbb{E}\left|\gamma_\eta(Z_i)\right|^2 \quad \text{and} \quad \Xi^\dagger := \sup_{\eta \in \mathcal{B}_Y} \mathbb{E}\left|\Gamma^\dagger_\eta(Z_i)\right|^2,$$

where

$$\gamma_\eta(z) := \tau(x,\eta) + \frac{a - e_o(x)}{e_o(x)\left(1 - e_o(x)\right)}\left\{(y-\eta)_- - \mu_a(x,\eta)\right\} \text{ and}$$
$$\gamma^\dagger_\eta(z) := \mu_0(x,\eta) + \frac{(1-a)}{1 - e_o\left(x\right)}\left\{(y-\eta)_- - \mu_0(x,\eta)\right\}.$$

Theorem 4.1. Suppose Assumption 2.1, Assumption 4.1, Assumption 4.2, and Assumption 4.3 hold. Let $\bar{K} = 3 + 2/\kappa$. If $\mathbb{E}\left|\gamma_\eta(Z_i)\right|^2 > c_o > 0$ for all $\eta \in \mathcal{B}_Y$, then for $\alpha \in (0,1)$, the following inequality holds:

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\mathrm{Reg}\left(\widehat{\pi}_n\right)\right]}{\sqrt{\mathrm{VC}(\Pi_n)/n}} \leq \frac{30}{\alpha}\sqrt{\Xi + \Xi^\dagger} + 72\sqrt{(\bar{K}/\alpha + 1)^2 + \Xi/\alpha^2}. \tag{4.3}$$

Theorem 4.1 complements Theorem 1 in Athey and Wager (2021) for 1-EWM policy.[3] The constant in Theorem 4.1 depends on α: it increases as α decreases, partly due to estimation error. Specifically, estimating the average welfare of the α-worst-affected group makes use of only an α-fraction of the total sample, leading to greater instability in welfare estimation.

Remark 4.2. Suppose that $\mathcal{B}_Y \subseteq [-\eta_B, \eta_B]$ for some $\eta_B > 0$. Under the strict overlap condition in Assumption 4.2, it follows that Ξ and $\Xi^\dagger$ can be upper bounded as

$$\Xi \leq \left(1 + \frac{2}{\kappa}\right)\left(\mathbb{E}|Y_i(0)|^2 + \mathbb{E}|Y_i(1)|^2 + \eta_B\right) \quad \text{and} \quad \Xi^\dagger \leq \left(1 + \frac{2}{\kappa}\right)\left(\mathbb{E}|Y_i(0)|^2 + \eta_B\right).$$

Remark 4.3. Recall that we learn the optimal policy by simultaneously solving out $\widehat{\pi}_n$ and $\widehat{\eta}_n$ from $\max_{(\pi,\eta)\in\Pi_n\times\mathcal{B}_Y} \widehat{\mathbb{V}}_n(\pi,\eta)$. Let $\widehat{\theta} \equiv (\widehat{\pi}, \widehat{\eta}) \in \Pi_n \times \mathcal{B}_Y$ denote any near-optimal solution satisfying

$$\widehat{\mathbb{V}}_n(\widehat{\theta}) \geq \sup_{\theta\in\Theta_n} \widehat{\mathbb{V}}_n(\theta) - o_P\left(r_n\right),$$

where $r_n = \sup_{\theta\in\Theta_n}\left|\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}_n(\theta)\right|$. In fact, Theorem 4.1 holds if the exact optimizer $\widehat{\pi}_n$ is replaced by any near-optimal welfare maximizer $\widehat{\pi}$ and $r_n = o_P(n^{-1/2})$. The term $o_P(r_n)$ enables us to find an approximate solution to $\max_{\theta\in\Theta_n} \widehat{\mathbb{V}}_n(\theta)$, which is particularly useful when the optimization is non-concave.

[3] Athey and Wager (2021) also allow for an approximate optimal policy.

4.3.1 Technical Comparisons with Kitagawa and Tetenov (2018); Athey and Wager (2021)

The regret bounds in Theorem 4.1 and those in Kitagawa and Tetenov (2018); Athey and Wager (2021) are all of order $\sqrt{\mathrm{VC}(\Pi_n)/n}$. In addition, Theorem 4.1 and Theorem 1 in Athey and Wager (2021) provide explicit expressions for the constants which require more delicate technical proofs than Kitagawa and Tetenov (2018, 2021).

Following Kitagawa and Tetenov (2018, 2021), the proof of the order of the regret bounds of $\widehat{\pi}_n$ relies on the lemma below.

Lemma 4.2. Suppose Assumption 2.1, Assumption 4.1 and Assumption 4.2 hold. If $b_o/2 > \zeta_e \wedge \kappa_\mu$, then

$$\mathrm{Reg}(\widehat{\theta}_n) \leq 2 \sup_{\theta \in \Theta_n} |(\mathbb{P}_n - P)\, g_\theta| + r_n,$$

where $r_n = o_P(n^{-1/2})$.

Lemma 4.2 implies that it is sufficient to study the concentration of the empirical process:

$$\mathbb{V}_n(\theta) - \mathbb{V}(\theta) = (\mathbb{P}_n - P) g_\theta \quad \text{over} \quad \theta \in \Theta_n.$$

In contrast to Kitagawa and Tetenov (2018, 2021) and Athey and Wager (2021), the score function for the α-expected welfare g_θ is nonlinear in θ rendering the VC dimension of the function class $\mathcal{G}_{\Theta_n} := \{g_\theta : \theta \in \Theta_n\}$ difficult to derive. Instead of exploiting the VC dimension of the corresponding function classes as in Kitagawa and Tetenov (2018, 2021) and Athey and Wager (2021), we directly upper bound the covering number of $\mathcal{G}_{\Theta_n}$ and then apply the classic empirical process maximal inequality, such as Theorem 2.14.1 in van der Vaart and Wellner (1998).

Lemma 4.3. If Assumption 4.3 holds, then there is an envelope function G for $\mathcal{G}_{\Theta_n}$ and constant $c_o > 0$ not depending on n and p such that

$$N\left(\epsilon \|G\|_{Q,2}, \mathcal{G}_{\Theta_n}, L^2(Q)\right) \leq (c_o/\epsilon)^{24\mathrm{VC}(\Pi_n)+48}, \quad \forall \epsilon > 0,$$

for all finite discrete probability measures Q on $\mathcal{Z}$.

Assumption 4.3 and Assumption 2.1 (2) ensure the existence of an envelope function that is bounded in $L^2(P)$. Applying Theorem 2.14.1 in van der Vaart and Wellner (1998) and Lemma 4.3, we conclude that there is a universal constant $c_o > 0$ not depending on n such that

$$\mathbb{E}_P\left[\sup_{\theta \in \Theta_n} |(\mathbb{P}_n - P) g_\theta|\right] \leq c_o \sqrt{\mathrm{VC}(\Pi_n)/n}. \tag{4.4}$$

Compared with Kitagawa and Tetenov (2018, 2021), one of the technical challenges addressed by Athey and Wager (2021) on 1-EWM policy lies in handling the doubly robust estimator of the welfare function. They show that as long as $\mathrm{VC}(\Pi_n)$ does not grow too

rapidly with n, the use of cross-fitting and ML/nonparametric estimation of nuisance parameters results in a regret bound of the order $\sqrt{\mathrm{VC}(\Pi_n)/n}$. Building on Kitagawa and Tetenov (2018, 2021) and Athey and Wager (2021) on 1-EWM policy, we establish an upper bound for α-EWM for any $\alpha \in (0,1)$ with an explicit expression for the constant c_o in Eq. (4.4). Similar to Athey and Wager (2021), we employ a classical chaining argument to derive an upper bound for the Rademacher complexity of the score function class. However, due to the nonlinearity of score function g_θ with respect to θ, the slicing technique used in Athey and Wager (2021) is difficult to implement. Instead, we introduce a new conditional semi-metric and apply the classical Dudley's chaining argument to directly bound the Rademacher complexity of $\mathcal{G}_{\Theta_n}$. We refer interested reader to Appendix D.5 for details.

5 Inference for the Optimal Welfare

In this section, we develop asymptotically valid inference for the optimal α-expected welfare. Compared with regret bounds, inference on optimal welfare is lacking even for 1-EWM except for the first-best policy; see Luedtke and van der Laan (2016, 2018); Shi et al. (2020), and Appendix B in the supplemental material to Kitagawa and Tetenov (2018).

We first impose conditions including the uniqueness of the optimal solution denoted as θ_o to ensure asymptotic normality of $\sup_{\theta\in\Theta} \widehat{\mathbb{V}}_n(\theta)$ based on which we construct Wald-type inference. We then summarize a general inference procedure that relaxes the uniqueness assumption. A detailed treatment of the general inference procedure is postponed to Appendix C.

For simplicity, we assume that the policy class does not change with the sample size n, i.e., $\Pi_n = \Pi$ for all n, and write $\Theta = \Pi \times \mathcal{B}_Y$. We define a metric space $(\Theta, \|\cdot\|)$, where

$$\|\theta_1 - \theta_2\| \equiv |\eta_1 - \eta_2| + \|\pi_1 - \pi_2\|_{P,2} = |\eta_1 - \eta_2| + \sqrt{\mathbb{E}|\pi_1(X_i) - \pi_2(X_i)|^2}.$$

for any $\theta_1, \theta_2 \in \Theta$. This premise will be upheld throughout the subsequent analysis.

5.1 Assumptions

We establish asymptotic normality under two assumptions, the bounded support assumption and the uniqueness assumption.

Assumption 5.1. (1) The outcome $Y_i = Y_i(A_i)$ has bounded support, i.e., $\mathbb{P}(|Y_i| \le c_o) = 1$ for some constant $c_o > 0$.

(2) The policy class Π has finite VC-dimension, i.e., $\mathrm{VC}(\Pi) < \infty$.

Assumption 5.1 is widely adopted in policy learning research, see, e.g., Kitagawa and Tetenov (2018, 2021); Rai (2018); Kallus and Zhou (2018); Luedtke and van der Laan (2016);

Luedtke and Chambaz (2020).[4] Assumption 5.1 (1) implies that the feasible set $\mathcal{B}_Y$ of the dual reformulation of $\mathbb{W}_\alpha(\pi)$ can be restricted to $[-c_o, c_o]$ and the regression functions $|\mu_a(x,\eta)| \leq 2c_o$ for all $\eta \in \mathcal{B}_Y$ and $a \in \{0,1\}$. Moreover, the functions $g_\theta(\cdot)$ are also uniformly bounded, i.e., $\sup_{\theta\in\Theta} \|g_\theta\|_\infty < \infty$.

Assumption 5.2 (Uniqueness)**.** There exists a $\theta_o \equiv (\pi_o, \eta_o) \in \Theta$ such that for all $\epsilon > 0$, $\mathbb{V}(\theta_o) > \sup\{\mathbb{V}(\theta) : \theta \in \Theta, \|\theta - \theta_o\| > \epsilon\}$.

Assumption 5.2 is a standard condition in extremum estimation. It ensures that $\theta_o \in \Theta$ is a unique and well-separated point of maximum of $\theta \mapsto \mathbb{V}(\theta)$. Lemma 14.4 in Kosorok (2008) gives some sufficient conditions for this assumption. If for all $\epsilon > 0$, $\mathbb{W}(\pi_o) > \sup_{\pi:\|\pi-\pi_o\|>\epsilon} \mathbb{W}(\pi)$ and $Y_i(\pi)$ has positive density at $\mathrm{VaR}_\alpha(Y_i(\pi))$ for all $\pi \in \Pi$, then Assumption 5.2 is satisfied. For policy learning, Assumption 5.2 is strong, although it is adopted in Wang et al. (2018), Section 2.3 of Kitagawa and Tetenov (2018), and Section 2.3 of Luedtke and Chambaz (2020).

Remark 5.1. For 1-EWM, uniqueness of the first-best optimal policy excludes a special class of distributions known as exceptional distributions. For α-EWM with $\alpha \in (0,1)$, we show in Lemma A.1 that the first best policy is given by $\pi_o = \mathbb{1}\{\tau(x,\eta_o) \geq 0\}$ with $\eta_o = \eta^*_{\mathrm{FB}}$ defined in Lemma A.1. Assumption 5.2 excludes the class of exceptional distributions for which $\mathbb{P}[\tau(X_i,\eta_o) = 0] > 0$. This is because Assumption 5.2 implies that $\theta_o = (\pi_o, \eta_o)$ is the unique and well separated maximizer. As a result,

$$\mathbb{1}\{\tau(X_i,\eta_o) \geq 0\} = \mathbb{1}\{\tau(X_i,\eta_o) > 0\}, \quad P\text{-a.s.},$$

and $\mathbb{P}(\tau(X_i,\eta_o) = 0) = 0$.

5.2 Asymptotic Normality

To establish asymptotic normality of $\widehat{\mathbb{V}}_n(\widehat{\theta}_n)$, consider the following decomposition:

$$\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}(\theta_o) = \underbrace{\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}_n(\widehat{\theta}_n)}_{=o_P(n^{-1/2})} + \underbrace{\mathbb{V}_n(\widehat{\theta}_n) - \mathbb{V}(\widehat{\theta}_n)}_{\approx \mathbb{V}_n(\theta_o) - \mathbb{V}(\theta_o)} + \underbrace{\mathbb{V}(\widehat{\theta}_n) - \mathbb{V}(\theta_o)}_{=-\mathrm{Reg}(\widehat{\pi}_n,\Pi)}. \tag{5.1}$$

Note that the first term on the RHS of Eq. (5.1) is $o_P(n^{-1/2})$ due to Lemma 4.1.

In the rest of this section, we will show that

(i) the second term on the RHS of Eq. (5.1) is asymptotically equivalent to $\mathbb{V}_n(\theta_o) - \mathbb{V}(\theta_o)$;

(ii) the third term on the RHS of Eq. (5.1) is of order $o_P(n^{-1/2})$.

[4] Although studies like Athey and Wager (2021) do not adopt this assumption for regret bounds, it substantially simplifies the technical analysis for statistical inference.

Consequently,

$$\begin{aligned}\sqrt{n}\left[\mathbb{V}_n(\widehat{\theta}_n) - \mathbb{V}(\theta_o)\right] &= \sqrt{n}\left[\mathbb{V}_n(\theta_o) - \mathbb{V}(\theta_o)\right] + o_P(1) \\ &= \sqrt{n}(\mathbb{P}_n - P)g_{\theta_o} + o_P(1).\end{aligned}$$

and asymptotic normality follows.

To show (i), we first prove $\|\widehat{\theta}_n - \theta_o\| = o_P(1)$ in Lemma 5.1 below. Since $\mathcal{G}_\Theta \equiv \{g_\theta : \theta \in \Theta\}$ is P-Donsker by Lemma 4.3, (i) follows.

Lemma 5.1. Under Assumption 2.1, Assumption 4.2, Assumption 5.1 and Assumption 5.2, it holds that $\|\widehat{\theta}_n - \theta_o\| = o_P(1)$.

To show (ii), we note that

$$\begin{aligned}\mathrm{Reg}(\widehat{\pi}_n) &= \mathbb{V}(\theta_o) - \mathbb{V}(\widehat{\theta}_n) = \mathbb{V}(\theta_o) - \widehat{\mathbb{V}}_n(\theta_o) + \widehat{\mathbb{V}}_n(\theta_o) - \widehat{\mathbb{V}}_n(\widehat{\theta}_n) + \widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}(\widehat{\theta}_n) \\ &\leq \mathbb{V}(\theta_o) - \mathbb{V}_n(\theta_o) + \mathbb{V}_n(\widehat{\theta}_n) - \mathbb{V}(\widehat{\theta}_n) + r_n \\ &= (\mathbb{P}_n - P)(g_{\widehat{\theta}_n} - g_{\theta_o}) + r_n,\end{aligned}$$

where the inequality follows from $\widehat{\mathbb{V}}_n(\theta_o) - \widehat{\mathbb{V}}_n(\widehat{\theta}_n) \leq 0$. Similar to Luedtke and Chambaz (2020), one can show that under mild conditions including boundedness and uniqueness, asymptotic equicontinuity arguments ensure that $(\mathbb{P}_n - P)(g_{\widehat{\theta}_n} - g_{\theta_o}) = o_P(n^{-1/2})$ for any policy class Π satisfying $\mathrm{VC}(\Pi) < \infty$.

Summing up, we obtain asymptotic normality of $\widehat{\mathbb{V}}_n(\widehat{\theta}_n)$.

Theorem 5.1. Suppose conditions in Lemma 5.1 hold. Then,

$$\begin{aligned}\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}_P(\theta_o) &= (\mathbb{V}_n - \mathbb{V}_P)(\theta_o) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{g_{\theta_o}(Z_i) - \mathbb{E}_P[g_{\theta_o}(Z_i)]\right\} + o_P(n^{-1/2}),\end{aligned}$$

where the function g_{θ_o}, defined in Eq. (3.2), is evaluated at $\theta = \theta_o$. In particular,

$$\sqrt{n}\left[\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}(\theta_o)\right] \rightsquigarrow N\left(0, \sigma_o^2\right),$$

where $\sigma_o^2 = \mathrm{Var}\left[g_{\theta_o}(Z_i)\right]$.

Remark 5.2. Drawing on Newey (1994); Luedtke and van der Laan (2016), and under the assumptions stated in Theorem 5.1 and other mild conditions, our optimal welfare estimator achieves semiparametric efficiency bound.

The next theorem presents a consistent estimator of the asymptotic variance σ_o^2.

Theorem 5.2. Consider the following estimator of σ_o^2:

$$\widehat{\sigma}_o^2 = \frac{1}{n}\sum_{i=1}^{n}\left[g_{\widehat{\theta}_n}\left(Z_i; \widehat{\mu}^{(-k(i))}, \widehat{e}^{(-k(i))}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} g_{\widehat{\theta}_n}\left(Z_i; \widehat{\mu}^{(-k(i))}, \widehat{e}^{(-k(i))}\right)\right]^2.$$

Under the conditions of Lemma 5.1, it holds that $\widehat{\sigma}_o^2 = \sigma_o^2 + o_P(1)$ and

$$\sqrt{n}\widehat{\sigma}_o^{-1}\left[\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}(\theta_o)\right] \rightsquigarrow N(0,1).$$

5.3 Uniform Inference

Appendix C develops uniform inference for the optimal welfare without Assumption 5.2. It improves upon the inference proposed in Appendix B in the supplemental material to Kitagawa and Tetenov (2018). We provide a summary of the procedures here and refer to interested reader to Appendix C for technical details.

Define the supremum functional $\psi : \ell^\infty(\Theta) \to \mathbb{R}$ as $\psi : h \mapsto \sup_{\theta\in\Theta} h(\theta)$. Consider the multiplier bootstrap $\widehat{\mathbb{G}}_n^* : \Theta \to \mathbb{R}$ defined as

$$\widehat{\mathbb{G}}_n^* : \theta \mapsto n^{-1/2}\sum_{i=1}^n \xi_i\left[\widehat{g}_\theta(Z_i) - \widehat{\mathbb{V}}_n(\theta)\right], \tag{5.2}$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. random variables independent of $(Z_i)_{i=1}^n$, with $\mathbb{E}(\xi_i) = 0$, $\mathbb{E}(\xi_i^2) = 1$ and $\mathbb{E}[\exp|\xi_i|] < \infty$. For given $\epsilon_n = o(1)$ with $n^{1/2}\epsilon_n \to \infty$, let

$$\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*) = \frac{\psi\big(\widehat{\mathbb{V}}_n + \epsilon_n\widehat{\mathbb{G}}_n^*\big) - \psi(\widehat{\mathbb{V}}_n)}{\epsilon_n}. \tag{5.3}$$

For any $\gamma \in (0,1)$, let c_γ denote the γ-empirical quantile of $\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*)$ which can be obtained from a large number of bootstrap samples. The one-sided confidence interval at the desired level γ is

$$\left[\sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta) - c_{1-\gamma}/\sqrt{n}, \infty\right), \tag{5.4}$$

with correct asymptotic coverage:

$$\lim_{n\to\infty}\inf_{P\in\mathcal{P}_n}\mathbb{P}\left[\mathbb{V}_P(\theta_o) \geq \sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta) - c_{1-\gamma}/\sqrt{n}\right] \geq 1-\gamma,$$

where $\mathcal{P}_n$ is a collection of distributions satisfying some regularity conditions specified in Assumption 5.1 in Appendix C. Define $q_{1-\gamma}$ as the $(1-\gamma)$-empirical quantile of $\left|\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*)\right|$ for any $\gamma > 0$. The corresponding two-sided confidence interval is

$$\left[\sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta) - q_{1-\gamma}/\sqrt{n}, \sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta) + q_{1-\gamma}/\sqrt{n}\right], \tag{5.5}$$

which attains the correct asymptotic coverage for any fixed distribution $P \in \mathcal{P}_n$:

$$\liminf_{n\to\infty}\mathbb{P}\left[\left|\sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}(\theta_o)\right| \leq q_{1-\gamma}/\sqrt{n}\right] \geq 1-\gamma.$$

6 Empirical Application and Simulations

This section presents extensive numerical results on the finite sample performance of our debiased estimator and proposed inference using both real data and synthetic data.[5]

6.1 The JTPA Study

Kitagawa and Tetenov (2018) apply 1-EWM method to experimental data from the National Job Training Partnership Act (JTPA) Study. The study randomized whether applicants are eligible to receive training and job-search assistance provided by the JTPA. The pre-treatment covariates included in the data are years of education (*edu*) and pre-program earnings (*prevearn*) and the outcome variable is an applicant's earnings 30 months after the assignment (*earnings*). The sample size is 9,223 and the propensity score is known to be 2/3. We adopt this data studied by Kitagawa and Tetenov (2018) and, similar to Kitagawa and Tetenov (2018), we analyze welfare from an intent-to-treat standpoint, considering hypothetically making available the training program to eligible individuals, who may decline it. For detailed data description and evaluation of average program effects, we refer the reader to Bloom et al. (1997).

We consider three policy classes: simple (treat all or none) and linear with and without squared and cubic *edu*. More specifically, the two linear policy classes take the form

$$\Pi_{\text{LES}} := \left\{ \{x : \beta_0 + \beta_1 edu + \beta_2 prevearn > 0\}, (\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3 \right\} \text{ and} \tag{6.1}$$

$$\Pi^3_{\text{LES}} := \left\{ \begin{array}{c} \{x : \beta_0 + \beta_1 edu + \beta_2 prevearn + \beta_3 edu^2 + \beta_4 edu^3 > 0\}, \\ (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \in \mathbb{R}^5 \end{array} \right\}. \tag{6.2}$$

We investigate $\alpha \in \mathcal{A} := \{0.25, 0.3, 0.4, 0.5, 0.8\}$. We recommend that researchers interested in the $\alpha = 1$ case consider the 1-EWM in Kitagawa and Tetenov (2018) directly. For each $\alpha \in \mathcal{A}$ and policy class, we estimate $\mu_a(x, \eta) = \mathbb{E}\left[(Y_i(a) - \eta)_- \mid X_i = x\right]$ for $a \in \{0, 1\}$ and a given η, using random forests (RF) developed by Athey et al. (2019). We then apply simulated annealing (SA), proposed by Kirkpatrick et al. (1983), to select the combination of parameters that (approximately) maximizes the objective function.[6] SA is a derivative-free probabilistic optimization algorithm aiming at finding approximate solutions by iteratively exploring the solution space and gradually decreasing the probability of accepting worse solutions as the algorithm progresses.[7]

[5] Data and codes for this section can be accessed at https://github.com/yqi3/alpha-EWM.

[6] We build RF using `regression_forest()` in `R` package `grf` and implement SA using `optim_sa()` in the `R` package `optimization` (Athey et al., 2019; Husmann et al., 2017). We use default tuning parameters for RF. For SA, the specifications are more problem-specific. A good strategy is to plot the loss function and inspect if there is sufficient evidence of convergence.

[7] Geman and Geman (1984) prove convergence of *generic* SA to a global optimum, provided that the probability of accepting worse solutions shrinks sufficiently slowly, and that all elements in the solution space are equally probable as the number of training epochs goes to infinity.

Estimation and inference results for $\mathbb{W}_\alpha(\pi_o)$ are organized in Table 1. The first two columns consist of the class of simple policies and serve as baselines for Π_{LES} and Π^3_{LES} in the third and fourth columns. Detailed expressions for the optimal policies can be found in Appendix I.1. The observed increase in $\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$ across panels reflects that, as α grows, the lower-tail subpopulation expands to include relatively better outcomes. This raises the average and thus increases the α-expected welfare. The percentage of treated individuals tends to increase with α as well. The 95% confidence intervals (CIs) constructed using normal inference in Algorithm 1 are reported in the third row of each panel in Table 1, and the 95% CIs from uniform inference obtained via multiplier bootstrap with $\epsilon = n^{-1/4}$ and $B = 100$ are presented in the last row of each panel. For each combination of α and policy class, the CI from uniform inference is wider than that from normal inference. While we cannot verify uniqueness, a simulation study calibrated to the JTPA sample in Section 6.2 finds that the Wald-type CIs achieve approximately 95% coverage, offering supporting evidence for their validity in this application.

	Treat None	**Treat All**	**Linear**	**Linear with edu^2 and edu^3**
Panel 1: $\alpha = 0.25$				
% treated	0%	100%	34.761%	32.896%
$\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$	\$376.968	\$451.027	\$530.630	\$546.300
95% CI (normal)	(\$298.567, \$455.368)	(\$372.626, \$529.427)	(\$439.331, \$621.930)	(\$446.461, \$646.138)
95% CI ($\epsilon = n^{-\frac{1}{4}}$)	(−\$48.098, \$802.033)	(\$154.412, \$747.641)	(\$146.400, \$914.860)	(\$155.773, \$936.826)
Panel 2: $\alpha = 0.3$				
% treated	0%	100%	50.992%	32.820%
$\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$	\$695.647	\$838.930	\$917.718	\$918.011
95% CI (normal)	(\$585.617, \$805.678)	(\$728.900, \$948.961)	(\$793.695, \$1041.741)	(\$776.708, \$1059.315)
95% CI ($\epsilon = n^{-\frac{1}{4}}$)	(\$152.922, \$1238.373)	(\$490.538, \$1187.322)	(\$457.579, \$1377.858)	(\$506.934, \$1329.088)
Panel 3: $\alpha = 0.4$				
% treated	0%	100%	82.392%	81.969%
$\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$	\$1647.506	\$1947.011	\$2038.321	\$2039.468
95% CI (normal)	(\$1468.631, \$1826.381)	(\$1768.137, \$2125.886)	(\$1845.888, \$2230.754)	(\$1840.260, \$2238.676)
95% CI ($\epsilon = n^{-\frac{1}{4}}$)	(\$995.201, \$2299.812)	(\$1519.072, \$2374.951)	(\$1477.132, \$2599.510)	(\$1516.364, \$2562.573)
Panel 4: $\alpha = 0.5$				
% treated	0%	100%	83.400%	83.379%
$\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$	\$2981.034	\$3419.311	\$3524.651	\$3527.108
95% CI (normal)	(\$2746.431, \$3215.638)	(\$3184.708, \$3653.915)	(\$3274.440, \$3774.861)	(\$3269.096, \$3785.121)
95% CI ($\epsilon = n^{-\frac{1}{4}}$)	(\$2233.145, \$3728.923)	(\$2910.270, \$3928.352)	(\$2951.684, \$4097.617)	(\$2898.115, \$4156.101)
Panel 5: $\alpha = 0.8$				
% treated	0%	100%	86.783%	79.204%
$\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$	\$8671.975	\$9522.451	\$9661.526	\$9690.607
95% CI (normal)	(\$8326.551, \$9017.398)	(\$9177.028, \$9867.874)	(\$9292.969, \$10030.082)	(\$9309.569, \$10071.646)
95% CI ($\epsilon = n^{-\frac{1}{4}}$)	(\$7816.114, \$9527.835)	(\$8876.617, \$10168.285)	(\$8940.210, \$10382.840)	(\$8983.668, \$10397.546)

Table 1: Estimated $\mathbb{W}_\alpha(\pi_o)$ for different α's and policy classes that condition on *edu* and *prevearn*. Baseline results for treating none or all of the individuals are shown in the first two columns. The third and fourth rows of each panel report the 95% CI based on normal and uniform inference, respectively.

Examining the point estimates of welfare, we see that for all $\alpha \in \mathcal{A}$, a simple policy of

treating all outperforms treating none. Moreover, relative to treating all, there is a considerable increase in the targeted welfare generated by the optimal policy of class Π_{LES}. Linear policies with edu^2 and edu^3 only bring tiny welfare improvements. Figures 2 and 3 highlight the optimal treatment regions. Following Kitagawa and Tetenov (2018), we bin the individuals by $(edu, prevearn)$, and the number of individuals with each combined characteristic is represented by the size of the corresponding dot.

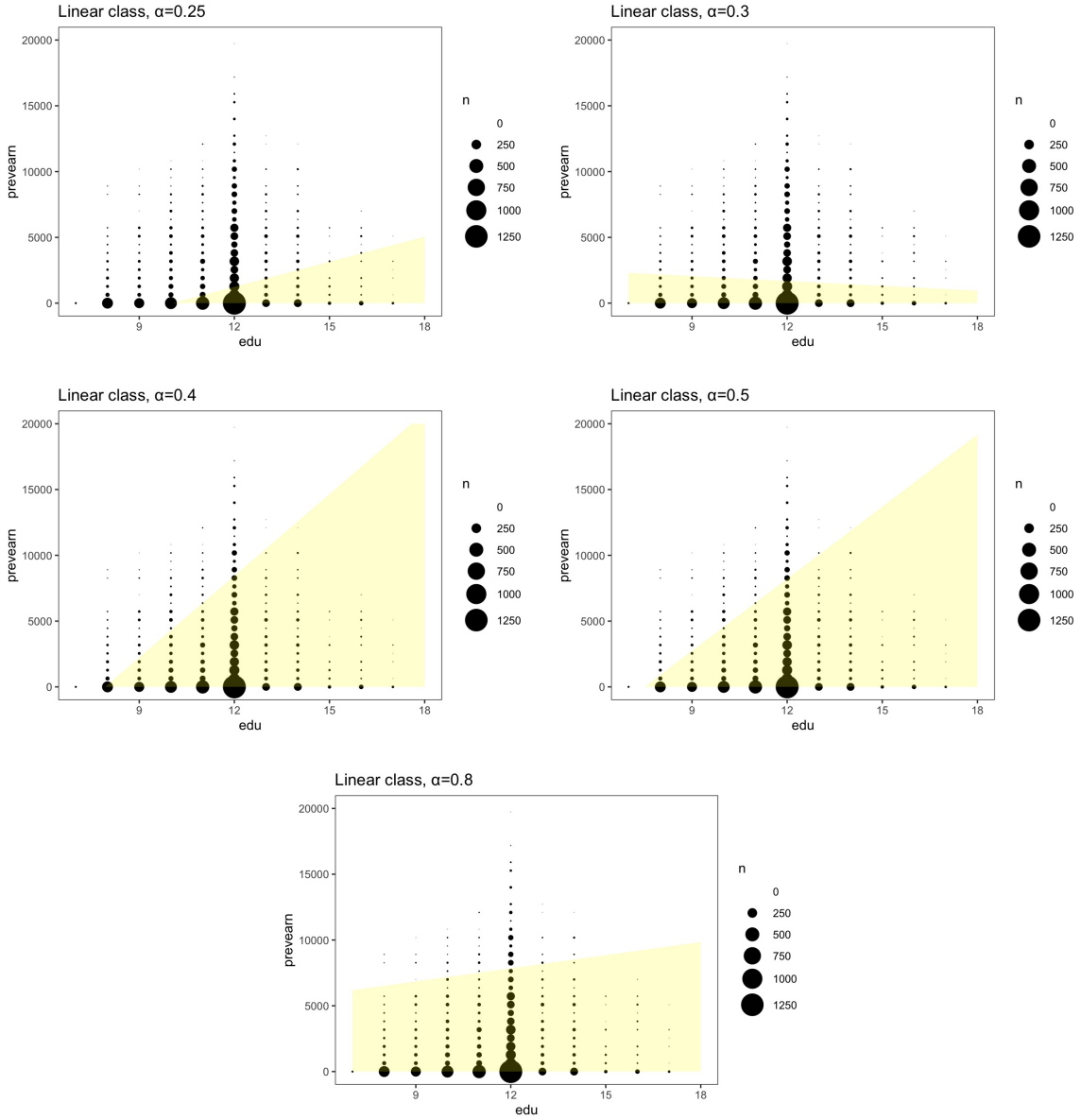

Figure 2: Optimal policies from the linear class Π_{LES} conditioning on *edu* and *prevearn*. The number of individuals with characteristics closest to each $(edu, prevearn)$ in the grid is represented by the size of the corresponding dot. $\alpha \in \{0.25, 0.3, 0.4, 0.5, 0.8\}$.

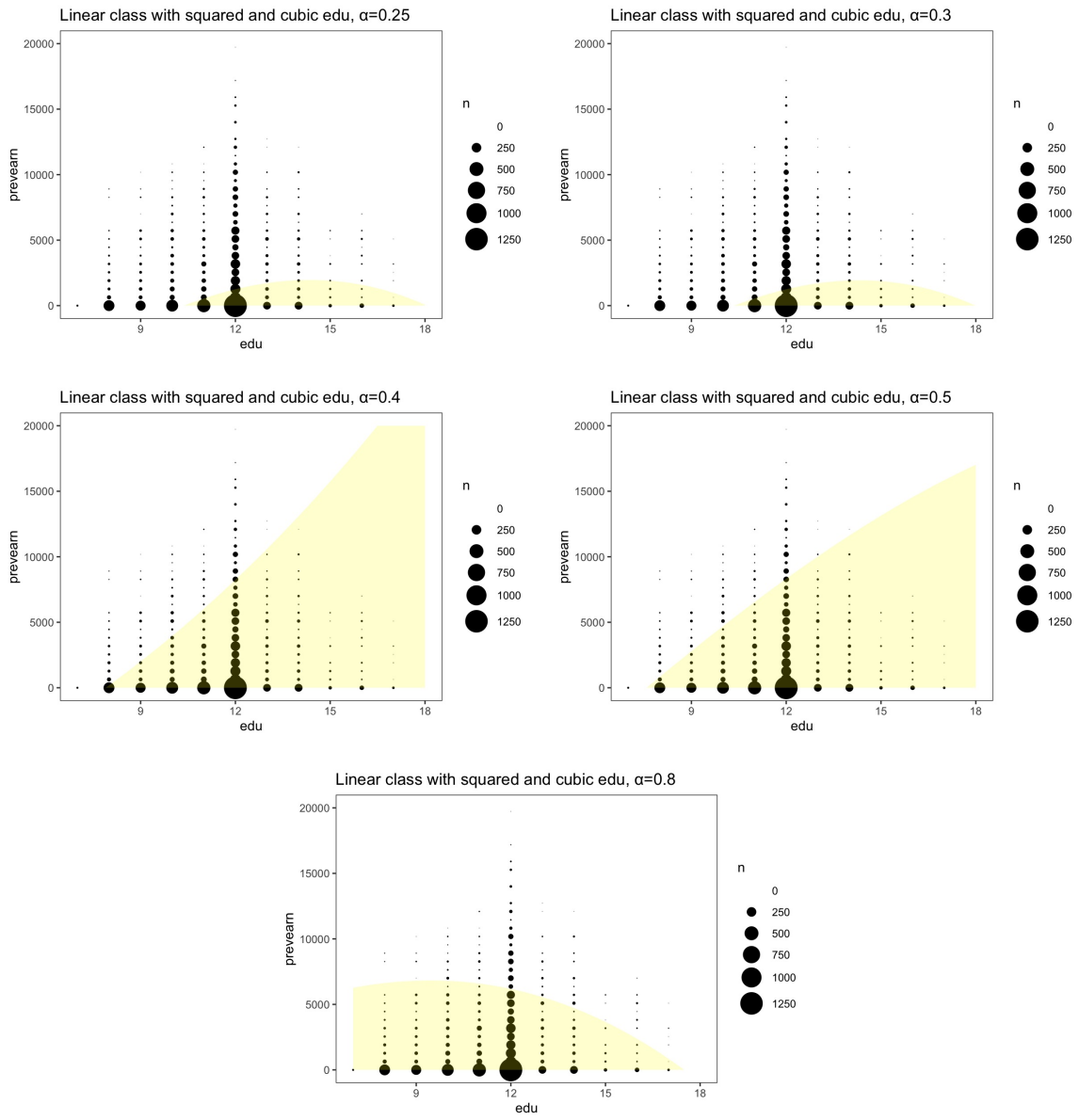

Figure 3: Optimal policies from the linear class Π^3_{LES} conditioning on *edu*, *prevearn*, edu^2, and edu^3. The number of individuals with characteristics closest to each $(edu, prevearn)$ in the grid is represented by the size of the corresponding dot. $\alpha \in \{0.25, 0.3, 0.4, 0.5, 0.8\}$.

Tables 2 and 3 examine welfare gains and losses as we switch between different targeting policies and estimate the resulting welfare of different targeted subpopulations. For example, the first row in Table 2 shows the estimated welfare of the worst-off 25% of the population when the optimal linear policies are targeting the worst-off 25%, 30%, 40%, 50%, and 80%, respectively. The diagonal entries (i.e., the row maximums) are highlighted as these optimal policies are targeting the actual subpopulations of interest. Tables 2 and 3 demonstrate a valuable strength of our method, as we are able to conduct rich policy evaluations by estimating the expected welfare at any α for any given policy. In other words, even when a

policy is not targeting the worst-affected $(\alpha \times 100)\%$, we can still evaluate its performance at α to obtain a clear picture of the trade-offs, which opens up possibilities for learning policies that promote greater equality across subpopulations.

From Table 2 below and Table 6 in Appendix I.1, adopting the linear policy that targets $\alpha' = 0.8$ leads to an 11.9% decrease in the average welfare of the worst-affected quarter of the population ($\alpha = 0.25$), compared to implementing the optimal linear policy *targeting* the worst-affected quarter ($\alpha = \alpha' = 0.25$). Conversely, adopting the policy targeting the worst-affected quarter ($\alpha' = 0.25$) only leads to a 5.3% decrease in the 0.8-expected welfare ($\alpha = 0.8$) relative to implementing the optimal policy targeting the worst-affected 80% ($\alpha = \alpha' = 0.8$). In Table 3, similar patterns emerge with the inclusion of edu^2 and edu^3 in treatment assignment. Based on Tables 2 and 3, Tables 6 and 7 in Appendix I.1 report the percentage welfare loss for every combination of actual α and α' for policy selection. A notable observation is that the bottom quarter of the population is particularly vulnerable when the policy targets some $\alpha' \geq 0.4$ instead. Thus, policymakers aspiring for greater equality should prioritize smaller levels of α, such as 0.25 or 0.3, as evidenced by the small percentage welfare losses in the first two columns of Tables 6 and 7, all of which are below 5.5%.

α of Interest \ α' for Policy Selection	0.25	0.3	0.4	0.5	0.8
0.25	530.630 (46.581)	525.116 (48.020)	500.874 (42.561)	495.241 (45.623)	467.467 (41.415)
0.3	898.609 (65.824)	917.718 (63.277)	908.589 (62.561)	896.640 (62.399)	862.059 (57.638)
0.4	1944.643 (108.718)	2020.792 (105.008)	2038.321 (98.180)	2035.307 (99.732)	1992.898 (96.564)
0.5	3331.114 (147.675)	3485.067 (141.436)	3522.112 (131.105)	3524.651 (127.658)	3493.405 (131.633)
0.8	9146.288 (215.680)	9451.340 (209.083)	9552.165 (185.883)	9588.269 (185.388)	9661.526 (188.039)

Table 2: Estimated $\mathbb{W}_\alpha(\pi_o)$ for different actual α's of interest and α''s for linear policy selection (policy class Π_{LES}). Standard errors are reported in parentheses, and all values are in USD.

α of Interest \ α' for Policy Selection	0.25	0.3	0.4	0.5	0.8
0.25	546.300 (50.938)	543.405 (51.359)	504.095 (44.861)	496.645 (45.001)	476.020 (42.162)
0.3	917.043 (68.319)	918.011 (72.094)	910.930 (62.953)	897.083 (61.785)	871.931 (61.71)
0.4	1972.299 (109.090)	1974.302 (109.322)	2039.468 (101.637)	2036.425 (100.647)	2004.521 (102.276)
0.5	3364.695 (144.252)	3369.302 (146.286)	3525.834 (130.284)	3527.108 (131.639)	3509.814 (134.797)
0.8	9197.693 (216.352)	9191.845 (214.680)	9555.919 (186.193)	9598.417 (185.961)	9690.607 (194.407)

Table 3: Estimated $\mathbb{W}_\alpha(\pi_o)$ for different actual α's of interest and α''s for linear policy selection with edu^2 and edu^3 (policy class Π^3_{LES}). Standard errors are reported in parentheses, and all values are in USD.

6.2 Simulations Based on WGAN-Generated JTPA Data

We next present simulation results based on a superpopulation generated using Wasserstein Generative Adversarial Networks (WGANs) to evaluate the finite-sample performance of our debiased estimator. We focus on this simulation setup in the main text because the generated data more closely resembles real-world data distributions, making it more illustrative of practical applications. For comparison, we also conduct two additional simulation studies inspired by the DGPs in Athey and Wager (2021), with adjustments that make the treatment assignment exogenous. Since the results across all three designs are qualitatively similar—our estimator consistently exhibits decreasing mean squared error as the sample size increases, and the coverage rates approach the nominal 95% level in larger samples—we relegate the latter two studies to Appendix I.3.

In all three simulation setups, the propensity scores are assumed to be known, i.e., $\widehat{e}(\cdot) = e(\cdot)$. Cases with unknown propensity scores can be analyzed analogously using an estimator $\widehat{e}(\cdot)$ that satisfies Assumption 4.2. Since uniform inference based on the multiplier bootstrap is computationally intensive, we report only the coverage rates based on confidence intervals constructed via Wald inference. We examine values of $\alpha \in \mathcal{A}$ considered in Section 6.1.

We employ WGANs developed by Athey et al. (2024) to construct a hypothetical superpopulation, referred to as WGAN-JTPA, consisting of one million observations based on the JTPA data in Section 6.1. As mentioned by Athey et al. (2024), a benefit of using WGAN-generated data for simulations is that this practice largely rules out the possibility for researchers to choose particular DGPs that favor their proposed methods. This subsection demonstrates robust performance of our debiased estimator even when the underlying superpopulation is built from real datasets like the JTPA, which has highly skewed outcome and covariate distributions. Appendix I.2 discusses the training process in more detail and presents some summary statistics.

While technical details of WGANs can be found in Athey et al. (2024), we highlight that to build the superpopulation, since we generate $X|A$ followed by $Y|(X, A)$ and apply the same generator on $(X, 1-A)$ to obtain $Y|(X, 1-A)$, both potential outcomes are available for each individual. As a result, we can directly compute the true expected welfare at any α induced by any policy, which is simply a tail average of post-treatment outcomes. For each $\alpha \in \mathcal{A}$, we run SA to find a linear policy $\pi_o \in \Pi_{\text{LES}}$ (as defined in (6.1)) that maximizes $\mathbb{W}_\alpha(\pi)$ and treat the resulting optimum $\mathbb{W}_\alpha(\pi_o)$ as the population truth.

As an illustration, we use WGAN-JTPA to compare the 0.25-EWM policy with the 1-EWM (mean-optimal) and equality-minded (standard Gini social welfare-optimal) policies. Inspired by Figure 3 in Kitagawa and Tetenov (2021), Figure 4 plots the between-quantile differences in post-treatment outcomes across these policies. The figure shows that both the 0.25-EWM and equality-minded policies raise the welfare of lower-ranked individuals while lowering the welfare of higher-ranked individuals relative to the 1-EWM policy at the population level, with the 0.25-EWM policy placing much greater emphasis on these adjustments.

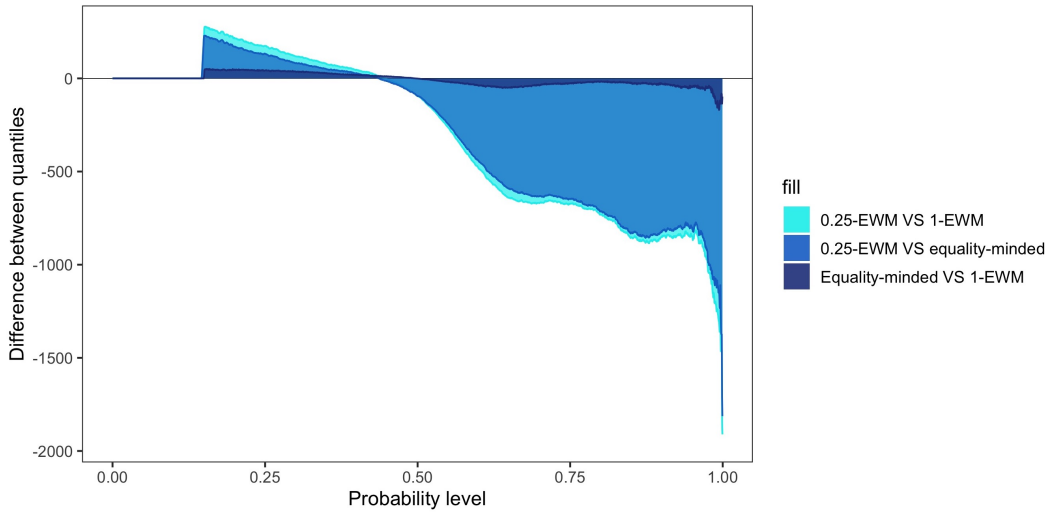

Figure 4: Between-quantile differences in outcomes for the 0.25-EWM, 1-EWM, and equality-minded policies using the WGAN-JTPA data.

In the simulations, for each replicate, we draw a sample of size $n \in \{2000, 5000, 10000\}$ without replacement from WGAN-JTPA. The propensity score is fixed at the population mean of A, which is approximately 0.66475.[8] For each pair (n, α), we apply Algorithm 1 to 1,000 sample draws and organize the results in Table 4. As shown by the marginal histogram for *earnings* in Figure 5 in Appendix I.2, WGAN-JTPA inherits the high skewness present in the original JTPA data. Consequently, larger sample sizes are required to achieve satisfactory coverage. From Table 4, our optimal welfare estimator achieves acceptable coverage when $n = 5{,}000$, which is a realistic sample size for both experimental and observational studies (for reference, the original JTPA sample used by Kitagawa and Tetenov (2018) contains 9,223 observations).

7 Concluding Remarks

The α-expected welfare function considered in this paper offers a flexible interpolation between the Rawlsian welfare ($\alpha \to 0$) and the empirical welfare maximization ($\alpha = 1$) approach proposed by Kitagawa and Tetenov (2018). Like Athey and Wager (2021) for the empirical welfare maximization, our development of the doubly robust scores facilitates asymptotic inference for the optimal welfare and allows practitioners flexibility in how they estimate the nuisance parameters. Besides learning the optimal policies, our estimation strategy also enables more thorough policy evaluations by computing the average welfare of the worst-affected subpopulation of any size (fraction of the population). In addition to establishing regret bounds for the debiased estimator, we also develop inference for the optimal α-expected

[8] This is very close to the mean of A in the actual JTPA data, 0.66497. In the JTPA Study, treatment was randomized with probability 2/3, and we assume randomized treatment in WGAN-JTPA as well.

Sample size	**2,000**	**5,000**	**10,000**
Panel 1: $\alpha = 0.25$, truth = 1119.195			
Avg. % treated using $\widehat{\pi}_n$	52.655%	54.893%	55.310%
Bias	191.791	82.014	43.564
Variance	46486.298	18778.912	9049.531
MSE	83269.985	25505.129	10947.367
95% Coverage	93.1%	93.7%	94.9%
Panel 2: $\alpha = 0.3$, truth = 1908.135			
Avg. % treated using $\widehat{\pi}_n$	53.739%	54.245%	56.121%
Bias	206.873	96.268	48.651
Variance	55137.705	22734.450	11799.126
MSE	97934.121	32001.932	14166.024
95% Coverage	92.0%	94.3%	94.8%
Panel 3: $\alpha = 0.4$, truth = 3460.773			
Avg. % treated using $\widehat{\pi}_n$	55.863%	57.153%	58.069%
Bias	229.273	101.223	48.033
Variance	59133.582	24046.124	13456.291
MSE	111699.467	34292.263	15763.427
95% Coverage	91.8%	93.9%	94.3%
Panel 4: $\alpha = 0.5$, truth = 4867.556			
Avg. % treated using $\widehat{\pi}_n$	58.165%	60.027%	61.596%
Bias	204.781	96.355	49.745
Variance	58786.097	22457.617	12335.452
MSE	100721.497	31741.832	14810.019
95% Coverage	92.3%	94.4%	95.3%
Panel 5: $\alpha = 0.8$, truth = 9475.336			
Avg. % treated using $\widehat{\pi}_n$	75.955%	83.083%	88.094%
Bias	210.888	92.727	52.521
Variance	80007.017	33274.611	16694.598
MSE	124480.959	41872.844	19453.079
95% Coverage	93.5%	93.9%	95.3%

Table 4: Simulation results based on WGAN-JTPA data (1,000 replications). All quantities inherit the units of USD from the empirical data; units are omitted for brevity.

welfare for any $\alpha \in (0, 1)$. Results from extensive numerical studies based on both JTPA data and simulated data demonstrate the efficacy and practical value of policy learning through α-EWM.

We are currently working on several extensions of this paper. Methodologically, it is important to develop statistical tests to compare whether one policy is superior to another. Practically, it would be beneficial to determine who is actually targeted by the optimal policy. For example, what characteristics do the worst-affected individuals have? Information like this could present a more comprehensive picture of the relevant population and promote the design of more equitable policies.

References

Adjaho, C. and Christensen, T. (2022). Externally valid treatment choice. *arXiv preprint arXiv:2205.05561*, 1.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Andrews, D. W. (1994). Empirical process methods in econometrics. *Handbook of econometrics*, 4:2247–2294.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2024). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 240(2):105076.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106:1039–1082.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997). The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of Human Resources*, 32(3):549–576.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. 42(4):1564–1597.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Cui, Y. and Han, S. (2023). Individualized treatment allocations with distributional welfare. *arXiv preprint arXiv:2311.15878*.

Duchi, J., Hashimoto, T., and Namkoong, H. (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664.

Durrett, R. (2019). *Probability: Theory and Examples*, volume 49. Cambridge university press.

Fan, Y., Park, H., and Xu, G. (2023). Quantifying distributional model risk in marginal problems via optimal transport. *arXiv preprint arXiv:2307.00779*.

Fang, E. X., Wang, Z., and Wang, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, 118(543):1733–1746.

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Firpo, S., Galvao, A. F., and Parker, T. (2023). Uniform inference for value functions. *Journal of Econometrics*, 235(2):1680–1699.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921.

Giné, E. and Nickl, R. (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge university press.

Greselin, F. and Zitikis, R. (2018). From the classical gini index of income inequality to a new zenga-type relative measure of risk: A modeller's perspective. *Econometrics*, 6(1):4.

Hong, H. and Li, J. (2018). The numerical delta method. *Journal of Econometrics*, 206(2):379–394.

Husmann, K., Lange, A., and Spiegel, E. (2017). The r package optimization: Flexible global optimization with simulated-annealing. *CRAN citation*.

Kallus, N. (2018). Balanced policy evaluation and learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8909–8920.

Kallus, N. and Zhou, A. (2018). Confounding-robust policy improvement. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9289–9299.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. *Statistical Causal Inferences and Their Applications in Public Health Research*, page 141.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219.

Kim, K. and Zubizarreta, J. (2023). Fair and robust estimation of heterogeneous treatment effects for policy learning. *arXiv preprint arXiv:2306.03625*.

Kirkpatrick, S., Gelatt Jr, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616. Supplementary materials and online appendix available at https://doi.org/10.3982/ECTA13288.

Kitagawa, T. and Tetenov, A. (2021). Equality-minded treatment choice. *Journal of Business & Economic Statistics*, 39(2):561–574.

Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Lei, L., Sahoo, R., and Wager, S. (2023). Policy learning under biased sample selection. *arXiv preprint arXiv:2304.11735*.

Luedtke, A. and Chambaz, A. (2020). Performance guarantees for policy learning. *Annales de L'Institut Henri Poincare Section (B) Probability and Statistics*, 56(3):2162–2188.

Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2):713–742.

Luedtke, A. R. and van der Laan, M. J. (2018). Parametric-rate inference for one-sided differentiable parameters. *Journal of the American Statistical Association*, 113(522):780–788.

Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–82.

Qi, Z., Pang, J.-S., and Liu, Y. (2023). On robustness of individualized decision rules. *Journal of the American Statistical Association*, 118(543):2143–2157.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210.

Rai, Y. (2018). Statistical inference for treatment assignment policies. *Unpublished Manuscript*.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Harvard University Press.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42.

Rockafellar, R. T., Uryasev, S. P., and Zabarankin, M. (2002). Deviation measures in risk analysis and optimization. *University of Florida, Department of Industrial & Systems Engineering Working Paper*, (2002-7).

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 6(1):34–58.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2021). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Shi, C., Lu, W., and Song, R. (2018). A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709.

Shi, C., Lu, W., and Song, R. (2020). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21(176):1–67.

Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, 50(197):3–17.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner (1998). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

van der Vaart, A. W. and Wellner, J. A. (2011). A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192.

Viviano, D. and Bradic, J. (2024). Fair policy targeting. *Journal of the American Statistical Association*, 119(545):730–743.

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhao, P. and Cui, Y. (2023). A semiparametric instrumented difference-in-differences approach to policy learning. *arXiv preprint arXiv:2310.09545*.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhou, Z., Athey, S., and Wager, S. (2023). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183.

A First Best α-EWM Policy

It is insightful to compare the first-best (FB) policies based on expected welfare function and the AVaR welfare function for $\alpha \in (0,1)$. In EWM, the FB policy is

$$\mathbb{1}\left\{x \in \mathcal{X} : \tau(x) > 0\right\},$$

where $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x\right]$ is the CATE. We now provide a similar representation of the FB policy in our set-up. The FB (optimal) policy is defined as

$$\pi^*_{\mathrm{FB}} \in \operatorname{argmax}_{\pi \in \Pi_o} \mathbb{W}_\alpha(\pi).$$

We assume the existence of π^*_{FB}, which maximizes the average welfare of the size-α lowest-ranked subpopulation.

Recall μ_1, μ_0, and τ defined in Section 3. Under Assumption 2.1, for any given η, $\tau(x,\eta)$ is identified. Moreover, let $\chi_1(\eta) \equiv \mathbb{E}\left[\mu_1(X_i,\eta)\mathbb{1}\{\tau(X_i,\eta) \geq 0\}\right]$ and $\chi_0(\eta) \equiv \mathbb{E}\left[\mu_0(X_i,\eta)\mathbb{1}\{\tau(X_i,\eta) < 0\}\right]$.

Lemma A.1. Suppose the functions $\chi_0(\cdot)$ and $\chi_1(\cdot)$ are continuous. Then, for each $\alpha \in (0,1]$, there is a constant η^*_{FB} depending on α such that the policy given by

$$\pi^*_{\mathrm{FB}}(x) = \mathbb{1}\left\{\tau\left(x, \eta^*_{\mathrm{FB}}\right) > 0\right\},$$

maximizes $\mathbb{W}_\alpha(\pi)$ over $\pi \in \Pi_o$.

Proof. From Lemma 3.1 and Remark 3.1, it follows that

$$\mathbb{W}_\alpha(\pi) = \mathrm{AVaR}_\alpha(Y_i(\pi)) = \sup_{\eta \in \mathcal{B}_Y} \left\{\frac{1}{\alpha}\mathbb{E}\left[(Y_i(\pi) - \eta)_-\right] + \eta\right\}. \tag{A.1}$$

Hence,

$$\begin{aligned}\sup_{\pi \in \Pi_o} \mathbb{W}_\alpha(\pi) &= \sup_{\pi \in \Pi_o} \sup_{\eta \in \mathcal{B}_Y} \left\{\frac{1}{\alpha}\mathbb{E}\left[(Y_i(\pi) - \eta)_-\right] + \eta\right\} \\ &= \sup_{\eta \in \mathcal{B}_Y} \sup_{\pi \in \Pi_o} \left\{\frac{1}{\alpha}\mathbb{E}\left[(Y_i(\pi) - \eta)_-\right] + \eta\right\}.\end{aligned}$$

For a fixed $\eta \in \mathcal{B}_Y$, consider the following maximization:

$$\sup_{\pi \in \Pi_o} \frac{1}{\alpha}\mathbb{E}\left[(Y_i(\pi) - \eta)_-\right] + \eta.$$

An optimal solution to this problem is given by $\pi_\eta^*(x) = \mathbb{1}\{\tau(x,\eta) \geq 0\}$. As a result,

$$\begin{aligned}
\sup_{\pi\in\Pi_o} \mathbb{W}_\alpha(\pi) &= \sup_{\eta\in\mathcal{B}_Y} \left\{ \frac{1}{\alpha}\mathbb{E}\left[\left(Y_i(\pi_\eta^*) - \eta\right)_-\right] + \eta \right\} \\
&= \sup_{\eta\in\mathcal{B}_Y} \left\{ \frac{1}{\alpha}\mathbb{E}\left[(Y_i(1) - \eta)_- \mathbb{1}\{\tau(X_i,\eta) \geq 0\} + (Y_i(0) - \eta)_- \mathbb{1}\{\tau(X_i,\eta) < 0\}\right] + \eta \right\} \\
&= \sup_{\eta\in\mathcal{B}_Y} \left\{ \frac{1}{\alpha}\mathbb{E}\left[\mu_1(X_i,\eta)\mathbb{1}\{\tau(X_i,\eta) \geq 0\} + \mu_0(X_i,\eta)\mathbb{1}\{\tau(X_i,\eta) < 0\}\right] + \eta \right\}.
\end{aligned}$$

Since χ_0 and χ_2 are continuous, the function $\frac{1}{\alpha}\mathbb{E}\left[\left(Y_i(\pi_\eta^*) - \eta\right)_-\right] + \eta$ is also continuous in η. Consequently, it attains its maximum at η_{FB}^* over the compact set $\mathcal{B}_Y$. Therefore, we conclude that $\mathbb{1}\{\tau(x,\eta_{\text{FB}}^*) \geq 0\}$ is the FB policy. □

When $\alpha = 1$, our FB policy $\pi_{\text{FB}}^*(\cdot)$ reduces to $\mathbb{1}\{\tau(x) \geq 0\}$, the FB policy under EWM. When $\alpha \in (0,1)$, $\pi_{\text{FB}}^*(\cdot)$ depends on the distribution of post-treatment outcomes through the optimal cutoff η_{FB}^*.

B Improved Rate Under the Margin Assumption

In this section, we demonstrate that the asymptotic regret bound presented in Theorem 4.1 can be further tightened under the margin assumption, a commonly adopted condition in the statistical learning literature. Throughout this subsection, we continue to uphold Assumption 5.1.

B.1 Curvature or Margin Assumption

Since θ_o is the unique maximizer of $\mathbb{V}(\theta)$, the first order derivative of $\mathbb{V}(\theta)$ should vanish at θ_o and the second-order derivative should be negative definite. Motivated by this intuition, we introduce the following curvature (or margin) assumption.

Assumption B.1 (Curvature)**.** Suppose there exist constants $\rho_o \geq 1$ and $c_o > 0$ such that for every θ in some nonempty neighborhood of θ_o, the following inequality holds:

$$\mathbb{V}(\theta_o) - \mathbb{V}(\theta) \geq c_o \left\|\theta - \theta_o\right\|^{\rho_o}.$$

Let $\check{\theta}_n$ denote a maximizer of the function $\mathbb{V}_n(\theta) = \mathbb{P}_n g_\theta$. Assumption B.1 plays a pivotal role in establishing the convergence rate of $\check{\theta}_n$ as well as establishing the oracle regret bound, i.e., the convergence rate of $\mathbb{V}(\theta_o) - \mathbb{V}(\check{\theta}_n)$. The parameter ρ_o is commonly referred to as the margin parameter in the statistical learning literature (see Tsybakov (2004); Schölkopf and Smola (2002)). From this perspective, Assumption B.1 serves as an analogue to the restricted eigenvalue condition in the Lasso framework. Let $\boldsymbol{X}$ be the design matrix, and let $\widehat{\beta}$ denote the Lasso estimator for β. The margin assumption helps to establish the relationship between

the prediction error $\left\|\boldsymbol{X}'(\widehat{\beta}-\beta)\right\|$ and the estimation error $\|\widehat{\beta}-\beta\|$. In the context of Lasso estimation, ρ_o is set to be one, whereas $\rho_o = 2$ in classical M-estimation theory (see van der Vaart and Wellner (1998); Kosorok (2008)).

In the following example, we verify Assumption B.1 for the linear rules introduced in Example 1.

Example 4 (Linear Rules)**.** We consider th policy class $\Pi = \{\pi_\beta = \mathbb{1}\{x'\beta > 0\} : \|\beta\| = 1\}$.[9] To verify Assumption B.1, we apply the primitive conditions stated in Assumption B.2. With a slight abuse of notation, we write $\theta = (\beta, \eta)$ and $g_\theta = g_{(\pi_\beta,\eta)}$, and let $\theta_o = (\beta_o, \eta_o)$ denote the maximizer of the function $(\beta, \eta) \mapsto \mathbb{V}(\pi_\beta, \eta)$.

Assumption B.2 (Curvature Assumption for Linear Rules)**.** (1) The function $\theta \mapsto \mathbb{V}(\theta)$ is twice continuously differentiable in a neighborhood of θ_o with a negative definite Hessian matrix $\nabla^2_\theta \mathbb{V}(\theta)$ evaluated at $\theta = \theta_o$.

(2) *Margin Assumption*: There are $t_* > 0$ and $\rho \geq 1$ such that $P(0 < |X_i'\beta_o| \leq t) \lesssim t^\rho$ for all $t \in (0, t_*)$.

(3) The support $\mathcal{X}$ of X_i is bounded.

Assumption B.2 (1) is a standard assumption in parametric M-estimation (see van der Vaart and Wellner (1998); van der Vaart (2000); Kosorok (2008); Kim and Pollard (1990); Shi et al. (2018)). In contrast, Assumption B.2 (2) is widely used in statistical and policy learning, as noted by Kitagawa and Tetenov (2018); Luedtke and Chambaz (2020); Tsybakov (2004); Zhao and Cui (2023). It is straightforward to see that Assumption B.2 (2) and (3) together imply $\|\pi_\beta - \pi_{\beta_o}\|_{L^2(P)} \lesssim \|\beta - \beta_o\|^{\rho/2}$. As a result, Assumption B.2 provides the necessary conditions to verify Assumption B.1 for linear policies, which can be established via a Taylor expansion:

$$\begin{aligned}\mathbb{V}(\theta) - \mathbb{V}(\theta_o) &< -c_o\left(\|\beta - \beta_o\|^2 + |\eta - \eta_o|^2\right) \\ &\leq -c_o\left(\|\pi_\beta - \pi_{\beta_o}\|^\rho_{L^2(P)} + |\eta - \eta_o|^2\right),\end{aligned}$$

where $c_o > 0$ is a constant does not depends on $\theta = (\beta, \eta)$.

B.2 Faster Rate

In this subsection, we derive a sharper oracle regret bound than the one presented in Theorem 4.1. For illustrative purposes, this subsection focuses on the oracle regret bound based on the true influence scores[10] $g_\theta(\cdot)$. Fundamentally, the convergence rate of the regret $\mathbb{V}(\check{\theta}_n) - \mathbb{V}(\theta_o)$ is largely determined by the modulus of continuity of the empirical process

[9] The restriction $\|\beta\| = 1$ ensures that if $\beta_1 \neq \beta_2$ with $\|\beta_1\| = \|\beta_2\| = 1$ then $\pi_{\beta_1} \neq \pi_{\beta_2}$.

[10] GX: Based on Eq. (D.2), the regret bound can be upper-bounded by the sum of the oracle regret bound combined with the nuisance parameter estimation error or the uniform coupling error, as established in Lemma 4.1.

$\sqrt{n}(\mathbb{P}_n - P)g_\theta$, indexed by θ. This can be effectively controlled using maximal inequalities under uniform entropy conditions, see van der Vaart and Wellner (1998, 2011); Chernozhukov et al. (2014).

To establish the improved oracle regret bound rate under Assumption 5.2, we introduce the following technical assumption. This helps circumvent measurability issues and enables the use of Talagrand's inequality to control the local empirical process effectively.

Assumption B.3. There is a countable subset Θ' of Θ satisfying that for any $\theta \in \Theta$, there is a sequence $(\theta_k)_{k=1}^{\infty}$ in Θ' such that $\lim_{k\to\infty} g_{\theta_k}(z) = g_\theta(z)$ for P-a.s. $z \in \mathcal{Z}$.

Theorem B.1. Suppose that Assumption 4.2, Assumption 5.1 (1), Assumption 5.2, Assumption B.1, and Assumption B.3 hold. If $\tau(x,\eta)$ is uniformly bounded, i.e., $\sup_{x,\eta} |\tau(x,\eta)| < \infty$, then there is a universal constant $c_o > 0$ not depending on n such that

$$\mathbb{E}_P\left[\mathrm{Reg}(\check{\theta}_n)\right] \leq c_o \left(\mathrm{VC}(\Pi)/n\right)^{\rho_o/(2\rho_o-1)}, \quad \forall n \in \mathbb{N}^+.$$

Remark B.1. Let us analyze the role of the margin parameter ρ_o. If we remove the assumption on the margin parameter (i.e., letting $\rho_o \to \infty$), the regret convergence rate becomes $O(\sqrt{\mathrm{VC}(\Pi)/n})$, identical to the rate in Theorem 4.1, and independent of ρ_o. Notably, the knowledge of the margin parameter ρ_o is not required, as it neither needs to be estimated nor plays a role in constructing the optimal policy.

C Uniform Inference for the Optimal Welfare

In this section, we develop inference for the optimal welfare without Assumption 5.2. It improves upon the inference proposed in Appendix B in the supplemental material to Kitagawa and Tetenov (2018). Throughout this section, we assume that $\mathrm{VC}(\Pi)$ is finite, i.e., Assumption 5.1 is satisfied.

To develop uniform inference, we define a distance d_Π to measure the dissimilarity between policies in Π, independent of the the underlying distribution P. To do so, let $\nu = \nu_1 \times \cdots \times \nu_p$ on $\mathbb{R}^p$ be a product finite measure. The distance d_Π is defined as

$$d_\Pi(\pi, \tilde{\pi}) = \int_{\mathbb{R}^p} |\pi(x) - \tilde{\pi}(x)| d\nu(x), \quad \forall \pi, \tilde{\pi} \in \Pi.$$

A typical choice for ν is the Lebesgue measure on $\mathbb{R}^p$. Moreover, we introduce a pseudometric d_Θ on Θ, defined by $d_\Theta(\theta, \tilde{\theta}) = d_\Pi(\pi, \tilde{\pi}) + |\eta - \tilde{\eta}|$ for all $\theta = (\pi, \eta)$ and $\tilde{\theta} = (\tilde{\pi}, \tilde{\eta})$. Furthermore, the estimated functions $\widehat{e}(\cdot)$ and $\widehat{\mu}_a(\cdot,\cdot)$ need to satisfy Assumption 4.2 uniformly across a collection of distributions $P \in \mathcal{P}_n$. This, in turn, requires the nonparametric/ML models used to estimate $e_o(\cdot)$ and $\mu_a(\cdot,\cdot)$ to be not excessively complex. To formalize this condition, let Δ_n, ψ_n, and $\tau_n \searrow 0$ be sequences that approach zero from above at a rate no faster than polynomial in n (e.g. $\Delta_n > n^{-c}$ for some $c > 0$). Let $\mathcal{M}_{n,a}$ and $\mathcal{D}_n$ denote the classes of

measurable functions $\check{\mu}_a, \check{e}$ such that $\|\check{\mu}_a - \mu_a\|_{P,2} \leq \tau_n/2$ and $\|\check{e} - e_o\|_{P,2} \leq \tau_n/2$. Finally, let

$$\mathcal{F}_n = \{g_\theta(\cdot; \check{\mu}, \check{e}) : \theta \in \Theta, \check{\mu}_a \in \mathcal{M}_{n,a}, \check{e} \in \mathcal{D}_n\},$$

where $g_\theta(z; \check{\mu}, \check{e})$ is defined in Eq. (3.2). We impose the following regularity conditions.

Assumption C.1. There exists $n_o \in \mathbb{N}^+$ and a constant $c_o > 0$ such that the following conditions hold for all $n \geq n_0$ and $P \in \mathcal{P}_n$.

(1) $|Y_i| \leq c_o$ P-a.s. and Assumption 2.1 holds.

(2) $X \in \mathbb{R}^p$ has density $f_P : \mathcal{X} \to \mathbb{R}_+$ such that $\|f_P\|_\infty \leq c_o$, with respect to ν.

(3) Suppose $\tau_n^2\sqrt{n} \leq \delta_n$, and the estimated functions $\widehat{\mu}_a(\cdot,\cdot) \in \mathcal{M}_{n,a}$ and $\widehat{e}(\cdot) \in \mathcal{D}_n$, with probability at least $1 - \Delta_n$. Let $a_n \geq n \vee e$ and $s_n \geq 1$ be two sequences such that

$$n^{-1/2}\left(\sqrt{s_n \log a_n} + n^{-1/4} s_n \log a_n\right) \leq \tau_n \quad \text{and}$$
$$\tau_n^{1/2}\sqrt{s_n \log a_n} + s_n n^{-1/4} \log a_n \cdot \log n \leq \psi_n.$$

The function class $\mathcal{F}_n$ is suitably measurable and its uniform covering entropy satisfies:

$$\sup_Q \log N\left(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_n, \|\cdot\|_{Q,2}\right) \leq s_n \log\left(a_n/\epsilon\right) \vee 0,$$

where F_1 is an envelope for $\mathcal{F}_n$ with $\|F_1\|_\infty \leq C$ for all n.

Define the supremum functional $\psi : \ell^\infty(\Theta) \to \mathbb{R}$ as $\psi : h \mapsto \sup_{\theta\in\Theta} h(\theta)$. We can verify that ψ is Hadamard directionally differentiable tangentially to $C_b(\Theta)$, which allows the application of generalized delta method, see Belloni et al. (2017); Fang and Santos (2019); Hong and Li (2018). Let $\Pi_P^\star := \arg\max_{\theta\in\Theta} \mathbb{V}_P(\theta)$. It is known that the directional derivative of ψ at $\mathbb{V}_P$ is $\psi'_P : C_b(\Theta) \to \mathbb{R}$ as $\psi'_P(h) = \sup_{\theta\in\Pi_P^\star} h(\theta)$.

To construct uniform inference, we follow the approach in Belloni et al. (2017); Fang and Santos (2019); Hong and Li (2018). It involves three steps. In the first step, we establish uniform weak convergence of the empirical process $\sqrt{n}(\widehat{\mathbb{V}}_n - \mathbb{V})$ to a Gaussian process in Lemma F.1 in Appendix F; in the second step, we apply the delta method to the supremum functional, validated by Lemma F.3: $\sqrt{n}\left[\sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta) - \sup_{\theta\in\Theta}\mathbb{V}(\theta)\right]$ to derive its limiting distribution in Theorem C.1; finally we estimate the limiting distribution by the numerical delta method introduced by Hong and Li (2018), see Lemma F.3 and Lemma F.4 in Appendix F.

Theorem C.1. Suppose $\mathrm{VC}(\Pi) < \infty$ and Assumption C.1 hold. Then

$$\sqrt{n}\left[\psi\left(\widehat{\mathbb{V}}_n\right) - \psi\left(\mathbb{V}_P\right)\right] \rightsquigarrow \psi'_P(\mathbb{G}_P) = \sup_{\theta\in\Pi_P^\star} \mathbb{G}_P(\theta),$$

where $\mathbb{G}_P : \theta \mapsto \mathbb{G}_P g_\theta$ is a mean zero tight Gaussian process on $\ell^\infty(\Theta)$ with covariance

function

$$\mathrm{Cov}_P(\theta_1, \theta_2) = \mathbb{E}\left[\mathbb{G}_P(\theta_1)\mathbb{G}_P(\theta_2)\right].$$

Moreover, the paths $\theta \mapsto \mathbb{G}_P(\theta)$ are a.s. uniformly continuous on (Θ, d_Θ), satisfying the following conditions:

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}_P\left[\sup_{\theta\in\Theta}|\mathbb{G}_P|\right] < \infty \quad \text{and} \quad \lim_{\delta\searrow 0}\sup_{P\in\mathcal{P}_n}\mathbb{E}_P\left[\sup_{d_\Theta(\theta,\bar{\theta})\le\delta}\left|\mathbb{G}_P(\theta) - \mathbb{G}_P(\bar{\theta})\right|\right] = 0.$$

When the maximizer of $\mathbb{V}_P$ is unique, i.e., $\Pi_P^\star = \{\theta_o\}$ is a singleton, Theorem C.1 implies that $\sqrt{n}\left[\psi\left(\widehat{\mathbb{V}}_n\right) - \psi\left(\mathbb{V}_P\right)\right]$ weakly converges to the normal distribution defined in Theorem 5.1. When $\Pi_P^\star$ is not a singleton, $\sqrt{n}\left[\psi\left(\widehat{\mathbb{V}}_n\right) - \psi\left(\mathbb{V}_P\right)\right]$ no longer converges weakly to normal distribution. Although Theorem C.1 establishes the asymptotic distribution of the estimator for the optimal welfare, conducting valid inference still requires information on the distribution of $\mathbb{G}_P$ and the directional derivative ψ'_P. We utilize the bootstrap approach to approximate the distribution of $\mathbb{G}_P$. In particular, we consider the multiplier bootstrap $\widehat{\mathbb{G}}_n^* : \Theta \to \mathbb{R}$ defined as

$$\widehat{\mathbb{G}}_n^* : \theta \mapsto n^{-1/2}\sum_{i=1}^n \xi_i\left[\widehat{g}_\theta(Z_i) - \widehat{\mathbb{V}}_n(\theta)\right],$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. random variables independent of $(Z_i)_{i=1}^n$, with $\mathbb{E}(\xi_i) = 0$, $\mathbb{E}(\xi_i^2) = 1$ and $\mathbb{E}\left[\exp|\xi_i|\right] < \infty$. We apply the numerical delta method proposed by Hong and Li (2018) to estimate the directional derivative $\psi'_P(\mathbb{G}_P)$.[11] This is justified by Theorem 3.1 in Hong and Li (2018) or Lemma F.3 in Appendix F. For given $\epsilon_n = o(1)$ with $n^{1/2}\epsilon_n \to \infty$, we estimate the $\psi'_P(\mathbb{G}_P)$ using the distribution of the random variable:

$$\widehat{\psi}'_n(\widehat{\mathbb{G}}_n^*) = \frac{\psi\left(\widehat{\mathbb{V}}_n + \epsilon_n\widehat{\mathbb{G}}_n^*\right) - \psi(\widehat{\mathbb{V}}_n)}{\epsilon_n}. \tag{C.1}$$

[11] Other methods than the numerical delta method introduced by Hong and Li (2018) can be used to estimate $\psi'_P(\mathbb{G}_P) = \sup_{\theta\in\Pi_P^\star}\mathbb{G}_P(\theta)$ as well; see, for example, Firpo et al. (2023).

D Proofs for results in the main text

D.1 Proof of Lemma 3.1

Proof. For $0 < \alpha < 1$, the results can be found in Theorem 6.2 of Shapiro et al. (2021). For $\alpha = 1$, we first note that $(Y_i(\pi) - \eta)_- + (Y_i(\pi) - \eta)_+ = Y_i(\pi)$. Therefore, we have

$$\begin{aligned}
\sup_{\eta \in \mathbb{R}} \mathbb{V}_1(\pi, \eta) &= \sup_{\eta \in \mathbb{R}} \left\{ \mathbb{E}\left[(Y_i(\pi) - \eta)_- \right] + \eta \right\} \\
&= \sup_{\eta \in \mathbb{R}} \left\{ \mathbb{E}\left[(Y_i(\pi) - \eta) - (Y_i(\pi) - \eta)_+ \right] + \eta \right\} \\
&= \mathbb{E}\left[Y_i(\pi) \right] - \inf_{\eta \in \mathbb{R}} \mathbb{E}\left[(Y_i(\pi) - \eta)_+ \right].
\end{aligned}$$

We note that $\eta \mapsto (Y_i(\pi) - \eta)_+$ is decreasing and converges to zero almost surely as $\eta \to \infty$. Moreover, we have $0 \leq (Y_i(\pi) - \eta)_+ \leq |Y_i(\pi)| + |\eta|$, applying the dominated convergence theorem yields:

$$\inf_{\eta \in \mathbb{R}} \mathbb{E}\left[(Y_i(\pi) - \eta)_+ \right] = \lim_{\eta \to \infty} \mathbb{E}\left[(Y_i(\pi) - \eta)_+ \right] = 0.$$

This shows that $\sup_{\eta \in \mathbb{R}} \mathbb{V}_1(\pi, \eta) = \mathbb{E}\left[Y_i(\pi) \right] = \lim_{\eta \to \infty} \mathbb{V}_1(\pi, \eta)$. □

D.2 Proof of Theorem 3.1

Proof. First, it is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\pi(X_i) \left(Y_i(1) - \eta \right)_- | X_i \right] &= \pi(X_i) \mathbb{E}\left[(Y_i(1) - \eta)_- | X_i \right] \\
&= \pi(X_i) \mu_1(X_i, \eta),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[(1 - \pi(X_i)) \left(Y_i(0) - \eta \right)_- | X_i \right] &= (1 - \pi(X_i)) \, \mathbb{E}\left[(Y_i(0) - \eta)_- | X_i \right] \\
&= \pi(X_i) \mu_0(X_i, \eta).
\end{aligned}$$

Applying the law of iterated expectations gives

$$\begin{aligned}
\mathbb{E}\left[(Y_i(\pi) - \eta)_- \right] &= \mathbb{E}\left[(1 - \pi(X_i)) \left(Y_i(0) - \eta \right)_- \right] + \mathbb{E}\left[\pi(X_i) \left(Y_i(1) - \eta \right)_- \right] \\
&= \mathbb{E}\left\{ \mathbb{E}\left[(1 - \pi(X_i)) \left(Y_i(0) - \eta \right)_- \mid X_i \right] \right\} \\
&+ \mathbb{E}\left\{ \mathbb{E}\left[\pi(X_i) \left(Y_i(1) - \eta \right)_- \mid X_i \right] \right\} \\
&= \mathbb{E}\left[\pi\left(X_i \right) \mu_1(X_i, \eta) \right] + \mathbb{E}\left[\left(1 - \pi\left(X_i \right) \right) \mu_0(X_i, \eta) \right].
\end{aligned}$$

This ends the proof of the first part of Eq. (3.1). Next, we consider the following derivation:

$$\begin{aligned}\mathbb{E}\left[A_i\left(Y_i(A_i)-\eta\right)_- \mid X_i\right] &= \mathbb{E}\left[A_i\left(Y_i(A_i)-\eta\right)_- \mid X_i, A_i=1\right]\mathbb{P}(A_i=1\mid X_i)\\ &=_{(1)} \mathbb{E}\left[\left(Y_i(1)-\eta\right)_- \mid X_i, A_i=1\right]e_o(X_i)\\ &= \mathbb{E}\left[\left(Y_i(1)-\eta\right)_- \mid X_i\right]e_o(X_i)\\ &= \mu_1(X_i,\eta)e_o(X_i),\end{aligned}$$

where Equation (1) follows from conditional independence. Therefore,

$$\mathbb{E}\left[\frac{A_i\pi(X_i)}{e_o(X_i)}(Y_i-\eta)_- \mid X_i\right] = \frac{\pi(X_i)}{e_o(X_i)}\mathbb{E}\left[A_i(Y_i-\eta)_- \mid X_i\right] = \pi(X_i)\mu_1(X_i,\eta).$$

Using the similar argument displayed above, one has

$$\mathbb{E}\left[\left(1-A_i\right)\left(Y_i(A_i)-\eta\right)_- \mid X_i\right] = \mu_0(X_i,\eta)\left[1-e_o(X_i)\right],$$

and hence

$$\mathbb{E}\left[\frac{(1-A_i)(1-\pi(X_i))}{(1-e_o(X_i))}(Y_i-\eta)_- \mid X_i\right] = \left(1-\pi(X_i)\right)\mu_0(X_i,\eta).$$

As a result, we have

$$\mathbb{E}\left[w\left(Z_i,\pi\right)\left(Y_i-\eta\right)_-\right] = \mathbb{E}\left[\pi(X_i)\mu_1(X_i,\eta)\right] + \mathbb{E}\left[\left(1-\pi(X_i)\right)\mu_0(X_i,\eta)\right],$$

and the desired result follows. □

D.3 Proof of Lemma 4.1

Proof. Let $g(x,a) = \frac{a-e_o(x)}{e_o(x)(1-e_o(x))}$, and define

$$\begin{aligned}\phi_i(\eta) &= \frac{1}{\alpha}\tau(X_i,\eta) + g(X_i,A_i)\left[(Y_i-\eta)_- - \mu_{A_i}(X_i,\eta)\right],\\ \widehat{\phi}_i(\eta) &= \frac{1}{\alpha}\widehat{\tau}\left(X_i,\eta\right) + \widehat{g}\left(X_i,A_i\right)\left[(Y_i-\eta)_- - \widehat{\mu}_{A_i}(X_i,\eta)\right],\\ \psi_i(\eta) &= \mu_0(X_i,\eta) + \frac{1-A_i}{\alpha\left(1-e\left(X_i\right)\right)}\left[(Y_i-\eta)_- - \mu_0\left(X_i,\eta\right)\right],\\ \widehat{\psi}_i(\eta) &= \widehat{\mu}_0(X_i,\eta) + \frac{1-A_i}{\alpha\left(1-\widehat{e}\left(X_i\right)\right)}\left[(Y_i-\eta)_- - \widehat{\mu}_0\left(X_i,\eta\right)\right].\end{aligned}$$

By the definition of $g(x,a)$, and we estimate it by $\widehat{g}(x,a) = \frac{a-\widehat{e}(x)}{\widehat{e}(x)(1-\widehat{e}(x))}$. Under Assumption 2.1 (2) and Assumption 4.2, we have

$$\sup_{x,a} |\widehat{g}(x,a) - g(x,a)| = o_P(1),$$
$$\left[\mathbb{E}\left|\widehat{g}(X_i,A_i) - g(X_i,A_i)\right|^2\right]^{1/2} = O(n^{-\zeta_e}).$$

We divide $\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}_n(\theta)$ into two parts:

$$\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \pi(X_i)\left[\widehat{\phi}_i(\eta) - \phi_i(\eta)\right] + \frac{1}{n}\sum_{i=1}^n \left[\widehat{\psi}_i(\eta) - \psi_i(\eta)\right].$$

The proof is divided into following two steps for bounding the two terms displayed above.

Step 1. We first bound the first summand by considering the following decomposition:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \pi(X_i)\left[\widehat{\phi}_i(\eta) - \phi_i(\eta)\right] \\
&= \frac{1}{n}\sum_{i=1}^n \pi(X_i)\left[(Y_i-\eta)_- - \mu_{A_i}(X_i,\eta)\right]\left[\widehat{g}^{(-k(i))}(X_i) - g(X_i)\right] \\
&+ \frac{1}{n}\sum_{i=1}^n \pi(X_i)\underbrace{\left[\widehat{\tau}^{(-k(i))}(X_i,\eta) - \tau(X_i,\eta) - g(X_i)\left(\widehat{\mu}_{A_i}^{(-k(i))}(X_i,\eta) - \mu_{A_i}(X_i,\eta)\right)\right]}_{=\widehat{\phi}_\eta^{(-k(i))}(Z_i)} \\
&- \frac{1}{n}\sum_{i=1}^n \pi(X_i)\left[\widehat{\mu}_{A_i}^{(-k(i))}(X_i,\eta) - \mu_{A_i}(X_i,\eta)\right]\left[\widehat{g}^{(-k(i))}(X_i) - g(X_i)\right].
\end{aligned}$$

Denote these three summands by $\mathit{\Pi}_1(\theta)$, $\mathit{\Pi}_2(\theta)$, and $\mathit{\Pi}_3(\theta)$. We will bound all three summands separately.

To bound the first term, it suffices to consider the contribution of each folder. For any folder $k \in [K]$, let

$$\mathit{\Pi}_1^{(k)}(\theta) = \frac{1}{n}\sum_{i\in\mathcal{I}_k} \pi(X_i)\left[(Y_i-\eta)_- - \mu_{A_i}(X_i,\eta)\right]\left[\widehat{g}^{(-k(i))}(X_i) - g(X_i)\right].$$

By Assumption 4.2, we have

$$\sup_{x\in\mathcal{X}}\left|\widehat{g}^{(-k(i))}(x) - g(x)\right| \le 1,$$

with probability tending to one. Moreover, $\mathbb{E}\left[(Y_i-\eta)_- - \mu_{A_i}(X_i,\eta) \mid X_i, A_i, \widehat{g}^{(-k(i))}\right] = 0$.

By Lemma G.1 and applying Theorem 2.14.1 in van der Vaart and Wellner (1998) gives that there is a universal constant $c_o > 0$ such that the following inequalities hold for all n

large enough:

$$\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}|\mathit{\Pi}_1^{(k)}(\theta)|\;\middle|\;\widehat{g}^{(-k)}\right]\leq c_o\sqrt{\mathrm{VC}(\Pi_n)/n}\sqrt{\mathbb{E}\left[\left|\widehat{g}^{(-k)}(Z)-g_o(Z)\right|^2|\widehat{g}^{(-k)}(\cdot)\right]}.$$

Therefore, given $\mathrm{VC}(\Pi_n)=o(n^{2\zeta_e})$ and by Jensen's inequality, taking expectation on both hand-sides gives

$$\mathbb{E}\left[\sup_{\theta\in\Theta_n}|\mathit{\Pi}_1^{(k)}(\theta)|\right]\leq\frac{c_o\sqrt{\mathrm{VC}(\Pi_n)/n}}{n^{\zeta_e}}=o(n^{-1/2}).$$

Next, we bound the second term $\mathit{\Pi}_2(\theta)$. By Assumption 2.1 and cross-fitting, one has

$$\mathbb{E}\left[\widehat{\phi}_\eta^{(-k(i))}(Z_i)\middle|X_i,\widehat{\tau}^{(-k(i))}(\cdot),\widehat{\mu}_{A_i}^{(-k(i))}(\cdot)\right]=0,$$

for all $\eta\in\mathcal{B}_Y$. Given $\widehat{g}^{(-k)}$, $\widehat{\mu}_a^{(-k)}$, the class of function $\mathcal{H}^{(-k)}\equiv\{z\mapsto\widehat{\phi}_\eta^{(-k)}(z):\eta\in\mathcal{B}_Y\}$ is Lipschitz in η, i.e., there is some constant $c_o>0$ such that

$$\left|\widehat{\phi}_{\eta_1}^{(-k)}(z)-\widehat{\phi}_{\eta_2}^{(-k)}(z)\right|\leq c_o|\eta_1-\eta_2|,$$

for all η_1,η_2. By Theorem 2.7.11 in van der Vaart and Wellner (1998), there is a universal $K>0$ such that

$$N(\epsilon,\mathcal{H}^{(-k)},L^2(Q))\leq N_{[]}(\epsilon,\mathcal{H}^{(-k)},L^2(Q))\leq N\left(\epsilon/c_o,\mathcal{B}_Y,\|\cdot\|\right),$$

for all finitely discrete distribution Q on $\mathcal{Z}$. Let $\mathcal{F}_n^{(-k)}=\Pi_n\otimes\mathcal{H}^{(-k)}$, and $\mathcal{F}_n^{(-k)}$ has an envelope function

$$\widehat{F}^{(-k)}(z)=\sup_{\eta\in\mathcal{B}_Y}\left|\widehat{\phi}_\eta^{(-k)}(z)\right|,$$

where $\|\widehat{F}^{(-k)}\|_\infty=o_P(1)$ by Assumption 4.2. Since $\sup_{\eta\in\mathcal{B}_Y}\|\widehat{\phi}_\eta^{(-k)}\|_\infty=o_P(1)$, then for any $\pi,\pi_1\in\Pi_n$ with $\|\pi-\pi_1\|_{P,2}\leq\epsilon/2$ and $\eta,\eta_1\in\mathcal{B}_Y$ such that $\|\widehat{\phi}_\eta^{(-k)}-\widehat{\phi}_{\eta_1}^{(-k)}\|_\infty\leq(\epsilon/2)\|\widehat{F}^{(-k)}\|_\infty$, one has

$$\begin{aligned}\left\|\pi\widehat{\phi}_\eta^{(-k)}-\pi_1\widehat{\phi}_{\eta_1}^{(-k)}\right\|_{P,2}&\leq\|\pi\|_{P,2}\left\|\widehat{\phi}_\eta^{(-k)}-\widehat{\phi}_{\eta_1}^{(-k)}\right\|_{P,2}+\|\pi-\pi_1\|_{P,2}\left\|\widehat{\phi}_{\eta_1}^{(-k)}\right\|_{P,2}\\&\leq\epsilon,\end{aligned}$$

with probability tending to one. Therefore, the following inequality holds with probability tending to one:

$$\begin{aligned}\log N\left(\epsilon,\mathcal{F}^{(-k)},L^2(Q)\right)\leq&\log N\left(\epsilon/2,\Pi_n,L^2(Q)\right)+\log N\left(\epsilon/2,\mathcal{H}^{(-k)},L^2(Q)\right)\\\leq&\,\mathrm{VC}(\Pi_n)\log(2/\epsilon)+\log\left(2c_o/\epsilon\right),\end{aligned}$$

for all finitely discrete distribution Q. Applying maximal inequality in van der Vaart and

Wellner (2011) or Chernozhukov et al. (2014) gives

$$\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}|\mathit{\Pi}_2^{(k)}(\theta)|\middle|\widehat{\tau}^{(-k}(\cdot),\widehat{\mu}_a^{(-k}(\cdot)\right]=O\left(\sqrt{\frac{\mathrm{VC}(\Pi_n)}{n_k}}\right)\sqrt{\mathbb{E}\left[\left|\widehat{F}^{(-k)}\right|^2\middle|\widehat{\tau}^{(-k}(\cdot),\widehat{\mu}_a^{(-k}(\cdot)\right]}.$$

Taking expectation on both hand sides yields and applying Jensen's inequality, we have

$$\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}|\mathit{\Pi}_2^{(k)}(\theta)\right]=o(n^{-1/2}).$$

Using the similar argument, we can establish an upper bound for $\mathit{\Pi}_3(\theta)$ as follows:

$$\mathbb{E}\left[\sup_{\theta\in\Theta_n}|\mathit{\Pi}_3(\theta)|\right]=o(n^{-1/2}).$$

Step 2. We bound the second term $n^{-1}\sum_{i=1}^n[\widehat{\psi}_i(\eta)-\psi_i(\eta)]$. Consider the following decomposition:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\left[\widehat{\psi}_i(\eta)-\psi_i(\eta)\right]&=\frac{1}{n}\sum_{i=1}^n\left[\widehat{\mu}_0^{(-k(i))}\left(X_i,\eta\right)-\mu_0\left(X_i,\eta\right)\right]\left[1-\frac{1-A_i}{1-e(X_i)}\right]\\&+\frac{1}{n}\sum_{i=1}^n(1-A_i)\left[(Y_i-\eta)_--\mu_0\left(X_i,\eta\right)\right]\left[\frac{1}{1-\widehat{e}^{(-k(i))}\left(X_i\right)}-\frac{1}{1-e\left(X_i\right)}\right]\\&-\frac{1}{n}\sum_{i=1}^n(1-A_i)\left[\widehat{\mu}_0^{(-k(i))}\left(X_i,\eta\right)-\mu_0\left(X_i,\eta\right)\right]\left[\frac{1}{1-\widehat{e}^{(-k(i))}\left(X_i\right)}-\frac{1}{1-e\left(X_i\right)}\right].\end{aligned}$$

Denote these three summands by $I_1^{(k)}(\eta)$, $I_2^{(k)}(\eta)$ and $I_3^{(k)}(\eta)$, and we can bound three summands using the similar argument in step 1 as follows,

$$\begin{aligned}&\sqrt{n}\mathbb{E}\left[\sup_{\eta\in\mathcal{B}_Y}\left|I_1^{(k)}(\eta)\right|\right]=O\left(n^{-\zeta_\mu}\right),\quad\sqrt{n}\mathbb{E}\left[\sup_{\eta\in\mathcal{B}_Y}\left|I_2^{(k)}(\eta)\right|\right]=O\left(n^{-\zeta_e}\right),\\&\sqrt{n}\mathbb{E}\left[\sup_{\eta\in\mathcal{B}_Y}\left|I_3^{(k)}(\eta)\right|\right]=O\left(n^{-\zeta_e-\zeta_\mu}\right).\end{aligned}$$

Therefore, combination of step 1 and step 2 shows

$$\sqrt{n}\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}\left|\widehat{\mathbb{V}}_n(\theta)-\mathbb{V}_n(\theta)\right|\right]=O(n^{-a_o}),$$

where $a_o=\zeta_\mu\wedge\zeta_e-b_o/2>0$. □

D.4 Proof of Lemma 4.2

Proof. By the definitions of $\mathbb{W}_\alpha(\pi)$ and $\mathbb{V}(\pi, \eta)$, the regret of π relative to the policy class Π_n can be written as

$$\mathrm{Reg}(\pi, \Pi_n) = \max_{\pi' \in \Pi_n} \left[\sup_{\eta \in \mathcal{B}_Y} \mathbb{V}(\pi', \eta) \right] - \sup_{\eta \in \mathcal{B}_Y} \mathbb{V}(\pi, \eta).$$

Noting that for $\widehat{\theta}_n \equiv (\widehat{\pi}_n, \widehat{\eta}_n)$, we obtain

$$\mathrm{Reg}(\widehat{\theta}_n) = \sup_{\pi' \in \Pi_n} \mathbb{W}_\alpha(\pi') - \mathbb{W}_\alpha(\widehat{\pi}_n) = \sup_{\theta' \in \Theta_n} \mathbb{V}\left(\theta'\right) - \mathbb{V}(\widehat{\theta}_n). \tag{D.1}$$

We consider the following expression:

$$\mathbb{V}(\theta) - \mathbb{V}(\widehat{\theta}_n) = \mathbb{V}(\theta) - \mathbb{V}_n(\widehat{\theta}_n) + \mathbb{V}_n(\widehat{\theta}_n) - \mathbb{V}(\widehat{\theta}_n).$$

Let $\check{\theta}_n = \mathrm{argmax}_{\theta \in \Theta_n} \mathbb{V}_n(\theta)$. By the definitions of $\check{\theta}_n$ and $\widehat{\theta}_n$, it follows that:

$$\begin{aligned}
\mathbb{V}(\theta) - \mathbb{V}_n(\widehat{\theta}_n) &\leq \mathbb{V}(\theta) - \mathbb{V}_n(\theta) + \underbrace{\mathbb{V}_n(\theta) - \mathbb{V}_n(\check{\theta}_n)}_{\leq 0} + \underbrace{\mathbb{V}_n(\check{\theta}_n) - \widehat{\mathbb{V}}_n(\check{\theta}_n)}_{= o_P(n^{-1/2})} \\
&\quad + \underbrace{\widehat{\mathbb{V}}_n(\check{\theta}_n) - \widehat{\mathbb{V}}_n(\widehat{\theta}_n)}_{\leq 0} + \underbrace{\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}_n(\widehat{\theta}_n)}_{o_P(n^{-1/2})} \\
&\leq \mathbb{V}(\theta) - \mathbb{V}_n(\theta) + r_n,
\end{aligned}$$

where $r_n = o_P(n^{-1/2})$ and $\sqrt{n}\mathbb{E}|r_n| \to 0$ by Lemma 4.1. Thus, for all $\theta \in \Theta_n$:

$$\begin{aligned}
0 \leq \mathbb{V}(\theta) - \mathbb{V}(\widehat{\theta}_n) &\leq \mathbb{V}_n(\widehat{\theta}_n) - \mathbb{V}(\widehat{\theta}_n) + \mathbb{V}(\theta) - \mathbb{V}_n(\theta) + r_n \\
&\leq 2 \sup_{\theta \in \Theta_n} |\mathbb{V}_n(\theta) - \mathbb{V}(\theta)| + r_n \\
&= 2 \sup_{\theta \in \Theta_n} |(\mathbb{P}_n - P)\, g_\theta| + r_n.
\end{aligned} \tag{D.2}$$

Without loss of generality, suppose that there exists $\theta_n^* \in \Theta_n$ such that $\mathbb{V}(\theta_n^*) = \max_{\theta \in \Theta_n} \mathbb{V}(\theta)$. If no such θ_n^* exists, the proof can be adapted using an ε-approximate optimizer, where $\varepsilon \to 0$. Substituting θ_n^* into the preceding expression yields

$$0 \leq \mathbb{V}(\theta_n^*) - \mathbb{V}(\widehat{\theta}_n) \leq 2 \sup_{\theta \in \Theta_n} |(\mathbb{P}_n - P) g_\theta| + r_n.$$

□

D.5 Proof of Theorem 4.1

Inspired by Lemma 2 in Athey and Wager (2021), this proof follows the classical chaining argument while incorporating a novel, conditionally-defined semi-distance.

New Conditional Semi-distance Recall that $g(x,a) = \frac{a-e_o(x)}{e_o(x)(1-e_o(x))}$, and g_θ defined in Eq. (3.2) can be rewritten as

$$\begin{aligned} g_\theta(z) = \frac{1}{\alpha} \underbrace{\left[\mu_0(x,\eta) + \frac{(1-a)}{1-e_o(x)}\left\{(y-\eta)_- - \mu_0(x,\eta)\right\}\right]}_{\equiv \gamma_\eta^\dagger(z)} + \eta \\ + \frac{1}{\alpha}\pi(x) \underbrace{\left[\tau(x,\eta) + g(x,a)\left\{(y-\eta)_- - \mu_a(x,\eta)\right\}\right]}_{\equiv \gamma_\eta(z)}. \end{aligned} \tag{D.3}$$

Since $\eta \mapsto \sum_{i=1}^n |\gamma_\eta(Z_i)|^2$ is continuous almost surely and $\mathcal{B}_Y$ is compact, then there is a $\eta_n \in \mathcal{B}_Y$ at which the function $\sum_{i=1}^n |\gamma_\eta(Z_i)|^2$ attains its maximum. Given $(Z_i)_{i=1}^n$, define a conditional 2-norm distance between two policies π_1 and π_2 as

$$D_n^2(\pi_1,\pi_2) = \frac{\sum_{i=1}^n |\gamma_{\eta_n}(Z_i)|^2 (\pi_1(X_i) - \pi_2(X_i))^2}{\sum_{i=1}^n |\gamma_{\eta_n}(Z_i)|^2}. \tag{D.4}$$

Let $N_{D_n}(\epsilon, \Pi_n, (Z_i)_{i=1}^n)$ denote the ϵ-covering number under distance D_n. For simplicity, let $\Gamma_i = \gamma_{\eta_n}(Z_i)$. To bound N_{D_n} by the ϵ-Hamming entropy, we can construct a sample $(X_j')_{j=1}^m$ with X_j' contained in the support of $(X_i)_{i=1}^n$ such that for all $i \in [n]$:

$$\left| |\{j \in [m] : X_j' = X_i\}| - m\Gamma_i^2 / \sum_{j=1}^n \Gamma_j^2 \right| \le 1.$$

As a result, one has

$$\left| \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\pi_1(X_j') \neq \pi_2(X_j')\} - \frac{\sum_{i=1}^n \Gamma_i^2 (\pi_1(X_i) - \pi_2(X_i))^2}{\sum_{i=1}^n \Gamma_i^2} \right| \le \frac{n}{m}.$$

It is clear that, for any policies π_1 and π_2, one has

$$\left| \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{\pi_1(X_j') \neq \pi_2(X_j')\} - D_n^2(\pi_1,\pi_2) \right| \le \frac{n}{m}.$$

Moreover, recall that the Hamming covering number does not depend on sample size, so letting $m \to \infty$, one has $N_{D_n}(\epsilon, \Pi_n, (Z_i)_{i=1}^n) \le N_H(\epsilon^2, \Pi_n)$.

Proof of Theorem 4.1. Recall $\Theta_n = \Pi_n \times \mathcal{B}_Y$. First we construct a sequence of ϵ-nets for

Π_n with decreasing scale. Without loss of generality, we assume $\mathcal{B}_Y = [-\eta_B, \eta_B]$ for some constant $\eta_B > 0$. For any $j \in \mathbb{N}^+$, construct the set $\mathcal{B}^{(j)} \subseteq \mathcal{B}_Y$ as

$$\mathcal{B}^{(j)} \equiv \left\{-\eta_B + k \cdot 2^{-j} : 1 \le k \le \left\lfloor \eta_B 2^{j+1} \right\rfloor\right\}.$$

Moreover, for each $j \in \mathbb{N}^+$, we also construct sets $\Pi_n^{(j)} \subset \Pi_n$ such that for any $\pi \in \Pi_n$ there is a $\pi_n^{(j)} \in \Pi_n^{(j)}$ such that $D_n(\pi, \pi_n^{(j)}) \le 2^{-j}$. We write $\Theta_n^{(j)} = \Pi_n^{(j)} \times \mathcal{B}^{(j)}$, and define the operators $\Psi_j : \Theta_n \to \Theta_n^{(j)}$ as $\Psi_j(\theta) = (\Psi_{\Pi,j}(\pi), \Psi_{\mathcal{B}_Y,j}(\eta))$, where $\Psi_{\Pi,j}(\pi) = \arg\min_{\pi_0 \in \Pi_n^{(j)}} D_n(\pi_0, \pi)$ and $\Psi_{\mathcal{B}_Y,j}(\eta) = \arg\min_{\eta_0 \in \mathcal{B}^{(j)}} |\eta - \eta_0|$. Let $J_0 = 1$ $J(n) = (\log n)(3 - 2b_o)/8$ and $J_+(n) = (\log n)(1 - b_o)$, and we consider the following decomposition:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i, \theta) = &\frac{1}{n}\sum_{i=1}^n \xi_i g\left(X_i, \Psi_{J_0}(\theta)\right) \\
&+ \sum_{j=J_0+1}^{J(n)} \frac{1}{n}\sum_{i=1}^n \xi_i \left[g\left(X_i, \Psi_j(\theta)\right) - g\left(X_i, \Psi_{j-1}(\theta)\right)\right] \\
&+ \sum_{j=J(n)+1}^{J_+(n)} \frac{1}{n}\sum_{i=1}^n \xi_i \left[g\left(X_i, \Psi_j(\theta)\right) - g\left(X_i, \Psi_{j-1}(\theta)\right)\right] \\
&+ \frac{1}{n}\sum_{i=1}^n \xi_i \left[g\left(X_i, \theta\right) - g\left(X_i, \Psi_{J_+(n)}(\theta)\right)\right].
\end{aligned} \tag{D.5}$$

Recall the expression of g_θ given in Eq. (D.3), define

$$\widehat{S}_n = \sup_{\theta \in \Theta_n} \frac{1}{n}\sum_{i=1}^n |g_\theta(Z_i)|^2, \quad \widehat{\Xi}_n = \sup_{\eta \in \mathcal{B}_Y} \frac{1}{n}\sum_{i=1}^n |\gamma_\eta(Z_i)|^2, \quad \widehat{\Xi}_n^\dagger = \sup_{\eta \in \mathcal{B}_Y} \frac{1}{n}\sum_{i=1}^n \left|\gamma_\eta^\dagger(Z_i)\right|^2.$$

By the definition of Eq. (D.3), it is clear that $\widehat{S}_n \le \frac{2}{\alpha^2}\left[\widehat{\Xi}_n + \widehat{\Xi}_n^\dagger\right] + 2\eta_B^2$. Moreover, it is helpful to restrict the proof on the event

$$\mathcal{A}_n = \left\{\inf_{\eta \in \mathcal{B}_Y} \frac{1}{n}\sum_{i=1}^n |\gamma_\eta(Z_i)|^2 > c_o/2 \ \text{ and } \ \widehat{\Xi}_n, \widehat{\Xi}_n^\dagger \le M_o\right\},$$

where $M_o > 0$ is a sufficient large constant. The function class $\{|\gamma_\eta| : \eta \in \mathcal{B}_Y\}$ is of VC-type with $L^2(P)$-bounded envelope function, as established in the proof of Lemma 4.3. Moreover, the assumption of Theorem 4.1 ensures that $\inf_{\eta \in \mathcal{B}_Y} \mathbb{E}|\gamma_\eta(Z_i)|^2 > c_0$. By the Glivenko–Cantelli Theorem (e.g., Theorem 2.4.3 in van der Vaart and Wellner (1998)), we have

$$\inf_{\eta \in \mathcal{B}_Y} \frac{1}{n}\sum_{i=1}^n |\gamma_\eta(Z_i)|^2 \xrightarrow{a.s.} \inf_{\eta \in \mathcal{B}_Y} \mathbb{E}|\gamma_\eta(Z_i)|^2.$$

Similarly, we can show $\widehat{\Xi}_n \le M_o$ and $\widehat{\Xi}_n^\dagger \le M_o$, almost surely. This shows that $\lim_{n\to\infty} \mathbb{P}(\mathcal{A}_n) =$

1 and further

$$\lim_{n\to\infty} \sqrt{n} \left\{ \mathbb{E}\left[\mathcal{R}_n(\Theta_n)\right] - \mathbb{E}\left[\mathcal{R}_n(\Theta_n)\mathbb{1}_{\mathcal{A}_n}\right] \right\} = 0.$$

It is noted that on the event $\mathcal{A}_n$, the conditional distance D_n on Π_n is well defined.

Therefore, throughout the remainder of the proof, we will assume that the event $\mathcal{A}_n$ has occurred whenever appropriate. We structure the proof into the following four steps.

Step 1. We upper bound the first term of Eq. (D.5). By applying a union bound with Hoeffding's inequality, one has for all $t \geq 0$,

$$\begin{aligned}\mathbb{P}_\xi\left[\sup_{\theta\in\Theta_n(J_0)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i g_\theta(Z_i)\right| \geq t\right] \leq & 2|\Theta_n^{(J_0)}| \sup_{\theta\in\Theta_n^{(J_0)}} \exp\left[-\frac{t^2/2}{n^{-1}\sum_{i=1}^n |g_\theta(Z_i)|^2}\right] \\ = & 2|\Theta_n^{(J_0)}| \exp\left[-t^2/(2\widehat{S}_n)\right].\end{aligned}$$

We note the following fact: if X is a non-negative random variable satisfying $\mathbb{P}(X \leq t_k) \leq 1 - 2^{-k}$ for all $k \in \mathbb{N}^+$, then $\mathbb{E}(X) \leq \sum_{k=1}^\infty 2^{-k} t_k$. Consequently, by setting $t_k = 2\widehat{S}_n^{1/2}\sqrt{k + \log 2|\Theta_n^{(J_0)}|}$ for all $k \in \mathbb{N}^+$, we have

$$\begin{aligned}\mathbb{E}_\xi\left[\sup_{\theta\in\Theta_n(J_0)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i g_\theta(Z_i)\right|\right] &\leq 2\widehat{S}_n^{1/2}\sum_{k=1}^\infty \frac{1}{2^k}\sqrt{\log|\Theta_n^{(J_0)}| + \log 2 + k} \\ &\leq 2\widehat{S}_n^{1/2}\sum_{k=1}^\infty \frac{1}{2^k}\sqrt{\log|\Theta_n^{(J_0)}|} + 2\widehat{S}_n^{1/2}\sum_{k=1}^\infty \frac{1}{2^k}\left(\sqrt{k} + \log 2\right) \\ &\leq 2\widehat{S}_n^{1/2}\sqrt{\log 2|\Theta_n^{(J_0)}|} + 3\widehat{S}_n^{1/2}.\end{aligned}$$

It is clear that

$$\begin{aligned}\log 2|\Theta_n^{(J_0)}| &= \log|\Pi_n^{(J_0)}| + \log|\mathcal{B}^{(j)}| + \log 2 \\ &\leq \log N_H(4^{-J_0}, \Pi_n) + \log\left(\eta_B 2^{J_0+1}\right) + \log 2 \\ &\leq (10\log 2) J_0 \mathrm{VC}(\Pi_n) + (J_0+2)\log 2 + \log(\eta_B),\end{aligned}$$

then

$$\mathbb{E}_\xi\left[\sup_{\theta\in\Theta_n(J_0)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i g_\theta(Z_i)\right|\right] \leq 2\widehat{S}_n^{1/2}\left[\sqrt{(10\log 2)J_0\mathrm{VC}(\Pi_n) + (J_0+2))\log 2 + \log(\eta_B)} + \frac{3}{2}\right].$$

By choosing $J_0 = 1$, the inequality above is reduced to

$$\mathbb{E}_\xi\left[\sup_{\theta\in\Theta_n(J_0)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i g_\theta(Z_i)\right|\right] \leq 2\widehat{S}_n^{1/2}\left[\sqrt{(10\mathrm{VC}(\Pi_n) + 3)\log 2 + \log(\eta_B)} + \frac{3}{2}\right].$$

From the proof of Lemma 4.3, it is evident that the function classes $\{g_\theta : \theta \in \Theta_n\}$ admit a uniform envelope function for all n, which is bounded in $L^2(P)$. Therefore, by applying Jensen's

inequality together with the Glivenko–Cantelli Theorem (e.g., Theorem 2.4.3 in van der Vaart and Wellner (1998)), we obtain

$$\mathbb{E}\widehat{S}_n^{1/2} \leq \left|\mathbb{E}\widehat{S}_n\right|^{1/2} \leq S_n^{1/2} \equiv \sup_{\theta\in\Theta_n} \sqrt{\mathbb{E}|g_\theta(Z_i)|^2} < \infty.$$

As a result, we have

$$\mathbb{E}\left[\sup_{\theta\in\Theta_n(J_0)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i g_\theta(Z_i)\right|\right] \leq 2S_n^{1/2}\left[\sqrt{(10\mathrm{VC}(\Pi_n)+3)\log 2 + \log(\eta_B)} + \frac{3}{2}\right].$$

Step 2. By the definition of the operators Ψ_j for all $j \in \mathbb{N}^+$, one has $D_n\left(\Psi_{\Pi,j}(\pi), \Psi_{\Pi,j+1}(\pi)\right) \leq 2^{-j}$ and $|\Psi_{\mathcal{B}_Y,j+1}(\eta) - \Psi_{\mathcal{B}_Y,j}(\eta)| \leq 2^{-j}$. It is not difficult to see that for all $z \in \mathcal{Z}$ and $\theta \in \Theta_n$, we have $|g(x,a)| \leq \frac{1}{\kappa}$ and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \left|g_{\Psi_j(\theta)}(Z_i) - g_{\Psi_{j+1}(\theta)}(Z_i)\right|^2 &\leq 2\left(\bar{K}/\alpha+1\right)^2 |\Psi_{\mathcal{B}_Y,j}(\eta) - \Psi_{\mathcal{B}_Y,j+1}(\eta)|^2 \\ &\quad + \frac{2}{\alpha^2} D_n^2\left(\Psi_{\Pi,j}(\pi), \Psi_{\Pi,j+1}(\pi)\right)\widehat{\Xi}_n \\ &\leq 2^{-2j+1}\left(\bar{K}/\alpha+1\right)^2 + 2^{-2j+1}\widehat{\Xi}_n.\end{aligned}$$

For notational simplicity, let $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ represent the conditional probability and expectation given $(Z_i)_{i=1}^n$, with randomness only from $(\xi_i)_{i=1}^n$. Then, by Hoeffding's inequality, for any $\lambda \geq 0$ and $\theta \in \Theta_n$, one has

$$\begin{aligned}&\mathbb{P}_\xi\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\left\{g_{\Psi_j(\theta)}(Z_i) - g_{\Psi_{j+1}(\theta)}(Z_i)\right\}\right| \geq \lambda\right] \\ &\leq 2\exp\left[-\frac{\lambda^2/2}{n^{-1}\sum_{i=1}^n \left|g_{\Psi_j(\theta)}(Z_i) - g_{\Psi_{j+1}(\theta)}(Z_i)\right|^2}\right] \\ &\leq 2\exp\left[-\frac{\lambda^2/2}{\bar{K}^2|\Psi_j(\eta) - \Psi_{j+1}(\eta)|^2 + D_n^2\left(\Psi_j(\pi), \Psi_{j+1}(\pi)\right)\widehat{\Xi}_n/\alpha^2}\right] \\ &\leq 2\exp\left[-\frac{\lambda^2/2}{4^{-j}\bar{K}^2 + 4^{-j}\widehat{\Xi}_n\alpha^2}\right] = 2\exp\left[-\frac{2^{2j-1}\lambda^2}{(\bar{K}/\alpha+1)^2 + \widehat{\Xi}_n/\alpha^2}\right],\end{aligned}$$

For any given $\delta > 0$, we choose λ_j for each $j \in \mathbb{N}^+$ as follows:

$$\lambda_j = 2^{-j+1/2}\sqrt{\left(\bar{K}/\alpha+1\right)^2 + \widehat{\Xi}_n/\alpha^2}\sqrt{\log|\Theta_n(j+1)|\,[2\log j + \log(2/\delta)]}.$$

Then, for all $j \in \mathbb{N}^+$,

$$\mathbb{P}_\xi \left[\sup_{\theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \left\{ g_{\Psi_j(\theta)}(Z_i) - g_{\Psi_{j+1}(\theta)}(Z_i) \right\} \right| \geq \lambda_j \right] \leq \delta / j^2.$$

It is clear that

$$\begin{aligned}
\log(|\Theta_n(j+1)|) &= \log |\Pi_n^{(j+1)}| + \log |\mathcal{B}^{(j+1)}| \\
&\leq \log N_{D_n} \left(2^{-j-1}, \Pi_n, (Z_i)_{i=1}^n \right) + (j+1) \log 2 \\
&\leq \log N_H (4^{-j-1}, \Pi_n) + (j+1) \log 2 \\
&\leq 10(j+1) \mathrm{VC}(\Pi_n) + (j+1) \log 2.
\end{aligned}$$

For any $\delta > 0$, one has

$$\mathbb{P}_\xi \left[\sup_{\theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \sum_{j=J_0}^{J_n - 1} \left[g_{\Psi_j(\theta)}(Z_i) - g_{\Psi_{j+1}(\theta)}(Z_i) \right] \right| \geq \sum_{j=J_0}^{\infty} \lambda_j \right] \leq 1 - \delta.$$

Therefore, by setting $\delta_k = 2^{-k}$ for all $k \in \mathbb{N}^+$, one has

$$\begin{aligned}
&\mathbb{E}_\xi \left[\sup_{\theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \sum_{j=J_0}^{J_n - 1} \left[g_{\Psi_j(\theta)}(Z_i) - g_{\Psi_{j+1}(\theta)}(Z_i) \right] \right| \right] \\
\leq & \sum_{j=J_0}^{\infty} \lambda_j \leq \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \left(18 \sqrt{\mathrm{VC}(\Pi_n)} + 5 \right),
\end{aligned}$$

where the last inequality holds due to $J_0 = 1$ and the following derivation:

$$\begin{aligned}
\sum_{j=J_0}^{\infty} \lambda_j &= \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \sum_{j=J_0}^{\infty} 2^{-j+1/2} \sqrt{[10(j+1)\mathrm{VC}(\Pi_n) + (j+1)\log 2] \cdot \log(2/\delta_j)} \\
&\quad + \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \sum_{j=J_0}^{\infty} 2^{-j+1/2} \sqrt{[10(j+1)\mathrm{VC}(\Pi_n) + (j+1)\log 2] \cdot 2 \log j} \\
&\leq \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \left[\sqrt{(10 \log 2)\mathrm{VC}(\Pi_n)} + \log 2 \right] \sum_{j=J_0}^{\infty} 2^{-j+\frac{1}{2}} (j+1) \\
&\quad + \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \left[\sqrt{20 \mathrm{VC}(\Pi_n)} + \sqrt{2 \log 2} \right] \sum_{j=J_0}^{\infty} 2^{-j+\frac{1}{2}} (j+1)^{1/2} \sqrt{\log j} \\
&\leq \frac{17}{4} \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \left[\sqrt{(10 \log 2)\mathrm{VC}(\Pi_n)} + \log 2 \right] \\
&\quad + \frac{151}{100} \sqrt{\left(\bar{K}/\alpha + 1 \right)^2 + \widehat{\Xi}_n / \alpha^2} \left[\sqrt{20 \mathrm{VC}(\Pi_n)} + \sqrt{2 \log 2} \right].
\end{aligned}$$

Then, it follows that

$$\mathbb{E}\left[\sup_{\theta\in\Theta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\sum_{j=J_0}^{J_n-1}\left[g_{\Psi_j(\theta)}(Z_i)-g_{\Psi_{j+1}(\theta)}(Z_i)\right]\right|\right]\leq\sqrt{(\bar{K}/\alpha+1)^2+\Xi/\alpha^2}\left(18\sqrt{\mathrm{VC}(\Pi_n)}+5\right).$$

Step 3. We verify that the third term in Eq. (D.5) with $J(n)\leq j<J_+(n)$ are asymptotically negligible. We note that $\Psi_{J(n)}(\theta)=\Psi_{J(n)}(\Psi_{J_+(n)}(\theta))$, applying a union bound with Hoeffding's inequality gives

$$\begin{aligned}
&\mathbb{P}_\xi\left[\sup_{\theta\in\Theta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\left[g_{\Psi_{J(n)}(\theta)}(X_i)-g_{\Psi_{J_+(n)}(\theta)}(X_i)\right]\right|\geq t\right]\\
=&\,\mathbb{P}_\xi\left[\sup_{\theta\in\Theta_n(J_+(n))}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\left[g_{\theta}(X_i)-g_{\Psi_{J(n)}(\theta)}(X_i)\right]\right|\geq t\right]\\
\leq&\,2|\Theta_n\left(J_+(n)\right)|\exp\left[-\frac{2^{2J(n)-1}t^2}{(\bar{K}/\alpha+1)^2+\widehat{\Xi}_n/\alpha^2}\right].
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
\log|\Theta_n\left(J_+(n)\right)|&=\log|\Pi_n^{J_+(n)}|+\log|\mathcal{B}^{J_+(n)}|\\
&\leq\log N_{D_n}\left(2^{-J_+(n)},\Pi_n,(Z_i)_{i=1}^n\right)+\log\eta_B+(J_+(n)+1)\log 2\\
&\leq\log N_H\left(4^{-J_+(n)},\Pi_n\right)+\log\eta_B+(J_+(n)+1)\log 2\\
&\leq(5\log 4)J_+(n)\cdot n^{b_o}+(J_+(n)+1)\log 2.
\end{aligned}$$

Thus, recall $J_+(n)=(\log n)(1-b_o)$ and $J(n)=(\log n)(3-2b_o)/8$, one has

$$\begin{aligned}
&\mathbb{E}_\xi\left[\sup_{\theta\in\Theta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\left[g_{\Psi_{J(n)}(\theta)}(X_i)-g_{\Psi_{J_+(n)}(\theta)}(X_i)\right]\right|\right]\\
\leq&\,|\Theta_n(J_+(n))|\,2^{-2J(n)}(\bar{K}^2+\widehat{\Xi}_n)^{1/2}\\
\leq&\,\frac{(5\log 4)J_+(n)\cdot n^{b_o}+(J_+(n)+1)\log 2}{4^{J(n)}}\sqrt{\left(\bar{K}/\alpha+1\right)^2+\widehat{\Xi}_n/\alpha^2}=o_P(1).
\end{aligned}$$

Since the function class $\left\{\gamma_\eta^2:\eta\in\mathcal{B}_Y\right\}$ is P-Glivenko-Cantelli, then

$$\sup_{\eta\in\mathcal{B}_Y}\frac{1}{n}\sum_{i=1}^{n}|\gamma_\eta(Z_i)|^2\xrightarrow{a.s.}\sup_{\eta\in\mathcal{B}_Y}\mathbb{E}\,|\gamma_\eta(Z_i)|^2=\Xi.$$

Applying dominated convergence theorem on the term $\sqrt{\left(\bar{K}/\alpha+1\right)^2+\widehat{\Xi}_n/\alpha^2}$ gives

$$\lim_{n\to\infty}\mathbb{E}\left[\sup_{\theta\in\Theta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\left[g_{\Psi_{J(n)}(\theta)}(X_i)-g_{\Psi_{J_+(n)}(\theta)}(X_i)\right]\right|\right]=0.$$

Step 4. The forth term in Eq. (D.5) can be upper bounded as

$$\sup_{\theta\in\Theta_n}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\left[g_\theta(X_i)-g_{\Psi_{J_+(n)}(\theta)}(X_i)\right]\right|\le\sup_{\theta\in\Theta_n}\left|\frac{1}{n}\sum_{i=1}^{n}\left[g_\theta(X_i)-g_{\Psi_{J_+(n)}(\theta)}(X_i)\right]^2\right|^{1/2}$$
$$\le 2^{-J_+(n)+1/2}\sqrt{\left(\bar{K}/\alpha+1\right)^2+\widehat{\Xi}_n/\alpha^2}\xrightarrow{P}0.$$

Applying Donsker's theorem gives $\widehat{\Xi}_n\xrightarrow{P}\sup_{\eta\in\mathcal{B}_Y}\mathbb{E}\left|\gamma_\eta(Z_i)\right|^2$. Since $J_+(n)=(1-b_o)\log n$, applying Jensen's inequality and the dominated convergence theorem yields

$$\mathbb{E}\left[\sup_{\theta\in\Theta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\left[g_\theta(X_i)-g_{\Psi_{J_+(n)}(\theta)}(X_i)\right]\right|\right]\to 0.$$

As a result, combining these four steps, we have for n large enough,

$$\begin{aligned}\sqrt{n}\mathbb{E}_P\left[\mathcal{R}_n(\Theta_n)\right]&\le\sqrt{(\bar{K}/\alpha+1)^2+\Xi/\alpha^2}\left(18\sqrt{\mathrm{VC}(\Pi_n)}+5\right)\\&\quad+2S_n^{1/2}\left[\sqrt{(10\mathrm{VC}(\Pi_n)+3)\log 2+\log(\eta_B)}+\frac{3}{2}\right]\\&\le\left[5.3S_n^{1/2}+18\sqrt{(\bar{K}/\alpha+1)^2+\Xi/\alpha^2}\right]\sqrt{\mathrm{VC}(\Pi_n)}\\&\quad+5\sqrt{(\bar{K}/\alpha+1)^2+\Xi/\alpha^2}+(3+2\log(\eta_B))\,S_n^{1/2}+3,\end{aligned}\tag{D.6}$$

Finally, as argued by Bartlett and Mendelson (2002) in the proof of their Theorem 8, we have

$$\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}(P-\mathbb{P}_n)g_\theta\right]\le 2\mathbb{E}_P\left[\mathcal{R}_n(\Theta_n)\right]\quad\text{and}\quad\mathbb{E}_P\left[\sup_{\theta\in\Theta_n}(\mathbb{P}_n-P)g_\theta\right]\le 2\mathbb{E}_P\left[\mathcal{R}_n(\Theta_n)\right].\tag{D.7}$$

Recall Eq. (D.2), we have:

$$0\le\mathrm{Reg}\left(\widehat{\pi}_n\right)=\mathbb{V}(\theta)-\mathbb{V}(\widehat{\theta}_n)\le\mathbb{V}_n(\widehat{\theta}_n)-\mathbb{V}(\widehat{\theta}_n)+\mathbb{V}(\theta)-\mathbb{V}_n(\theta)+r_n$$

Taking expectations on both sides and combining this with Lemma 4.1, Eq. (D.6), and Eq. (D.7), we have

$$\limsup_{n\to\infty}\frac{\mathbb{E}_P\left[\mathrm{Reg}\left(\widehat{\pi}_n\right)\right]}{\left[5.3S_n^{1/2}+18\sqrt{(\bar{K}/\alpha+1)^2+\Xi/\alpha^2}\right]\sqrt{\mathrm{VC}(\Pi_n)/n}}\le 4.$$

Moreover, since $S_n\le\frac{2}{\alpha^2}\left[\Xi+\Xi^\dagger\right]$,

$$\limsup_{n\to\infty}\frac{\mathbb{E}_P\left[\mathrm{Reg}\left(\widehat{\pi}_n\right)\right]}{\sqrt{\mathrm{VC}(\Pi_n)/n}}\le\frac{30}{\alpha}\sqrt{\Xi+\Xi^\dagger}+72\sqrt{(\bar{K}/\alpha+1)^2+\Xi/\alpha^2}.$$

□

D.6 Proof of Lemma 4.3

Proof of Lemma 4.3. Recall that $g_\theta = \sum_{j=0}^5 g_{j,\theta}$, where the functions $g_{j,\theta} : z \mapsto g_{j,\theta}(z)$ are defined in Eq. (G.1), and $\mathcal{G}_{\Theta_n} \subset \oplus_{i=0}^5 \mathcal{G}_{j,\Theta_n}$ where $\mathcal{G}_{j,\Theta_n} = \{g_{j,\theta} : \theta \in \Theta_n\}$. We first construct envelope functions G_j for each $\mathcal{G}_{j,\Theta_n}$, provided Assumption 4.3 holds. Since $\mathcal{B}_Y$ is compact, there is $\eta_B > 0$ such that $\mathcal{B}_Y \subset [-\eta_B, \eta_B]$. It is clear that for all x and $\eta \in \mathcal{B}_Y$:

$$\begin{aligned}
|\mu_a(x,\eta)| &= \left|\mathbb{E}\left[(Y_i(a)-\eta)_- \mid X_i = x, A_i = a\right]\right| \\
&\le \mathbb{E}\left[\left|(Y_i(a)-\eta)_-\right| \mid X_i = x, A_i = a\right] \\
&\le \mathbb{E}\left[|Y_i(a)| \mid X_i = x, A_i = a\right] + \eta_B \equiv G_a(x),
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the second inequality holds due to $\mathcal{B}_Y \subset [-\eta_B, \eta_B]$. Moreover, it is easy to see G_a are $L^2(P)$-bounded for $a \in \{0,1\}$ due to Assumption 4.3 and G_a are envelope functions for $\mathcal{G}_{a,\Theta_n}$ for $a \in \{0,1\}$. Note that $1/e_o \le \kappa^{-1}$ and $1/(1-e_o) \le (1-\kappa)^{-1}$, letting $\bar{K} = \kappa^{-1} \vee (1-\kappa)^{-1}$, one has $|g_{3,\theta}(z)| \le G_3(z) \equiv \bar{K} G_0(z)$ and $|g_{5,\theta}(z)| \le G_5(z) \equiv \bar{K} G_1(z)$ for all z and $\theta \in \Theta_n$. Finally, for $j = 2, 4$,

$$|g_{j,\theta}(z)| \le G_j(z) \equiv \bar{K}(|y| + \eta_B),$$

where G_j are obviously $L^2(P)$-bounded.

By Theorem 2.6.7 in van der Vaart and Wellner (1998) and Lemma G.1, there are constants $c_o > 0$ such that

$$\sup_Q N\left(\epsilon \|G_j\|, \mathcal{G}_{j,\Theta_n}, L^2(Q)\right) \le (c_o/\epsilon)^{2\mathrm{VC}\left(\mathcal{G}_{j,\Theta_n}\right)}, \quad \forall \epsilon \in (0,1).$$

Let $G = \sum_{j=0}^5 G_j$ that is also $L^2(P)$-bounded, and an application of Lemma G.2 gives

$$\sup_Q N\left(\epsilon \|G\|_{Q,2}, \mathcal{G}_{\Theta_n}, L^2(Q)\right) \le (c_o/\epsilon)^{24\mathrm{VC}(\Pi_n)+24},$$

where supremum is taken over all discrete probability measures Q on $\mathcal{Z}$. □

D.7 Proof of Lemma 5.1

Proof of Lemma 5.1. Lemma 4.3 and Theorem 2.5.2 in van der Vaart and Wellner (1998) implies $\mathcal{G}_\Theta = \{g_\theta : \theta \in \Theta\}$ is P-Donsker and hence P-Glivenko-Cantelli. Consequently,

$$\sup_{\theta\in\Theta} |\mathbb{V}_n(\theta) - \mathbb{V}(\theta)| = \sup_{\theta\in\Theta} |(\mathbb{P}_n - P) g_\theta| = o_P(1).$$

Consider the following derivation:

$$\begin{aligned}\mathbb{V}_n(\widehat{\theta}_n) - \mathbb{V}_n(\theta_o) = &\underbrace{\mathbb{V}_n(\widehat{\theta}_n) - \widehat{\mathbb{V}}_n(\widehat{\theta}_n)}_{=o_P(n^{-1/2})} + \underbrace{\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \widehat{\mathbb{V}}_n(\check{\theta}_n)}_{\geq 0} \\ &+ \underbrace{\widehat{\mathbb{V}}_n(\check{\theta}_n) - \mathbb{V}_n(\check{\theta}_n)}_{=o_P(n^{-1/2})} + \underbrace{\mathbb{V}_n(\check{\theta}_n) - \mathbb{V}_n(\theta_o)}_{\geq 0},\end{aligned}$$

where the first and third terms are $o_P(n^{-1/2})$ by Lemma 4.1, and the second and fourth terms are guaranteed to be greater than zero according to the definitions of $\widehat{\mathbb{V}}_n$ and $\mathbb{V}_n$. This shows $\mathbb{V}_n(\widehat{\theta}_n) \geq \mathbb{V}_n(\theta_o) - o_P(1)$, and hence Theorem 5.7 in van der Vaart (2000) implies $\|\widehat{\theta}_n - \theta_o\| = o_P(1)$. □

D.8 Proof of Theorem 5.1

Proof of Theorem 5.1. Since $\Pi_n = \Pi$ for all n, it follows from Lemma 4.3 that $\Theta = \Pi \times \mathcal{B}_Y$ is Donsker. Leveraging Lemma 4.1 and an argument analogous to Theorem 1 in Luedtke and Chambaz (2020), we can establish that $(\mathbb{P}_n - P)(g_{\widehat{\theta}_n} - g\theta_o) = o_P(n^{-1/2})$ and $\mathbb{V}(\widehat{\theta}_n) - \mathbb{V}(\theta_o) = o_P(n^{-1/2})$. Consequently, from Eq. (5.1), we have:

$$\begin{aligned}\widehat{\mathbb{V}}_n(\widehat{\theta}_n) - \mathbb{V}(\theta_o) &= (\mathbb{V}_n - \mathbb{V})(\theta_o) + (\mathbb{P}_n - P)(g_{\widehat{\theta}_n} - g_{\theta_o}) + (\widehat{\mathbb{V}}_n - \mathbb{V})(\widehat{\theta}_n) + \mathbb{V}(\widehat{\theta}_n) - \mathbb{V}(\theta_o) \\ &= (\mathbb{V}_n - \mathbb{V})(\theta_o) + o_P(n^{-1/2}) \\ &= (\mathbb{P}_n - P)g_{\theta_o} + o_P(n^{-1/2}).\end{aligned}$$

The desired result follows from the central limit theorem. □

E Proofs of Results for Improved Rates under Margin Assumption

The proof of Theorem B.1 relies on Lemma E.1, which provides control over the continuity modulus of the empirical process $\theta \mapsto \mathbb{G}_n g_\theta$.

Lemma E.1. Suppose Assumption 5.1 (1) holds. There is a universal constant $c_o > 0$ not depending on n such that for every $\theta \in \Theta$, for any $\delta > 0$ small enough, one has

$$\mathbb{E}\left[\sup_{\theta' \in \Theta : \|\theta' - \theta\| \leq \delta} |\mathbb{G}_n(g_{\theta'} - g_\theta)|\right] \leq c_o(\mathrm{VC}(\Pi)^{1/2} + n^{-1/2}\mathrm{VC}(\Pi))\delta.$$

Proof. Fix $\theta \in \Theta$, we write $\mathcal{G}_{j,\delta}^- \equiv \{g_{j,\theta'} - g_{j,\theta} : \|\theta' - \theta\| < \delta, \theta' \in \Theta\}$ for $0 \leq j \leq 5$, and all the functions in these classes are uniformly bounded due to Assumption 5.1 (1) and Assumption 2.1. We study the first term. Fix any $\delta > 0$. There is a universal constant $K > 0$

such that for $\theta' \equiv (\pi', \eta') \in \Theta$ with $\|\theta' - \theta\| < \delta$,

$$\begin{aligned}|g_{0,\theta'}(z) - g_{0,\theta}(z)| &\leq \mu_0(x,\eta')\left(\pi' - \pi\right)(x) + \pi(x)\left[\mu_0(x,\eta') - \mu_0(x,\eta)\right] \\ &\leq \sup_{x\in\mathcal{X},\eta\in\mathcal{B}_Y} |\mu_0(x,\eta)| \left|(\pi' - \pi)(x)\right| + \delta.\end{aligned}$$

Let $G_o(z) \equiv \delta\left(1 + \sup_{x\in\mathcal{X},\eta\in\mathcal{B}_Y} |\mu_0(x,\eta)|\right)$. Since $\mathrm{VC}(\mathcal{G}^-_{0,\delta}) \leq 2\mathrm{VC}(\Pi) + 3$, then there are constants $A > 0$ such that

$$\sup_Q \log N(\epsilon\|G_o\|, \mathcal{G}^-_{0,\delta}, L^2(Q)) \lesssim \mathrm{VC}(\Pi)\log\left(A/\epsilon\right),$$

for all finitely discrete measure Q. We note that $\sup_{f\in\mathcal{G}^-_{1,\delta}} Pf^2 \lesssim \delta^2 \leq \|G_o\|^2_{P,2}$, and an application of Corollary 5.1 in Chernozhukov et al. (2014) yields

$$\begin{aligned}\mathbb{E}_P\left[\|\mathbb{G}_n\|_{\mathcal{G}^-_{0,\delta}}\right] &\lesssim \sqrt{\mathrm{VC}(\Pi)\delta^2\log A} + \frac{\mathrm{VC}(\Pi)\|G_o\|_\infty}{\sqrt{n}}\log A \\ &\lesssim \delta\left[\mathrm{VC}(\Pi)^{1/2} + n^{-1/2}\mathrm{VC}(\Pi)\right].\end{aligned}$$

Using the identical argument, we can show

$$\mathbb{E}_P\left[\|\mathbb{G}_n\|_{\mathcal{G}^-_{j,\delta}}\right] \lesssim \left(\mathrm{VC}(\Pi)^{1/2} + n^{-1/2}\mathrm{VC}(\Pi)\right)\delta, \quad \forall 1 \leq j \leq 5.$$

The desired result follows from

$$\mathbb{E}\left[\sup_{\theta\in\Theta:\|\theta-\theta_o\|<\delta} \mathbb{G}_n\left(g_\theta - g_{\theta_o}\right)\right] \leq \sum_{j=0}^{5} \mathbb{E}_P\left[\|\mathbb{G}_n\|_{\mathcal{G}^-_{j,\delta}}\right] \lesssim \delta\left[\mathrm{VC}(\Pi)^{1/2} + n^{-1/2}\mathrm{VC}(\Pi)\right].$$

□

Proof of Theorem B.1. By Assumption B.1, there is a small constant $\delta_o > 0$ such that

$$\{\theta : \mathbb{V}(\theta_o) - \mathbb{V}(\theta) \leq c_o\delta^{\rho_o}\} \subset \{\theta : \|\theta - \theta_o\| \leq \delta\}, \quad \forall \delta < \delta_o.$$

Hence, to obtain the convergence rate of $\|\check{\theta}_n - \theta_o\|$, we only need to study the concentration of $\mathbb{V}(\check{\theta}_n) - \mathbb{V}(\theta_o)$. The rest of the proof is highly inspired by Theorem 2 in Massart and Nédélec (2006). Let Θ' be a countable dense subset of Θ. Let

$$\epsilon_n = \left[(\mathrm{VC}(\Pi)/n)^{1/2} + \mathrm{VC}(\Pi)/n\right]^{\rho_o/(2\rho_o - 1)},$$

and there must be $\theta'_o \in \Theta'$ such that $\mathbb{V}(\theta_o) - \mathbb{V}(\theta'_o) \leq \epsilon_n^2$. We start from the identity

$$\begin{aligned}\mathbb{V}(\theta_o) - \mathbb{V}(\check{\theta}_n) &= \ell(\theta_o, \theta'_o) - \mathbb{P}_n(g_{\theta'_o} - g_{\check{\theta}_n}) + (\mathbb{P}_n - P)(g_{\theta'_o} - g_{\check{\theta}_n}) \\ &\leq \epsilon_n^2 + (\mathbb{P}_n - P)\left(g_{\theta'_o} - g_{\check{\theta}_n}\right).\end{aligned}$$

Let $x = c_o t^{1/2} \epsilon_n$, where K is a constant to be chosen later and

$$V_n(x) = \sup_{\theta \in \Theta'} \frac{(\mathbb{P}_n - P)(g_{\theta'_o} - g_\theta)}{P(g_{\theta'_o} - g_\theta) + \epsilon_n^2 + x^2}.$$

Since $\mathbb{V}(\theta_o) = P g_{\theta_o} \geq P g_{\theta'_o} = \mathbb{V}(\theta'_o)$, then

$$\mathbb{V}(\theta_o) - \mathbb{V}(\check{\theta}_n) \leq \mathbb{V}(\theta_o) - \mathbb{V}(\theta'_o) + V_n(x) \left[\mathbb{V}(\theta_o) - \mathbb{V}(\check{\theta}_n) + x^2 + \epsilon_n^2\right].$$

On the event $V_n(x) < \frac{1}{2}$, one has

$$\mathbb{V}(\theta_o) - \mathbb{V}(\check{\theta}_n) < 2\left[\mathbb{V}(\theta_o) - \mathbb{V}(\theta'_o)\right] + \epsilon_n^2 + x^2 \leq 3\epsilon_n^2 + x^2,$$

and hence

$$\mathbb{P}\left[\mathbb{V}(\theta_o) - \mathbb{V}(\check{\theta}_n) \geq 3\epsilon_n^2 + x^2\right] \leq \mathbb{P}\left[V_n(x) \geq 1/2\right].$$

Since $\tau(x)$ is uniformly bounded, it is clear that there is some sufficiently large $c_o > 0$ such that

$$\sup_{z \in \mathcal{Z}} |g_\theta(z) - g_{\theta_o}(z)| \leq c_o \|\theta - \theta_o\|.$$

As a result, the class $\{g_{\theta_o} - g_\theta : \theta \in \Theta\}$ is uniformly bounded, and hence

$$\sup_{\theta \in \Theta'} \mathrm{Var}\left[\frac{(g_{\theta_o} - g_\theta)(Z_i)}{P(g_{\theta_o} - g_\theta) + x^2}\right] \leq c_o x^{-4} \quad \text{and} \quad \sup_{\theta \in \Theta'} \left\| \frac{(g_{\theta_o} - g_\theta)(Z_i)}{P(g_{\theta_o} - g_\theta) + x^2} \right\|_\infty \leq c_o x^{-2}.$$

Applying the Talagrand's inequality yields that the follow inequality holds

$$V_n(x) < \mathbb{E}\left[V_n(x)\right] + \sqrt{\frac{K\left(x^{-2} + 4\mathbb{E}\left[V_n(x)\right]\right)t}{nx^2}} + \frac{2c_o x^{-2} t}{3n}$$

with probability greater than $1 - e^{-t}$. By the definition of $x = c_o t^{1/2} \epsilon_n$, applying Lemma A.5 in Massart and Nédélec (2006) and Lemma E.1 gives

$$\begin{aligned}
\mathbb{E}[V_n(x)] &\leq \mathbb{E}\left[\sup_{\theta \in \Theta' : \|\theta - \theta_o\| < \delta/c_o} \frac{(\mathbb{P}_n - P)(g_{\theta_o} - g_\theta)}{\mathbb{V}(\theta_o) - \mathbb{V}(\theta) + x^2}\right] \\
&\leq \mathbb{E}\left[\sup_{\theta \in \Theta' : \mathbb{V}(\theta_o) - \mathbb{V}(\theta) < \delta} \frac{(\mathbb{P}_n - P)(g_{\theta_o} - g_\theta)}{\mathbb{V}(\theta_o) - \mathbb{V}(\theta) + x^2}\right] \leq 4n^{-1/2} x^{-2} \varphi_n(x) \\
&= 4n^{-1/2}(c_o t^{1/2} \epsilon_n)^{-2} c_o \left(\mathrm{VC}(\Pi)^{1/2} + n^{-1/2}\mathrm{VC}(\Pi)\right) \epsilon_n^{1/\rho_o}.
\end{aligned}$$

By the definition of ϵ_n, we can choose $c_o > 0$ large enough, and there is N_o such that $\mathbb{E}\left[V_n(x)\right] < 1/100$ for all $n \geq N_o$. Choosing c_o large enough, it follows that

$$\frac{2c_o x^{-2} t}{3n} < \frac{1}{100} \quad \text{and} \quad \sqrt{\frac{c_o\left(x^{-2} + 4\mathbb{E}\left[V_n(x)\right]\right)t}{nx^2}} < \frac{1}{100}.$$

As a result, $\mathbb{P}\left[V_n(x) < 1/2\right] \geq 1 - e^{-t}$, and

$$\mathbb{P}\left[\mathbb{V}\left(\theta_o\right) - \mathbb{V}(\check{\theta}_n) \geq 3\epsilon_n^2 + x^2\right] \leq e^{-t}.$$

By the definition of x and $t \geq 1$, there must be a large $c_o > 0$ not depending on n such that

$$\mathbb{P}\left[\mathbb{V}\left(\theta_o\right) - \mathbb{V}(\check{\theta}_n) \geq c_o t\epsilon_n^2\right] \leq e^{-t}.$$

Since $\mathbb{V}\left(\theta_o\right) - \mathbb{V}(\check{\theta}_n) \geq 0$, an application of Lemma 2.2.13 in Durrett (2019) gives

$$\mathbb{E}_P\left[\ell(\theta_o, \check{\theta}_n)\right] \lesssim \left(\mathrm{VC}(\Pi)/n\right)^{\frac{\rho_o}{2\rho_o - 1}}.$$

□

F Proofs of Results for Uniform Inference for the Optimal Welfare

Let $\ell^\infty(\Theta)$ denote the space of all uniformly bounded functions from Θ to $\mathbb{R}$. Let $C_b(\Theta)$ denote the space of continuous and uniformly bounded functions on Θ.

F.1 Proof of Theorem C.1

As stated in Appendix C, Theorem C.1 directly follows from the uniform weak convergence of $\sqrt{n}(\widehat{\mathbb{V}}_n - \mathbb{V})$ and the uniformly valid functional delta method. Lemma F.1 establishes this uniform weak convergence, while Lemma F.2 verifies that the supremum functional is Hadamard directionally differentiable, thereby enabling the application of the delta method to construct inference for the optimal welfare.

Lemma F.1. Under the same assumptions in Theorem C.1, the following asymptotic approximation holds uniformly for all $P \in \mathcal{P}_n$:

$$\sqrt{n}\big(\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}_P(\theta)\big)_{\theta\in\Theta} = (\mathbb{G}_n g_\theta)_{\theta\in\Theta} + o_P(1), \quad \text{in } \ell^\infty(\Theta).$$

Moreover, we obtain the uniform weak convergence of $\sqrt{n}\big(\widehat{\mathbb{V}}_n - \mathbb{V}_P\big) \rightsquigarrow \mathbb{G}_P$, namely

$$\sqrt{n}\big(\widehat{\mathbb{V}}_n(\theta) - \mathbb{V}_P(\theta)\big)_{\theta\in\Theta} \rightsquigarrow (\mathbb{G}_P g_\theta)_{\theta\in\Theta}, \quad \text{in } \ell^\infty(\Theta),$$

uniformly in $P \in \mathcal{P}_n$, where $\mathbb{G}_P : \theta \mapsto \mathbb{G}_P g_\theta$ is defined in Theorem C.1. The process $\sqrt{n}(\widehat{\mathbb{V}}_n - \mathbb{V}_P)$ is stochastically equicontinuous uniformly over $P \in \mathcal{P}_n$.

Proof of Lemma F.1. Lemma A.1 in Rai (2018) implies that (Π, d_Π) is totally bounded, and its covering number satisfies $N(\epsilon, \Pi, d_\Pi) \leq C(e/\epsilon)^{\mathrm{VC}(\Pi)}$ for some universal constant $C > 0$.

To establish this theorem, we apply Theorem 5.1 from Belloni et al. (2017). Given Assumption C.1, it remains to verify Assumptions 5.1 and 5.2 in Belloni et al. (2017).

Assumption 5.1 in Belloni et al. (2017) is readily verified in our setting, as $\mathbb{V}_P(\theta)$ is identified by a linear moment condition and is uniformly bounded over all $P \in \mathcal{P}_n$.

Next, we verify Assumption 5.2 in Belloni et al. (2017) holds. Since $|Y_i| \leq c_o$ under all $P \in \mathcal{P}_n$, without loss of generality, we assume $\mathcal{B}_Y = [-c_o, c_o]$. We note that $\eta \in \mathcal{B}_Y$, where $\mathcal{B}_Y$ is bounded and $e_P \in (\delta, 1-\delta)$ for all $P \in \mathcal{P}_n$. Moreover, for all $\eta, \tilde{\eta} \in \mathcal{B}_Y$, one has $|(y-\eta) - (y-\tilde{\eta})_-| \leq |\eta - \tilde{\eta}|$ and

$$\begin{aligned} |\mu_{a,P}(z,\eta) - \mu_{a,P}(z,\tilde{\eta})| &= \mathbb{E}_P\left[(Y_i(a)-\eta)_- - (Y_i(a)-\tilde{\eta})_- | X_i = x\right] \\ &\leq |\eta - \tilde{\eta}|. \end{aligned}$$

Then it is easy to show $g_\theta(z, \mu_P, e_P)$ is Lipschitz continuous in θ, i.e., there is a constant C such that

$$\left|g_\theta(z,\mu_P,e_P) - g_{\tilde{\theta}}(z,\mu_P,e_P)\right| \leq C\left[|\tilde{\pi}(x) - \pi(x)| + |\eta - \tilde{\eta}|\right].$$

Therefore, by Assumption C.1 (2), there is a constant $C > 0$ such that the following inequality holds for all $\theta, \tilde{\theta}$ and $P \in \mathcal{P}_n$:

$$\|g_{\theta,P} - g_{\tilde{\theta},P}\|_{P,2} \leq C\left[\|\pi - \tilde{\pi}\|_{P,2} + |\eta - \tilde{\eta}|\right] \leq C d_\Theta(\theta, \tilde{\theta}).$$

□

Lemma F.2. The functional $\psi : h \mapsto \sup_\Theta h(\theta)$ mapping $\ell^\infty(\Theta)$ to $\mathbb{R}$ is Hadamard directionally differentiable at $\mathbb{V}_P$ with with the linear derivative map $\psi'_P : h \mapsto \sup_{\theta \in \Pi^\star_P} h(\theta)$. Specifically, for any sequences $\{h_n\} \subset \ell^\infty(\Theta)$ and $\{t_n\}$ such that $h_n \to h \in \ell^\infty(\Theta)$ and $t_n \searrow 0$, it holds that

$$\lim_{n\to\infty} \left| \frac{\psi(\mathbb{V}_P + t_n h_n) - \psi(\mathbb{V}_P)}{t_n} - \psi'_P(h) \right| = 0.$$

Proof. Since $h_n \to h$ in $\ell^\infty(\Theta)$, it is clear that

$$\left| \frac{\psi(\mathbb{V}_P + t_n h_n) - \psi(\mathbb{V}_P + t_n h)}{t_n} \right| \leq \sup_{\theta\in\Theta} |h_n(\theta) - h(\theta)| \to 0.$$

By the triangle inequality, to show this lemma, it suffices to show

$$\lim_{n\to\infty} \left| \frac{\psi(\mathbb{V}_P + t_n h) - \psi(\mathbb{V}_P)}{t_n} - \psi'_P(h) \right| = 0.$$

For any $\delta > 0$, define $\Theta_\delta = \{\theta \in \Theta : \mathbb{V}_P(\theta) + \delta > \sup_{\theta\in\Theta} \mathbb{V}_P(\theta)\}$. Since $h_n \in C_b(\Theta)$, we

let $\delta_n = 2t_n\|h\|_\infty$ and it is clear that

$$\frac{\sup_{\theta\in\Theta}\{\mathbb{V}_P(\theta)+t_nh(\theta)\}-\mathbb{V}_P(\theta_o)}{t_n} = \frac{\sup_{\theta\in\Theta_{\delta_n}}\{\mathbb{V}_P(\theta)+t_nh(\theta)\}-\mathbb{V}_P(\theta_o)}{t_n}.$$

The term on the RHS satisfies

$$\sup_{\theta\in\Theta_P^\star} h(\theta) \le \frac{\sup_{\theta\in\Theta_{\delta_n}}\{\mathbb{V}_P(\theta)+t_nh(\theta)\}-\mathbb{V}_P(\theta_o)}{t_n} \le \sup_{\theta\in\Theta_{\delta_n}} h(\theta). \tag{F.1}$$

We finish the proof by using contradiction to show $\sup_{\theta\in\Theta_{\delta_n}} h(\theta) \to \sup_{\theta\in\Theta_P^\star} h(\theta)$. Suppose that there is $\varepsilon_0 > 0$ such that

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_{\delta_n}} h(\theta) - \max_{\theta\in\Theta_P^*} h(\theta) > \varepsilon_0.$$

Without loss of generality, we assume $\sup_{\theta\in\Theta_{\delta_n}} h(\theta) - \max_{\theta\in\Theta_P^*} h(\theta) > \varepsilon_0$ for all n. For all n, let $\theta_n \in \Theta_{\delta_n}$ such that $h(\theta_n) > \sup_{\theta\in\Theta_{\delta_n}} h(\theta) - 1/n$. Since Θ is totally bounded, $\{\theta_n\}$ has a subsequence $\{\theta_{n_k}\}_{k\ge1}$ that converges to $\bar\theta_0 \in \Theta$. We note that $\mathbb{V}_P : \theta \mapsto \mathbb{V}_P(\theta)$ is continuous, then $\mathbb{V}_P(\theta_{n_k}) \to \mathbb{V}_P(\bar\theta_0)$. By the definition of Θ_{δ_n}, $|\mathbb{V}_P(\theta_o) - \mathbb{V}_P(\theta_{k_n})| \le \delta_{k_n}$ and letting $n\to\infty$ yields $\mathbb{V}_P(\bar\theta_0) = \mathbb{V}_P(\theta_o) = \sup_{\theta\in\Theta}\mathbb{V}_P(\theta)$ and $\bar\theta_0 \in \Theta_P^*$. Since $h \in C_b(\Theta)$ is continuous, $h(\bar\theta_0) - \max_{\theta\in\Theta_P^*} h(\theta) > \varepsilon_0/2$ for n large enough. Thus, $h(\bar\theta_0) > \max_{\theta\in\Theta_P^*} h(\theta)$, which contradicts $\bar\theta_0 \in \Theta_P^*$.

Therefore, by Eq. (F.1) and letting $n\to\infty$ gives

$$\lim_{n\to\infty} \frac{\sup_{\theta\in\Theta_{\delta_n}}\{\mathbb{V}_P(\theta)+t_nh(\theta)\}-\mathbb{V}_P(\theta_o)}{t_n} = \sup_{\theta\in\Theta_P^*} h(\theta) = \mathbb{V}_P'(h).$$

□

F.2 Proof of Lemmas F.3 and F.4

Recall the numerical derivative $\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*)$ as defined in Eq. (5.3). We establish that this quantity consistently estimates $\psi_P'(\mathbb{G}_P)$ for any fixed $P \in \mathcal{P}_n$. Recall that $\{\xi_i\}_{i=1}^n$ are i.i.d. random variables independent of $(Z_i)_{i=1}^n$, with $\mathbb{E}(\xi_i) = 0$, $\mathbb{E}(\xi_i^2) = 1$ and $\mathbb{E}[\exp|\xi_i|] < \infty$.

Lemma F.3. Under the same assumptions in Theorem C.1, then $\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*) \xrightarrow{P} \psi_P'(\mathbb{G}_P)$, for any fixed $P \in \mathcal{P}_n$.

Proof of Lemma F.3. The result follows directly from Theorem 3.1 in Hong and Li (2018). □

Next, we show that the one-sided confidence interval in Eq. (5.4) is uniformly valid over $P \in \mathcal{P}_n$, whereas the two-sided confidence interval in Eq. (5.5) is valid for any fixed $P \in \mathcal{P}_n$. Recall c_γ denoted the γ-empirical quantile of $\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*)$ and $q_{1-\gamma}$ denotes the $(1-\gamma)$-empirical quantile of $\left|\widehat{\psi}_n'(\widehat{\mathbb{G}}_n^*)\right|$ for any $\gamma > 0$.

Lemma F.4. Under the same assumptions in Theorem C.1, then

$$\lim_{n\to\infty}\inf_{P\in\mathcal{P}_n}\mathbb{P}\left[\mathbb{V}_P(\theta_o)\geq\sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta)-c_{1-\gamma}/\sqrt{n}\right]\geq 1-\gamma. \tag{F.2}$$

Moreover, for any fixed $P\in\mathcal{P}_n$

$$\liminf_{n\to\infty}\mathbb{P}\left[\left|\sup_{\theta\in\Theta}\widehat{\mathbb{V}}_n(\theta)-\mathbb{V}(\theta_o)\right|\leq q_{1-\gamma}/\sqrt{n}\right]\geq 1-\gamma. \tag{F.3}$$

Proof of Lemma F.4. The validity of the two-sided confidence interval, as stated in Eq. (F.3), follows directly from Lemma F.3. The uniform validity of the one-sided confidence interval in Eq. (F.2) can be established either by applying Theorem 3.5 in Hong and Li (2018), or by adapting the proof of Theorem 3 in Rai (2018). Noting the convexity of ψ_P and invoking Lemma F.5, the desired result follows by the same argument used in Rai (2018). □

The following lemma verifies the validity of multiplier bootstrap in our context.

Lemma F.5. Under the same assumptions in Theorem C.1, we have

$$\sup_{P\in\mathcal{P}_n}\sup_{h\in\mathrm{BL}_1(\ell^\infty(\Theta))}\left|\mathbb{E}_{B_n}[h(\widehat{\mathbb{G}}_n^*)]-\mathbb{E}[h(\mathbb{G}_P)]\right|=o_P(1),$$

where $\mathbb{E}_{B_n}$ denotes the expectation over the multiplier weights $(\xi_i)_{i=1}^n$ holding $(Z_i)_{i=1}^n$ fixed.

Proof. Define $\mathbb{G}_n^*$ denote the stochastic process $\theta\mapsto n^{-1}\sum_{i=1}^n\xi_i[g_\theta(Z_i)-\mathbb{V}_P(\theta)]$. It is clear that

$$\begin{aligned}\sup_{h\in\mathrm{BL}_1(\ell^\infty(\Theta))}\left|\mathbb{E}_{B_n}[h(\widehat{\mathbb{G}}_n^*)]-\mathbb{E}[h(\mathbb{G}_P)]\right|\leq&\sup_{h\in\mathrm{BL}_1(\ell^\infty(\Theta))}\left|\mathbb{E}_{B_n}[h(\widehat{\mathbb{G}}_n^*)]-\mathbb{E}_{B_n}[h(\mathbb{G}_n^*)]\right|\\&+\sup_{h\in\mathrm{BL}_1(\ell^\infty(\Theta))}\left|\mathbb{E}_{B_n}[h(\mathbb{G}_n^*)]-\mathbb{E}[h(\mathbb{G}_P)]\right|.\end{aligned}$$

Thus, it is sufficient to show

$$\begin{aligned}\sup_{h\in\mathrm{BL}_1(\ell^\infty(\Theta))}\left|\mathbb{E}_{B_n}[h(\widehat{\mathbb{G}}_n^*)]-\mathbb{E}_{B_n}[h(\mathbb{G}_n^*)]\right|&=o_P(1)\\\sup_{h\in\mathrm{BL}_1(\ell^\infty(\Theta))}\left|\mathbb{E}_{B_n}[h(\mathbb{G}_n^*)]-\mathbb{E}[h(\mathbb{G}_P)]\right|&=o_P(1).\end{aligned}$$

First, we note that

$$\begin{aligned}\sup_{h\in \mathrm{BL}_1(\ell^\infty(\Theta))} \left|\mathbb{E}_{B_n}[h(\widehat{\mathbb{G}}_n^*)] - \mathbb{E}_{B_n}[h(\mathbb{G}_n^*)]\right| &= \sup_{h\in \mathrm{BL}_1(\ell^\infty(\Theta))} \left|\mathbb{E}_{B_n}[h(\widehat{\mathbb{G}}_n^*) - h(\mathbb{G}_n^*)]\right| \\ &\leq \mathbb{E}_{B_n}\left[2 \wedge \sup_{\theta\in\Theta}\left|n^{-1/2}\sum_{i=1}^n \xi_i(\widehat{g}_\theta - g_\theta)(Z_i)\right|\right] \\ &+ \mathbb{E}_{B_n}\left[2 \wedge \sup_{\theta\in\Theta}\left|n^{-1/2}\sum_{i=1}^n \xi_i(\widehat{\mathbb{V}}_n - \mathbb{V}_P)(\theta)\right|\right].\end{aligned} \tag{F.4}$$

The sequence $(\xi_i)_{i=1}^n$ is independent of $(\widehat{g}_\theta - g_\theta(Z_i))_{i=1}^n$ and and, by Assumption 4.2, we have $\sup_{\theta,z}|\widehat{g}_\theta(z) - g_\theta(z)| = o_P(1)$. Using an argument similar to the proof of Lemma 4.1, it follows that the first term on the RHS of Eq. (F.4) is $o_P(1)$. Moreover, by Lemma 4.1, $\sup_{\theta\in\Theta}|(\widehat{\mathbb{V}}_n - \mathbb{V}_n)(\theta)| = o_P(n^{-1/2})$. Consequently, the second term on the right-hand side of Eq. (F.4) also converges to zero in probability.

Therefore, to end the proof, it suffices to show

$$\sup_{h\in \mathrm{BL}_1(\ell^\infty(\Theta))} |\mathbb{E}_{B_n}[h(\mathbb{G}_n^*)] - \mathbb{E}[h(\mathbb{G}_P)]| = o_P(1).$$

Since the function class $\{g_\theta : \theta \in \Theta\}$ is P-Donsker, this result follows from Theorem 2.9.6 in van der Vaart and Wellner (1998) or Theorem B.2 in Belloni et al. (2017). □

G Auxiliary Lemmas

Lemma G.1. Define functions indexed by θ as

$$\begin{aligned} g_{0,\theta}(z) &= \pi(x)\mu_0(x,\eta), \quad g_{1,\theta}(z) = (1-\pi(x))\,\mu_1(x,\eta), \\ g_{2,\theta}(z) &= \frac{(1-a)(1-\pi(x))(y-\eta)_-}{1-e_o(x)}, \\ g_{3,\theta}(z) &= -\frac{(1-a)(1-\pi(x))\mu_0(x,\eta)}{1-e_o(x)}, \\ g_{4,\theta}(z) &= \frac{\pi(x)a(y-\eta)_-}{e_o(x)}, \quad g_{5,\theta}(z) = -\frac{\pi(x)a\mu_1(x,\eta)}{e_o(x)}. \end{aligned} \tag{G.1}$$

Let $\mathcal{G}_{j,\Theta_n} \equiv \{g_{j,\theta} : \theta \in \Theta_n\}$ and $\mathcal{G}_{j,\theta}^- \equiv \{g_{j,\theta} - g_{j,\theta_o} : \theta \in \Theta_n\}$ for $0 \leq j \leq 5$, where the function $g_{j,\theta}$ are defined in Eq. (G.1). Then, for $0 \leq j \leq 5$,

$$\mathrm{VC}(\mathcal{G}_{j,\theta}) \leq 2\mathrm{VC}(\Pi_n) + 2 \quad \text{and} \quad \mathrm{VC}(\mathcal{G}_{j,\theta}^-) \leq 2\mathrm{VC}(\Pi_n) + 3.$$

Proof. By Theorem 2.6.18 in van der Vaart and Wellner (1998), to finish the proof, it suffices

to consider the VC-dimension of $\mathcal{G}_{j,\Theta_n}$. The subgraph of $g_{0,\theta}$ is the union of disjoint sets

$$C_\theta^+ = \{(x,t) : \pi(x) > 0\} \cap \{(x,t) : \mu_0(x,\eta) > t\},$$
$$C_\theta^- = \{(x,t) : \pi(x) \leq 0\} \cap \{(x,t) : t < 0\}.$$

First, we note that Π_n is of VC-index $\text{VC}(\Pi_n)$. Since the subgraph of $x \mapsto \mu_0(x,\eta_1)$ is contained in the subgraph of $x \mapsto \mu_0(x,\eta_2)$ if $\eta_1 \leq \eta_2$, then the collection of sets that take the form of $\{(x,t) : \mu_0(x,\eta) > t\}$ has VC-index 2. As a result, $\{C_\theta^+ : \theta \in \Theta_n\}$ has VC-index at most $\text{VC}(\Pi_n)+1$. Similarly, $\{C_\theta^- : \theta \in \Theta_n\}$ has VC-index at most $\text{VC}(\Pi_n)+1$. Therefore, $\{g_{0,\theta} : \theta \in \Theta_n\}$ is VC with index $2\text{VC}(\Pi_n)+1$.

Using the similar argument, one has $\text{VC}\left(\mathcal{G}_{j,\theta}\right) \leq 2\text{VC}(\Pi_n)+2$ for $1 \leq j \leq 5$. The result for $\text{VC}(\mathcal{G}_{j,\Theta_n}^-)$ follows from Theorem 2.6.18 in van der Vaart and Wellner (1998). □

Lemma G.2 (Theorem 3 in Andrews (1994))**.** Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two function classes with envelope functions F_1 and F_2, respectively. If we set

$$\mathcal{F}_1 \oplus \mathcal{F}_2 \equiv \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$$
$$\mathcal{F}_1 \otimes \mathcal{F}_2 \equiv \{f_1 \cdot f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\},$$

then $\mathcal{F}_1 \oplus \mathcal{F}_2$ and $\mathcal{F}_1 \otimes \mathcal{F}_2$ admit envelope functions $F_1 + F_2$ and $F_1 \cdot F_2$, respectively. Their covering number are upper bounded as

$$N\left(\epsilon\|F_1+F_2\|_{Q,2}, \mathcal{F}_1 \oplus \mathcal{F}_2, L^2(Q)\right) \leq N\left(\epsilon\|F_1\|_{Q,2}, \mathcal{F}_1, L^2(Q)\right) N\left(\epsilon\|F_1\|_{Q,2}/2, \mathcal{F}_1, L^2(Q)\right),$$
$$\sup_Q N\left(\epsilon\|F_1F_2\|_{Q,2}/2, \mathcal{F}_1 \otimes \mathcal{F}_2, L^2(Q)\right) \leq \left[\sup_Q N\left(\epsilon\|F_1\|_{Q,2}, \mathcal{F}_1, L^2(Q)\right)\right]\left[\sup_Q N\left(\epsilon\|F_2\|_{Q,2}, \mathcal{F}_2, L^2(Q)\right)\right].$$

H Algorithm for Welfare Optimization, Estimation and Inference

Algorithm 1 Welfare Optimization, estimation and inference of with cross-fitting

Input: Level $\alpha \in (0,1)$, estimators $\widehat{e}$, $\widehat{\mu}_1$, and $\widehat{\mu}_0$, and a K-fold random partition of the dataset $\{(X_i, Y_i, A_i)\}_{i=1}^n$, denoted as $\cup_{k=1}^K \mathcal{I}_k$, where $|\mathcal{I}_k| = n/K$.
Run simulated annealing to find $\widehat{\pi}_n, \widehat{\eta}_n$ that maximize the mean of the doubly robust scores $\Gamma_i := g_\theta\left(Z_i; \widehat{\mu}_i, \widehat{e}_i\right)$ and report $\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$ and its CI, where for a given (π, η),
for $k \in [K]$ **do**
  Using $\{(X_i, Y_i, A_i)\}_{i \in \mathcal{I}_k^c}$ and pseudo-outcome $\check{Y}_i(\eta) = (Y_i - \eta)_-$, construct
  $\widehat{e}^{-k(i)}(x)$ with $\{(X_i, A_i) : i \in \mathcal{I}_k^c\}$,
  $\widehat{\mu}_1^{-k(i)}(x, \eta)$ with $\{(X_i, \check{Y}_i(\eta), A_i) : i \in \mathcal{I}_k^c \wedge A_i = 1\}$, and
  $\widehat{\mu}_0^{-k(i)}(x, \eta)$ with $\{(X_i, \check{Y}_i(\eta), A_i) : i \in \mathcal{I}_k^c \wedge A_i = 0\}$.
  for $i \in \mathcal{I}_k$ **do**
    Evaluate $\widehat{e}_i := \widehat{e}^{-k(i)}(X_i), \widehat{\mu}_{1,i} := \widehat{\mu}_1^{-k(i)}(X_i, \eta)$, $\widehat{\mu}_{0,i} := \widehat{\mu}_0^{-k(i)}(X_i, \eta)$, and compute
    the doubly robust score $\Gamma_i = g_\theta\left(Z_i; \widehat{\mu}_i, \widehat{e}_i\right)$.
  end for
end for
Return $\widehat{\pi}_n$, $\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n) = \frac{1}{n}\sum_{i=1}^n \Gamma_i$, and $\left[\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n) \pm \Phi^{-1}((1+\gamma)/2)\widehat{\text{se}}\right]$ as γ-CI, where $\widehat{\text{se}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^n \left(\Gamma_i - \widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)\right)^2}$.

I Empirical Application and Simulation Studies: Supplementary Materials

This section provides additional details for the empirical analysis of the JTPA Study in Section 6.1 and for the simulations based on WGAN-JTPA in Section 6.2. In addition, we present results from two further simulation studies, using DGPs similar to those in Athey and Wager (2021) with some modifications.

I.1 Additional Results from the JTPA Study

This subsection complements Section 6.1. Expressions for the optimal policies under different combinations of $\alpha \in \mathcal{A}$ and policy class are organized in Table 5. We normalize the policy coefficient associated with *prevearn* to have an absolute value of 1.

Based on the welfare point estimates in Tables 2 and 3, Tables 6 and 7 compute the percentage losses in welfare as we switch between the optimal policy targeting an α of interest to policies targeting other levels of α'. We highlight the diagonal entries as these policies are targeting the actual subpopulations of focus, therefore having zero loss in welfare (as compared to themselves). Larger welfare losses tend to appear when the actual α and the α' for policy selection differ more. $\alpha = 0.25$ is particularly vulnerable if the policy is instead targeting some $\alpha' \geq 0.4$.

	Linear	**Linear with edu^2 and edu^3**
	$\alpha_0 = 0.25$	
Optimal policy	$\mathbb{1}[-6371.583 + 634.221edu - prevearn > 0]$	$\mathbb{1}[-18085.19 + 2272.77edu - 24.88edu^2 - 2.52edu^3 - prevearn > 0]$
% treated	34.761%	32.896%
	$\alpha_0 = 0.3$	
Optimal policy	$\mathbb{1}[3163.752 - 123.104edu - prevearn > 0]$	$\mathbb{1}[-17881.079 + 2235.937edu - 22.299edu^2 - 2.598edu^3 - prevearn > 0]$
% treated	50.992%	32.820%
	$\alpha_0 = 0.4$	
Optimal policy	$\mathbb{1}[-16400.524 + 2069.530edu - prevearn > 0]$	$\mathbb{1}[-10421.477 + 943.370edu + 41.482edu^2 + 0.795edu^3 - prevearn > 0]$
% treated	82.392%	81.969%
	$\alpha_0 = 0.5$	
Optimal policy	$\mathbb{1}[-13704.005 + 1825.869edu - prevearn > 0]$	$\mathbb{1}[-15844.957 + 2096.331edu + 9.463edu^2 - 1.361edu^3 - prevearn > 0]$
% treated	83.400%	83.379%
	$\alpha_0 = 0.8$	
Optimal policy	$\mathbb{1}[3849.726 + 333.043edu - prevearn > 0]$	$\mathbb{1}[-871.769 + 1532.005edu - 65.590edu^2 - 1.093edu^3 - prevearn > 0]$
% treated	86.783%	79.204%

Table 5: Optimal policies under different combinations of α and policy class.

α of Interest \ α' for Policy Selection	0.25	0.3	0.4	0.5	0.8
0.25	0.00%	1.04%	5.61%	6.67%	11.90%
0.3	2.08%	0.00%	0.99%	2.30%	6.06%
0.4	4.60%	0.86%	0.00%	0.15%	2.23%
0.5	5.49%	1.12%	0.07%	0.00%	0.89%
0.8	5.33%	2.18%	1.13%	0.76%	0.00%

Table 6: Percentage welfare loss for every combination of actual α and α' for policy selection, relative to implementing the optimal linear policy targeting the worst-affected $(\alpha \times 100)\%$.

α of Interest \ α' for Policy Selection	0.25	0.3	0.4	0.5	0.8
0.25	0.00%	0.53%	7.73%	9.09%	12.86%
0.3	0.11%	0.00%	0.77%	2.28%	5.02%
0.4	3.29%	3.20%	0.00%	0.15%	1.71%
0.5	4.60%	4.47%	0.04%	0.00%	0.49%
0.8	5.09%	5.15%	1.39%	0.95%	0.00%

Table 7: Percentage welfare loss for every combination of actual α and α' for policy selection, relative to implementing the optimal linear policy with edu^2 and edu^3 targeting the worst-affected $(\alpha \times 100)\%$.

I.2 Simulations Using the WGAN-JTPA Superpopulation Data: Details

We employ the `wgan` package in `Python` developed by Athey et al. (2024) to construct an artificial superpopulation that closely mimics the JTPA data in Bloom et al. (1997). Following the instructions in Athey et al. (2024), we first generate the covariates conditional on the

treatment status, i.e., $(edu, prevearn)|A$, then generate the outcome conditional on both the treatment status and the covariates, i.e., $earnings|(edu, prevearn, A)$. We set a constraint that *earnings* and *prevearn* are lower bounded by 0, and since *edu* takes integer values between 7 and 18, we set it to be a categorical variable. In the training step where neural networks are utilized, we set the batch size to 4,096, the maximum number of training epochs to 1,000 and the learning rate for both the generator and the critic to 0.001. To obtain the population counterfactuals, the generator for $earnings|(edu, prevearn, A)$ is re-applied on $(edu, prevearn, 1-A)$. Table 8 presents summary statistics for WGAN-JTPA, and Figures 5 and 6 display graphical comparisons between the JTPA and WGAN-JTPA data.

	$\boldsymbol{A=0}$ (33.503% of WGAN-JTPA)		$\boldsymbol{A=1}$ (66.497% of WGAN-JTPA)	
	mean	s.d.	mean	s.d.
earnings	13647.5	12227.77	14648.81	12904.37
edu	11.48	1.55	11.50	1.63
prevearn	2657.61	3678.91	2695.75	3709.31

Table 8: Summary statistics for WGAN-JTPA.

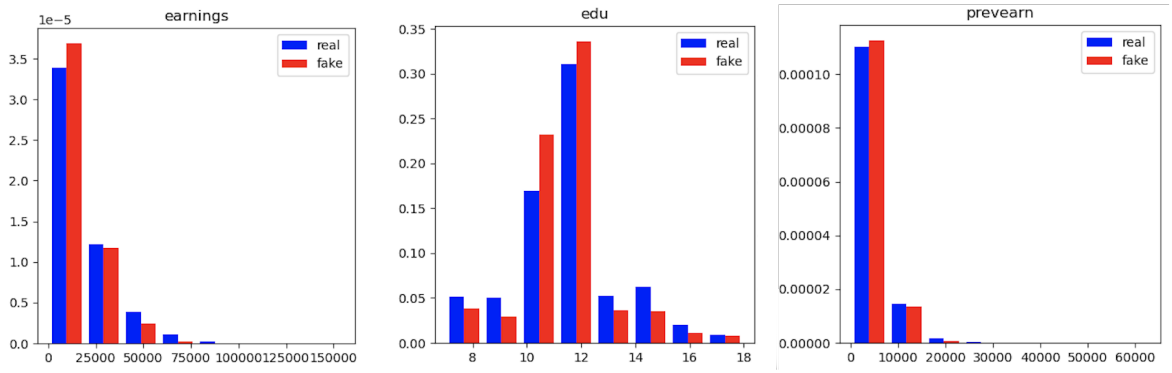

Figure 5: Marginal histograms for JTPA and WGAN-JTPA data.

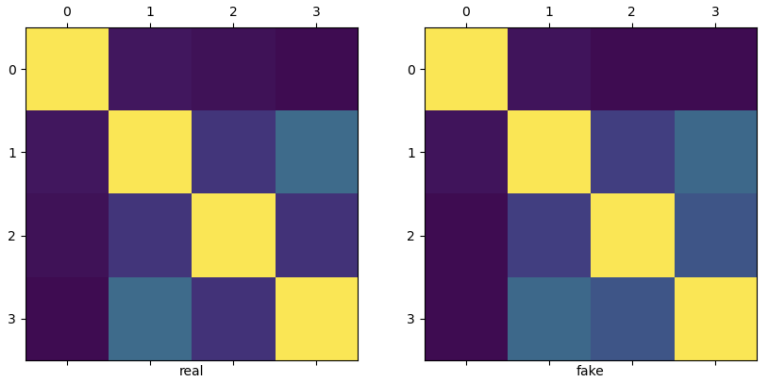

Figure 6: Between-variable correlations for JTPA and WGAN-JTPA data.

At the population level, Table 9 echoes Tables 2 and 3 in the main text by evaluating the α-expected welfare under policies targeting different α''s. $\alpha' = 1$ is equivalent to a mean-optimal policy. Similar to the JTPA estimation results in Section 6.1, there are notable changes in welfare across α's, indicating a potential risk of welfare impairment for the most disadvantaged when implementing a policy that targets the population mean, or a large α' in general.

α of Interest \ α' for Policy Section	0.25	0.3	0.4	0.5	0.8	1
0.25	1119.195	1119.145	1119.145	1119.145	1044.962	1029.394
0.3	1908.118	1908.135	1908.135	1908.135	1827.813	1808.950
0.4	3460.527	3460.773	3460.773	3460.773	3385.985	3365.862
0.5	4866.580	4867.556	4867.556	4867.556	4810.727	4792.908
0.8	9323.006	9328.851	9328.851	9328.851	9475.336	9472.923
1	14346.024	14351.932	14351.932	14351.932	14638.593	14643.594

Table 9: $\mathbb{W}_\alpha(\pi_o)$ for every combination of actual α and α' for policy selection using WGAN-JTPA. All values are in USD.

I.3 Two Simulation Studies Based on DGPs in Athey and Wager (2021)

Section 5.2 of Athey and Wager (2021) uses simulated data to exhibit the welfare improvements of their learned policies, which optimize the population mean outcome. We emulate their specifications of the outcome and CATE, while making treatment exogenous with a known propensity score 2/3. Below are our DGPs, with $n \in \{300, 500, 1000, 1500\}$:

$$X \sim N(0, I_{4\times 4}),\ \epsilon|X \sim N(0,1),\ A \sim \text{Bernoulli}(2/3),\ Y = 10 + (X_3 + X_4)_+ + A\tau(X) + \epsilon,$$

where $\tau(\cdot)$ has two specifications:

$$\tau(X) = \left((X_1)_+ + (X_2)_+ - 1\right)/2, \text{ or} \tag{I.1}$$
$$\tau(X) = \text{sign}(X_1 X_2)/2. \tag{I.2}$$

We construct two size-one-million superpopulations, one for each specification of $\tau(\cdot)$, and we restrict the policy class to linear rules of the form

$$\Pi_{\text{LES}} := \left\{\{x : \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 > 0\},\ (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \in \mathbb{R}^5\right\}.$$

Since we can generate Y_i for both $A_i = 0$ and $A_i = 1$, we have full knowledge of the true outcome distribution induced by any $\pi \in \Pi_{\text{LES}}$. To obtain the population truth for each $\alpha \in \mathcal{A}$ and specification of $\tau(\cdot)$, we run SA to select a $\pi_o \in \Pi_{\text{LES}}$ that maximizes the α-AVaR of the outcome distribution and take the resulting maximum.

At the population level, Tables 10 and 11 present the α-expected welfare under different α'-EWM policies. In Table 10, the changes in welfare across columns are noticeably small, which implies that different targeting policies generally have minimal impact on the welfare

of the disadvantaged subpopulation when $\tau(\cdot)$ is specified as (I.1). Table 11 shows slightly greater changes in welfare across columns, when $\tau(\cdot)$ is specified as (I.2).

α of Interest \ α' for Policy Selection	0.25	0.3	0.4	0.5	0.8	1
0.25	9.09461	9.09461	9.09461	9.09461	9.09457	9.09457
0.3	9.22146	9.22146	9.22146	9.22146	9.22142	9.22142
0.4	9.44289	9.44289	9.44289	9.44289	9.44287	9.44287
0.5	9.63925	9.63925	9.63925	9.63925	9.63925	9.63925
0.8	10.18965	10.18965	10.18965	10.18965	10.18967	10.18967
1	10.67678	10.67678	10.67678	10.67678	10.67682	10.67682

Table 10: $\mathbb{W}_\alpha(\pi_o)$ for every combination of actual α and α' for policy selection using the DGP in Section I.3; τ is specified as (I.1) and the superpopulation size is one million.

α of Interest \ α' for Policy Selection	0.25	0.3	0.4	0.5	0.8	1
0.25	9.04758	9.04754	9.04749	9.04521	9.04414	8.97402
0.3	9.17424	9.17431	9.17430	9.17252	9.17163	9.10875
0.4	9.39510	9.39533	9.39537	9.39463	9.39403	9.34376
0.5	9.59071	9.59105	9.59113	9.59143	9.59110	9.55145
0.8	10.13745	10.13808	10.13820	10.14084	10.14116	10.12731
1	10.61981	10.62073	10.62059	10.62466	10.62532	10.62593

Table 11: $\mathbb{W}_\alpha(\pi_o)$ for every combination of actual α and α' for policy selection using the DGP in Section I.3; τ is specified as (I.2) and the superpopulation size is one million.

Similar to Figure 4 in the main text, we plot the between-quantile differences in post-treatment outcomes to compare the 0.25-EWM policy with the 1-EWM and equality-minded policies. Figure 7 corresponds to $\tau(\cdot)$ as (I.1), and Figure 8 corresponds to $\tau(\cdot)$ as (I.2). Interestingly, in Figure 7, the equality-minded optimal policy is identical to the 1-EWM policy. In contrast, Figure 8 shows that the 0.25-EWM and equality-minded policies both enhance the welfare of lower-ranked observations while reducing the welfare of higher-ranked observations in comparison to the 1-EWM policy, with the 0.25-EWM policy focusing more on these adjustments. In Figure 7, such changes made by the 0.25-EWM policy are smaller in magnitude and more volatile.

For each $\tau(\cdot)$, we run Algorithm 1 with $K = 2$ on 1,000 random samples, each drawn without replacement from the corresponding superpopulation, for every combination of α and n. μ_1 and μ_0 are estimated using random forests with default tuning parameters. As demonstrated by the simulation results in Tables 12 and 13, our debiased estimator $\widehat{\mathbb{W}}_\alpha(\widehat{\pi}_n)$ performs satisfactorily even when n is as small as 500.

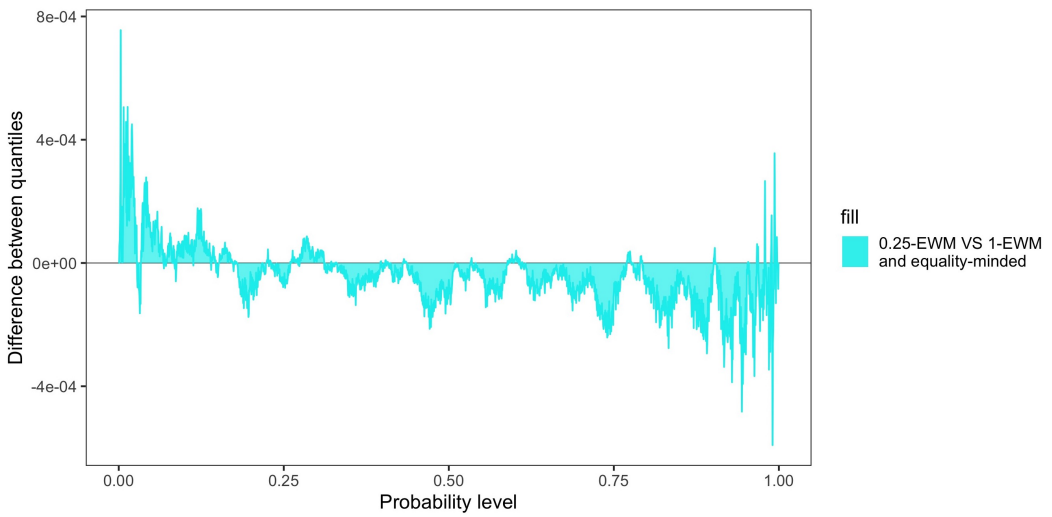

Figure 7: Between-quantile differences in outcomes for the 0.25-EWM, 1-EWM, and equality-minded policies using the DGP in Section I.3; τ is specified as (I.1) and the superpopulation size is one million.

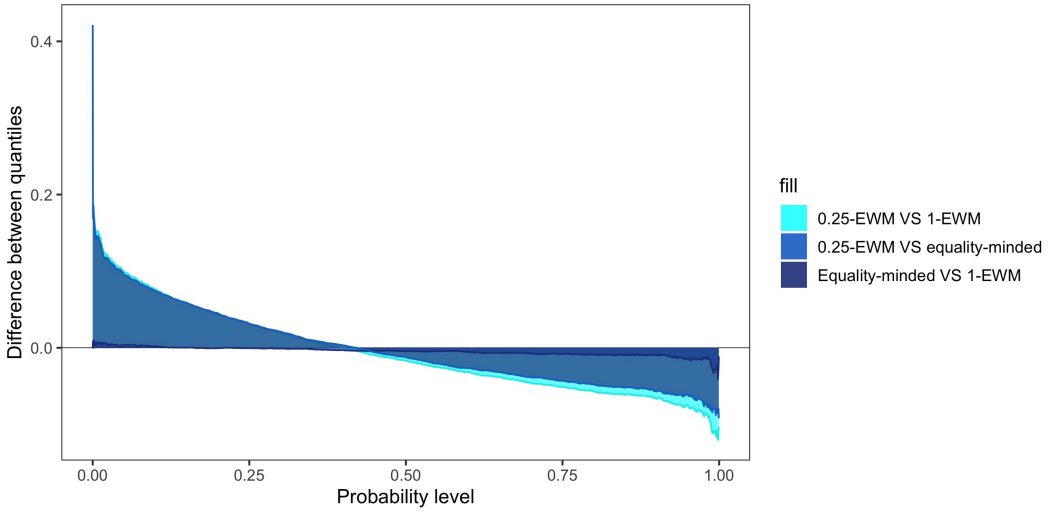

Figure 8: Between-quantile differences in outcomes for the 0.25-EWM, 1-EWM, and equality-minded policies using the DGP in Section I.3; τ is specified as (I.2) and the superpopulation size is one million.

Sample size	**300**	**500**	**1,000**	**1,500**
Panel 1: $\alpha = 0.25$, truth = 9.095				
Avg. % treated using $\widehat{\pi}_n$	33.041%	31.500%	35.254%	34.629%
Bias	0.016	−0.016	−0.012	−0.026
Var	0.017	0.010	0.005	0.004
MSE	0.017	0.010	0.005	0.004
95% Coverage	92.5%	94.7%	95.5%	94.6%
Panel 2: $\alpha = 0.3$, truth = 9.221				
Avg. % treated using $\widehat{\pi}_n$	34.131%	31.689%	35.655%	34.373%
Bias	0.012	−0.020	−0.017	−0.022
Var	0.015	0.010	0.005	0.003
MSE	0.015	0.010	0.005	0.004
95% Coverage	92.6%	93.7%	94.9%	94.4%
Panel 3: $\alpha = 0.4$, truth = 9.443				
Avg. % treated using $\widehat{\pi}_n$	35.185%	32.422%	35.103%	33.843%
Bias	0.011	−0.019	−0.018	−0.016
Var	0.013	0.009	0.004	0.003
MSE	0.013	0.009	0.004	0.004
95% Coverage	95.4%	94.1%	95.7%	93.9%
Panel 4: $\alpha = 0.5$, truth = 9.639				
Avg. % treated using $\widehat{\pi}_n$	34.373%	31.907%	35.307%	35.474%
Bias	0.011	−0.022	−0.021	−0.016
Var	0.012	0.008	0.004	0.003
MSE	0.012	0.008	0.004	0.003
95% Coverage	94.7%	94.9%	94.3%	94.1%
Panel 5: $\alpha = 0.8$, truth = 10.190				
Avg. % treated using $\widehat{\pi}_n$	38.436%	36.471%	36.443%	35.965%
Bias	0.007	−0.019	−0.015	−0.022
Var	0.011	0.007	0.003	0.002
MSE	0.011	0.007	0.003	0.003
95% Coverage	96.3%	94.3%	94.5%	94.2%

Table 12: Simulation results based on the DGP in Appendix I.3 (1,000 replications); τ is specified as (I.1).

Sample size	**300**	**500**	**1,000**	**1,500**
Panel 1: $\alpha = 0.25$, truth = 9.048				
Avg. % treated using $\widehat{\pi}_n$	33.613%	30.019%	21.031%	24.145%
Bias	0.058	0.028	−0.012	−0.009
Var	0.015	0.010	0.005	0.004
MSE	0.018	0.011	0.005	0.004
95% Coverage	91.7%	93.4%	95.2%	94.4%
Panel 2: $\alpha = 0.3$, truth = 9.174				
Avg. % treated using $\widehat{\pi}_n$	33.831%	31.193%	22.572%	24.439%
Bias	0.057	0.025	−0.007	−0.011
Var	0.013	0.009	0.005	0.003
MSE	0.016	0.009	0.005	0.003
95% Coverage	92.6%	94.0%	95.7%	96.7%
Panel 3: $\alpha = 0.4$, truth = 9.395				
Avg. % treated using $\widehat{\pi}_n$	37.306%	33.988%	22.097%	28.648%
Bias	0.055	0.020	−0.013	−0.014
Var	0.011	0.007	0.004	0.003
MSE	0.014	0.008	0.004	0.003
95% Coverage	93.2%	94.4%	95.6%	95.7%
Panel 4: $\alpha = 0.5$, truth = 9.591				
Avg. % treated using $\widehat{\pi}_n$	37.571%	34.613%	27.974%	30.885%
Bias	0.056	0.019	−0.015	−0.012
Var	0.011	0.007	0.003	0.002
MSE	0.014	0.007	0.004	0.003
95% Coverage	92.1%	95.4%	96.8%	95.9%
Panel 5: $\alpha = 0.8$, truth = 10.141				
Avg. % treated using $\widehat{\pi}_n$	44.033%	43.809%	38.756%	39.753%
Bias	0.067	0.030	−0.007	−0.012
Var	0.009	0.005	0.003	0.002
MSE	0.013	0.006	0.003	0.002
95% Coverage	94.3%	96.8%	96.8%	95.1%

Table 13: Simulation results based on the DGP in Appendix I.3 (1,000 replications); τ is specified as (I.2).