Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice[*]

Toru Kitagawa[†] and Aleksey Tetenov[‡]

First Version: March 9, 2015
This Version: November 27, 2017

Abstract

One of the main objectives of empirical analysis of experiments and quasi-experiments is to inform policy decisions that determine the allocation of treatments to individuals with different observable covariates. We study the properties and implementation of the *Empirical Welfare Maximization (EWM) method,* which estimates a treatment assignment policy by maximizing the sample analog of average social welfare over a class of candidate treatment policies. The EWM approach is attractive in terms of both statistical performance and practical implementation in realistic settings of policy design. Common features of these settings include: (i) feasible treatment assignment rules are constrained exogenously for ethical, legislative, or political reasons, (ii) a policy maker wants a simple treatment assignment rule based on one or more eligibility scores in order to reduce the dimensionality of individual observable characteristics, and/or (iii) the proportion of individuals who can receive the treatment is *a priori* limited due to a budget or a capacity constraint. We show that when the propensity score is known, the average social welfare attained by EWM rules converges at least at $n^{-1/2}$ rate to the maximum obtainable welfare uniformly over a minimally constrained class of data distributions, and this uniform convergence rate is minimax optimal. We examine how the uniform convergence rate depends on the richness of the class of candidate decision rules, the distribution of conditional treatment effects, and the lack of knowledge of the propensity score. We offer easily implementable algorithms for computing the EWM rule and an application using experimental data from the National JTPA Study.

[*] We would like to thank Susan Athey, Kei Hirano, Guido Imbens, Shin Kanaya, Charles Manski, Joerg Stoye, and seminar participants at Academia Sinica, Bern, Brown, BC, BU, Cambridge, CEMFI, CORE, Duke, GRIPS, JHU, NYU, UPenn, UPF, UVA and Warwick for beneficial comments. We would also like to thank the participants at 2014 Bristol Econometric Study Group, 2014 SETA Conference, Hakone Microeconometrics Conference, and ICEEE 2015 for beneficial comments and discussions. Financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) and the European Research Council (Starting grant No. 715940) is gratefully acknowledged.

[†] Cemmap/University College London, Department of Economics. Email: t.kitagawa@ucl.ac.uk

[‡] University of Bristol, Email: a.tetenov@bristol.ac.uk

1 Introduction

Treatment effects often vary with observable individual characteristics. An important objective of empirical analysis of experimental and quasi-experimental data is to determine the individuals who should be treated based on their observable characteristics. Empirical researchers often use regression estimates of individual treatment effects to infer the set of individuals who benefit or do not benefit from the treatment and to suggest who should be targeted for treatment. This paper advocates the *Empirical Welfare Maximization* (EWM) method, which offers an alternative way to choose optimal treatment assignment based on experimental or observational data from program evaluation studies. We study the frequentist properties of the EWM treatment choice rule and show its optimality in terms of welfare convergence rate, which measures how quickly the average welfare attained by practicing the estimated treatment choice rule converges to the maximal welfare attainable with the knowledge of the true data generating process. We also argue that the EWM approach is well-suited for policy design problems, since it easily accommodates many practical policy concerns, including (i) feasible treatment assignment rules being constrained exogenously for ethical, legislative, or political reasons, (ii) the policy maker facing a budget or capacity constraint that limits the proportion of individuals who can receive one of the treatments, or (iii) the policy maker wanting to have a simple treatment assignment rule based on one or more indices (eligibility scores) to reduce the dimensionality of individual characteristics.

Let the data be a size n random sample of $Z_i = (Y_i, D_i, X_i)$, where $X_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ refers to observable pre-treatment covariates of individual i, $D_i \in \{0,1\}$ is a binary indicator of the individual's treatment assignment, and $Y_i \in \mathbb{R}$ is her/his post-treatment observed outcome. The population from which the sample is drawn is characterized by P, a joint distribution of $(Y_{0,i}, Y_{1,i}, D_i, X_i)$, where $Y_{0,i}$ and $Y_{1,i}$ are potential outcomes that would have been observed if i's treatment status were $D_i = 0$ and $D_i = 1$, respectively. We assume *unconfoundedness,* meaning that in the data treatments are assigned independently of the potential outcomes $(Y_{0,i}, Y_{1,i})$ conditionally on observable characteristics X_i. Based on this data, the policy-maker has to choose a conditional treatment rule that determines whether individuals with covariates X in a target population will be assigned to treatment 0 or to treatment 1. We restrict our analysis to non-randomized treatment rules. The set of treatment rules could then be indexed by their *decision sets* $G \subset \mathcal{X}$ of covariate values, which determine the group of individuals $\{X \in G\}$ to whom treatment 1 is assigned. We denote the collection of candidate treatment rules by $\mathcal{G} = \{G \subset \mathcal{X}\}$.

The goal of our analysis is to empirically select a treatment assignment rule that gives the highest *welfare* to the target population. We assume that the joint distribution of $(Y_{0,i}, Y_{1,i}, X_i)$

of the target population is identical to that of the sampled population.[1] We consider the *additive* welfare criterion defined by the average of the individual outcomes in the target population.[2] When treatment rule G is applied to the target population, the social welfare defined by the sum of individual outcomes in the population is proportional to

$$W(G) \equiv E_P \left[Y_1 \cdot 1\{X \in G\} + Y_0 \cdot 1\{X \notin G\} \right] \tag{1.1}$$

where $E_P(\cdot)$ is the expectation with respect to P. Our framework could incorporate a broad range of social preferences by suitably redefining the outcome variable. Setting Y to be a concave transformation of one's measure of wealth leads to an inequality-averse social welfare of Atkinson (1970). When multiple outcome variables enter into the individual utility (e.g., consumption and leisure), Y can be set to a known function of these outcomes. The cost of treatment can be incorporated into the social welfare by redefining the individual potential outcome Y_d to be the outcome minus the cost of treatment d.

Denoting the conditional mean treatment response by $m_d(x) \equiv E[Y_d|X = x]$ and the conditional average treatment effect by $\tau(x) \equiv m_1(x) - m_0(x)$, we could also express the welfare criterion as

$$W(G) = E_P(m_0(X)) + E_P\left[\tau(X) \cdot 1\{X \in G\}\right]. \tag{1.2}$$

Assuming unconfoundedness, equivalence of the distributions of $(Y_{0,i}, Y_{1,i}, X_i)$ between the target and sampled populations, and the overlap condition for the propensity score $e(X) = E_P[D|X]$ in the sampled population, the welfare criterion (1.1) can be written equivalently as

$$\begin{aligned} W(G) &= E_P\left[\frac{YD}{e(X)} \cdot 1\{X \in G\} + \frac{Y(1-D)}{1-e(X)} \cdot 1\{X \notin G\}\right] \qquad (1.3) \\ &= E_P(Y_0) + E_P\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right) \cdot 1\{X \in G\}\right]. \end{aligned}$$

Hence, if the probability distribution of observables (Y, D, X) is fully known to the decision-maker, an optimal treatment rule from the utilitarian perspective can be written as

$$G^* \in \arg\max_{G \in \mathcal{G}} W(G). \tag{1.4}$$

Or, equivalently, as a maximizer of the welfare gain relative to $E_P(Y_0)$:

$$G^* \in \arg\max_{G \in \mathcal{G}} E_P\left[\tau(X) \cdot 1\{X \in G\}\right], \text{ or} \tag{1.5}$$

$$G^* \in \arg\max_{G \in \mathcal{G}} E_P\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right) \cdot 1\{X \in G\}\right]. \tag{1.6}$$

[1] In Remark 2.2, we consider a setting where the target and the sampled populations have identical conditional treatment effects, but different marginal distributions of X.

[2] In the econometrics literature of treatment choice, the additive social welfare is often referred to as a utilitarian social welfare.

The main idea of *Empirical Welfare Maximization (EWM)* is to solve a sample analog of the population maximization problem (1.4),

$$\begin{aligned}\hat{G}_{EWM} &\in \arg\max_{G\in\mathcal{G}} W_n(G), && (1.7)\\ \text{where } W_n(G) &= E_n\left[\frac{Y_iD_i}{e(X_i)}\cdot 1\left\{X_i\in G\right\} + \frac{Y_i(1-D_i)}{1-e(X_i)}\cdot 1\left\{X_i\notin G\right\}\right]\end{aligned}$$

and $E_n(\cdot)$ is the sample average. One notable feature of our framework is that the class of candidate treatment rules $\mathcal{G}=\{G\subset\mathcal{X}\}$ is not as rich as the class of all subsets of $\mathcal{X}$, and it may not include the *first-best decision set*

$$G^*_{FB}\equiv\left\{x\in\mathcal{X}:\tau(x)\geq 0\right\}, \tag{1.8}$$

which maximizes the population welfare (1.1) if any assignment rule is feasible to implement. Our framework with a constrained class of feasible assignment rules allows us to incorporate several types of exogenous constraints that generally restrict the complexity of feasible treatment assignment rules. For instance, when assigning treatments to individuals in the target population, it may not be realistic to implement a complex treatment assignment rule due to logistic, legal, ethical, or political restrictions.

The largest welfare that could be obtained by any treatment rule in class $\mathcal{G}$ is

$$W^*_{\mathcal{G}}\equiv\sup_{G\in\mathcal{G}} W(G), \tag{1.9}$$

which is the *second-best* welfare if $W^*_{\mathcal{G}} < W(G^*_{FB})$. In line with Manski (2004) and the subsequent literature on statistical treatment rules, we evaluate the performance of estimated treatment rules $\hat{G}\in\mathcal{G}$ in terms of their average welfare loss (regret) relative to the maximum feasible welfare $W^*_{\mathcal{G}}$

$$W^*_{\mathcal{G}} - E_{P^n}\left[W(\hat{G})\right] = E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G})\right]\geq 0, \tag{1.10}$$

where the expectation E_{P^n} is taken over different realizations of the random sample. This criterion measures the average difference between the best attainable population welfare and the welfare attained by implementing estimated policy $\hat{G}$. Since we assess the statistical performance of $\hat{G}$ by its welfare value $W(\hat{G})$, we do not require $\arg\max_{G\in\mathcal{G}} W(G)$ to be unique or $\hat{G}$ to converge to a specific set.

Assuming that the propensity score $e(X)$ is known and bounded away from zero and one, as is the case in randomized experiments, we derive a non-asymptotic distribution-free upper bound of $E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G}_{EWM})\right]$ as a function of sample size n and a measure of complexity of $\mathcal{G}$. Based on this bound, we show that the average welfare of the EWM treatment rule converges to $W^*_{\mathcal{G}}$

at rate $O(n^{-1/2})$ uniformly over a minimally constrained class of probability distributions. We also show that this uniform convergence rate of $\hat{G}_{EWM}$ is optimal in the sense that no estimated treatment choice rule of any kind can attain a faster uniform convergence rate compared to the EWM rule, i.e., minimax rate optimality of $\hat{G}_{EWM}$. For further refinement of this theoretical result, we analyze how this uniform convergence rate improves if the first-best decision rule G^*_{FB} is feasible, i.e., $G^*_{FB} \in \mathcal{G}$, and if the class of data generating processes is constrained by the *margin assumption,* which restricts the distribution of conditional treatment effects in a neighborhood of zero. We show that $\hat{G}_{EWM}$ remains minimax rate optimal with these additional restrictions.

When the data are from an observational study, the propensity score is usually unknown, so it is not feasible to implement the EWM rule (1.7). As a feasible version of the EWM rule, we consider *hybrid EWM* approaches that plug in estimators of the regression equations or the propensity score in the sample analogs of (1.5) or (1.6). Specifically, with estimated regression functions $\hat{m}_d(x) = \hat{E}(Y_d|X = x) = \hat{E}(Y|X = x, D = d)$, we define the *m-hybrid* rule as

$$\hat{G}_{m-hybrid} \in \arg\max_{G\in\mathcal{G}} E_n\left[\hat{\tau}^m\left(X_i\right) \cdot 1\left\{X_i \in G\right\}\right], \tag{1.11}$$

where $\hat{\tau}^m\left(X_i\right) \equiv \hat{m}_1\left(X_i\right) - \hat{m}_0\left(X_i\right)$. Similarly, with the estimated propensity score $\hat{e}(x)$, we define an *e-hybrid* rule as

$$\hat{G}_{e-hybrid} \in \arg\max_{G\in\mathcal{G}} E_n\left[\hat{\tau}^e_i \cdot 1\left\{X_i \in G\right\}\right], \tag{1.12}$$

where $\hat{\tau}^e_i \equiv \left[\frac{Y_i D_i}{\hat{e}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{e}(X_i)}\right] \cdot 1\left\{\varepsilon_n \leq \hat{e}\left(X_i\right) \leq 1 - \varepsilon_n\right\}$ with a converging positive sequence $\varepsilon_n \to 0$ as $n \to \infty$. We investigate the performance of these hybrid approaches in terms of the uniform convergence rate of the welfare loss and clarify how this rate is affected by the estimation uncertainty in $\hat{m}_d(\cdot)$ and $\hat{e}(\cdot)$.

Since the welfare criterion function involves optimization over a class of sets, estimation of the EWM and hybrid treatment rules could present challenging computational problems when $\mathcal{G}$ is rich, similarly to the maximum score estimation (Manski (1975), Manski and Thompson (1989)). We argue, however, that exact maximization of EWM criterion is now practically feasible for many problems in economics using widely-available optimization software and an approach proposed by Florios and Skouras (2008), which we extend and improve upon.

To illustrate EWM in practice, we compare EWM and plug-in treatment rules computed from the experimental data of the National Job Training Partnership Act Study analyzed by Bloom et al. (1997).

1.1 Related Literature

Our paper contributes to a growing literature on statistical treatment rules in econometrics, including Manski (2004), Dehejia (2005), Hirano and Porter (2009), Stoye (2009, 2012), Chamberlain (2011), Bhattacharya and Dupas (2012), Tetenov (2012), and Kasy (2017). Manski (2004) proposes to assess the welfare properties of statistical treatment rules by their maximum regret and derives finite-sample regret bounds for *Conditional Empirical Success (CES)* rules. CES rules take a finite partition of the covariate space and, separately for each set in this partition, assign the treatment that yields the highest sample average outcome. CES rules can be viewed as a type of EWM rules for which $\mathcal{G}$ consists of all unions of the sets in the partition and the empirical welfare criterion uses the sample propensity score. Manski shows that with the partition fixed, their welfare regret converges to zero at least at $n^{-1/2}$ rate. We show that this rate holds for a broader class of EWM rules and that it cannot be improved uniformly without additional restrictions on P.

Stoye (2009) shows that in the absence of ex-ante restrictions on how outcome distributions vary with covariates, finite-sample minimax regret is attained by rules that take the finest partition of the covariate space and operate independently for each covariate value. This important result implies that with continuous covariates, minimax regret does not converge to zero with sample size because the first-best treatment rule may be arbitrarily "wiggly" and difficult to approximate from countable data. Our approach does not give rise to Stoye's non-convergence result because we restrict the complexity of $\mathcal{G}$ and define regret relative to the maximum attainable welfare in $\mathcal{G}$ instead of the unconstrained first-best welfare. However, we do not derive exact finite-sample minimax regret rules in the more complex setting of our paper.

Treatment choice has substantial similarities with *classification.* In a binary classification problem, the researcher observes a random sample (Y_i, X_i), where $Y_i \in \{-1, +1\}$ denotes which of two classes an observation belongs to. A *classifier* aims to predict the correct classification of future observations based on observed covariates X. A treatment rule similarly seeks to "classify" individuals into those who should and should not be treated based on their covariates. Treatment choice, however, differs from classification in a few significant ways: (1) observed outcomes can be real-valued rather than binary, (2) we only observe one of the two potential outcomes, and not the correct classification of individuals in the sample, (3) policy settings often impose constraints on practicable treatment rules or on the proportion of the population that could be treated.

The earliest works noting the connection between treatment choice and classification are Zadrozny (2003) and Beygelzimer and Langford (2009). They propose algorithms that transform a sample from a treatment choice problem into a sample from a standard binary classification problem. Treatment rules could be then generated using any existing binary classification algorithm. Beygelzimer and Langford (2009) show that the welfare loss of any treatment rule G with respect to the first-best $(W(G^*_{FB}) - W(G))$ is bounded above by a classification regret with respect to the first-best classifier. It implies that if the regret of the classification algorithm converges to zero, then consistency of $W(\hat{G})$ to $W(G^*_{FB})$ holds. They do not consider any restrictions on $\mathcal{G}$ and do not study the welfare loss convergence rates. Instead, we consider maximizing $W_n(G)$ over a constrained class of policies without converting it into a classification problem.

The idea of optimizing the sample analog of a population decision problem is known as the *Empirical Risk Minimization (ERM) Principle* in classification (see Vapnik (1998) and references therein). The similarity between treatment choice and classification allows us to draw on recent results by Devroye et al. (1996), Tsybakov (2004), Massart and Nédélec (2006), Audibert and Tsybakov (2007), and Kerkyacharian et al. (2014), among others. We extend these convergence rate results for ERM classifiers to the treatment choice problem, accommodating the differences between classification and treatment choice and addressing issues specific to treatment choice. Establishing uniform convergence rates of the welfare regret of the EWM rule and its minimax rate optimality constitute the main theoretical contributions of this paper.

The analysis of *individualized treatment rules* has also received considerable attention in biostatistics. Qian and Murphy (2011) propose a plug-in approach using $E(Y_d|X)$ estimated by penalized least squares. They derive welfare convergence rate of $n^{-1/2}$ or better (with a margin condition), assuming that $E(Y_d|X)$ is well approximated by a sparse representation. Zhao et al. (2012) propose estimation of the treatment rule using a Support Vector Machine. This approach substitutes the EWM treatment choice objective function by a convex surrogate. They derive the welfare convergence rates that depend on the dimension of the covariates, similarly to nonparametric plug-in rules. These approaches are computationally attractive but cannot be used to choose from a constrained set of treatment rules or under a capacity constraint. Dudík et al. (2011) and Zhang et al. (2012) consider maximizing a doubly-robust estimate of the welfare over a set of policies and show by simulation that this approach outperforms the e-hybrid EWM approach in terms of welfare. Athey and Wager (2017) analytically characterize advantages of the doubly-robust approach by showing an improved constant term in the welfare regret upper bounds.

Several works in econometrics consider the *plug-in* approach to treatment choice using estimated regression equations,

$$\hat{G}_{plug-in} = \{x : \hat{\tau}^m(x) \geq 0\}, \quad \hat{\tau}^m(x) = \hat{m}_1(x) - \hat{m}_0(x), \tag{1.13}$$

where $\hat{m}_d(x)$ is a parametric or a nonparametric estimator of $E(Y_d|X = x)$. Hirano and Porter

(2009) establish local asymptotic minimax optimality of plug-in rules for parametric and semi-parametric models of treatment response. Under an aggregate budget constraint, Bhattacharya and Dupas (2012) consider nonparametric plug-in rules with propensity score weighted estimators of the regression equations and derive some of their properties. Armstrong and Shen (2015) consider statistical inference for the first-best decision rule G^*_{FB} from the perspective of inference for conditional moment inequalities. Empirical researchers often assess who should be treated by stratifying the population on an estimated predictor of Y_0, which leads to biased treatment effect estimates (Abadie et al. (2017)). Kasy (2016) considers estimation of a partially-ordered welfare ranking over treatment assignment policies with a set-identified welfare criterion.

To assess treatment effect heterogeneity, estimation and inference for conditional treatment effects based on parametric or nonparametric regressions are often reported, but the stylized output of statistical inference (e.g., confidence intervals, p-values) fails to offer the policy maker a direct guidance on what treatment rule to follow. In contrast, our EWM approach offers the policy maker a specific treatment assignment rule designed to maximize social welfare.

A treatment assignment rule could also be obtained by specifying a prior distribution for P and solving for a Bayes decision rule (see Dehejia (2005) and Chamberlain (2011) for Bayesian approaches to the treatment choice problem). Kasy (2017) proposes a nonparametric Bayesian approach to policy estimation for a range of public policy applications. In contrast to the Bayesian approach, the EWM approach utilizes only the empirical distribution of the data and does not require a prior distribution over the data generating processes.

Elliott and Lieli (2013) and Lieli and White (2010) also proposed maximizing the sample analog of a utilitarian decision criterion similar to EWM. They consider the problem of forecasting binary outcomes based on observations of (Y_i, X_i), as in Manski and Thompson (1989), where a forecast leads to a binary decision.

2 Theoretical Properties of EWM

2.1 Setup and Assumptions

Throughout our investigation of theoretical properties of EWM, we maintain the following assumptions.

Assumption 2.1.
(UCF) Unconfoundedness: $(Y_1, Y_0) \perp D|X$.
(BO) Bounded Outcomes: There exists $M < \infty$ such that the support of outcome variable Y is

contained in $[-M/2, M/2]$.
(SO) Strict Overlap: There exist $\kappa \in (0, 1/2)$ such that the propensity score satisfies $e(x) \in [\kappa, 1-\kappa]$ for all $x \in \mathcal{X}$.
(VC) VC-class: A class of decision sets $\mathcal{G}$ has a finite *VC-dimension*[3] $v < \infty$ and is countable.[4]

The assumption of unconfoundedness (selection on observables) holds if data are obtained from an experimental study with a randomized treatment assignment. In observational studies, unconfoundedness is a non-testable and often controversial assumption. Our analysis could be applied to the observational studies in which unconfoundedness is credible. The second assumption (BO) implies boundedness of the treatment effects, i.e.,

$$P_X(|\tau(X)| \leq M) = 1,$$

where P_X is the marginal distribution of X and $\tau(\cdot)$ is the conditional treatment effect $\tau(X) = E\left(Y_1 - Y_0|X\right)$. Implementing EWM does not require knowledge of M and this assumption is imposed mainly for analytical convenience. The third assumption (SO) is a standard assumption in the treatment effect literature. It is satisfied in randomized controlled trials by design, but it may be violated in observational studies if almost all the individuals are in the same group (treatment or control) for some values of X. We let $\mathcal{P}(M, \kappa)$ denote the class of distributions of (Y_0, Y_1, D, X) that satisfy Assumption 2.1 (UCF), (BO), and (SO).

The fourth assumption (VC) restricts the complexity of the class of candidate treatment rules $\mathcal{G}$ in terms of its VC-dimension. If X has a finite support, then the VC-dimension v of any class $\mathcal{G}$ does not exceed the number of support points. If some of X is continuously distributed, Assumption 2.1 (VC) requires $\mathcal{G}$ to be smaller than the Borel σ-algebra of $\mathcal{X}$. The following examples illustrate several practically relevant classes of the feasible treatment rules satisfying Assumption 2.1 (VC).

Example 2.1. *(Linear Eligibility Score) Suppose that a feasible assignment rule is constrained to those that assign the treatment according to an eligibility score. By the eligibility score, we*

[3] Let $\mathbf{x}^l \equiv \{x_1, \ldots, x_l\}$ be a finite set with $l \geq 1$ points in $\mathcal{X}$. Given a class of subsets $\mathcal{G}$ in $\mathcal{X}$, define $N(\mathbf{x}^l) = |\{\mathbf{x}^l \cap G : G \in \mathcal{G}\}|$ be the number of different subsets of $\mathbf{x}^l$ picked out by $G \in \mathcal{G}$. The VC-diemension $v \geq 1$ of $\mathcal{G}$ is defined by the largest l such that $\sup_{\mathbf{x}^l} N(\mathbf{x}^l) = 2^l$ holds (Vapnik (1998)). The VC-dimension is commonly used to measure the complexity of a class of sets in the statistical learning literature (see Vapnik (1998), Dudley (1999, Chapter 4), and van der Vaart and Wellner (1996) for extensive discussions). Note that the VC-dimension is smaller by one compared to the *VC-index* used to measure the complexity of a class of sets in the empirical process theory, e.g., van der Vaart and Wellner (1996).

[4] Countability of $\mathcal{G}$ is imposed to simplify measurability issues in proving our theoretical results. In Examples 2.1-2.3 below, we formulate $\mathcal{G}$ to be uncountable, whereas any practical implementation will only use a countable subset of $\mathcal{G}$ in search of the EWM rule.

mean a scalar-valued function of the individual's observed characteristics that determines whether one receives the treatment based on whether the eligibility score exceeds a certain threshold. The main objective of data analysis is therefore to construct an eligibility score that yields a welfare-maximizing treatment rule. Specifically, we assume that the eligibility score is constrained to being linear in a subvector of $x \in \mathbb{R}^{d_x}$, $x_{sub} \in \mathbb{R}^{d_{sub}}$, $d_{sub} \leq d_x$. The class of decision sets generated by Linear Eligibility Scores (LES) is defined as

$$\mathcal{G}_{LES} \equiv \left\{ \left\{ x \in \mathbb{R}^{d_x} : \beta_0 + x_{sub}^T \beta_{sub} \geq 0 \right\} : \left(\beta_0, \beta_{sub}^T \right) \in \mathbb{R}^{d_{sub}+1} \right\}. \tag{2.1}$$

We accordingly obtain an EWM assignment rule by maximizing

$$W_n(\beta) \equiv E_n \left[\frac{Y_i D_i}{e(X_i)} \cdot 1 \left\{ \beta_0 + X_{sub,i}^T \beta_{sub} \geq 0 \right\} + \frac{Y_i(1-D_i)}{1-e(X_i)} \cdot 1 \left\{ \beta_0 + X_{sub,i}^T \beta_{sub} < 0 \right\} \right]$$

in $\beta = \left(\beta_0, \beta_{sub}^T \right) \in \mathbb{R}^{d_{sub}+1}$. It is well known that the class of half-spaces spanned by $\left(\beta_0, \beta_{sub}^T \right) \in \mathbb{R}^{d_{sub}+1}$ has the VC-dimension $v = d_{sub} + 1$, so the requirement of finite VC-dimension in Assumption 2.1 (VC) holds. In Section C of Kitagawa and Tetenov (2017c), we discuss how to compute $\hat{G}_{EWM}$ when the class of decision sets is given by $\mathcal{G}_{LES}$. A plug-in rule based on a parametric linear regression also selects a treatment rule from $\mathcal{G}_{LES}$, but their welfare does not converge to the maximum welfare $W^*_{\mathcal{G}_{LES}}$ if the regression equations are misspecified, whereas the welfare of $\hat{G}_{EWM}$ always does (as shown in Theorem 2.1 below).

Example 2.2. *(Generalized Eligibility Score) Let $f_j(\cdot)$, $j = 1, \ldots, m$, and $g(\cdot)$ be known functions of $x \in \mathbb{R}^{d_x}$. Consider a class of assignment rules generated by Generalized Eligibility Scores (GES),*

$$\mathcal{G}_{GES} \equiv \left\{ \left\{ x \in \mathbb{R}^{d_x} : \ \sum\nolimits_{j=1}^{m} \beta_j f_j(x) \geq g(x) \right\}, \ (\beta_1, ..., \beta_m) \in \mathbb{R}^m \right\}.$$

The class of decision sets $\mathcal{G}_{GES}$ generalizes the linear eligibility score rules (2.1), as it allows for eligibility scores that are nonlinear in x, i.e., $\mathcal{G}_{GES}$ can accommodate decision sets that partition the space of covariates by nonlinear boundaries. It can be shown that $\mathcal{G}_{GES}$ has the VC-dimension $v = m + 1$ (Theorem 4.2.1 in Dudley (1999)).

Example 2.3. *(Intersection Rule of Multiple Eligibility Scores) Consider a situation where there are $L \geq 2$ eligibility scores. Let $\mathcal{G}_{GES,l}$, $l = 1, \ldots, L$, be classes of decision sets such that each of them is generated by contour sets of the l-th eligibility score. Suppose that a feasible decision rule is constrained to those that assign the treatment if the individual has all the L eligibility scores exceeding thresholds. In this case, the class of decision sets is constructed by the intersections, $\mathcal{G} \equiv \bigcap_{l=1}^{L} \mathcal{G}_{GES,l} = \left\{ \bigcap_{l=1}^{L} G_l : G_l \in \mathcal{G}_{GES,l}, l = 1, \ldots, L \right\}$. An intersection of a finite number of*

VC-classes is a VC-class with a finite VC-dimension (Theorem 4.5.4 in Dudley (1999)). We can also consider a class of treatment rules that assigns a treatment if at least one of the L eligibility scores exceeds a threshold. In this case, instead of intersections, the class of decision sets is formed by the unions of $\{\mathcal{G}_{GES,l}, l = 1, \ldots, L\}$, which is also known to have a finite VC-dimension (Theorem 4.5.4 in Dudley (1999))

2.2 Uniform Rate Optimality of EWM

To analyze statistical performance of EWM rules, we focus on a non-asymptotic upper bound of the worst-case welfare loss $\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{G}_{EWM})\right]$ and examine how it depends on sample size n and VC-dimension v. This finite sample upper bound allows us to assess the uniform convergence rate of the welfare and to examine how richness (complexity) of the class of candidate decision rules affects the worst-case performance of EWM. The main reason that we focus on the uniform convergence rate rather than a pointwise convergence rate is that the pointwise convergence rate of the welfare loss can vary depending on a feature of the data distribution and fails to provide a guaranteed learning rate of an optimal policy when no additional assumption, other than Assumption 2.1, is available.

For heuristic illustration of the derivation of the uniform convergence rate, consider the following inequality, which holds for any $\tilde{G} \in \mathcal{G}$:

$$\begin{aligned} W(\tilde{G}) - W(\hat{G}_{EWM}) &= W(\tilde{G}) - W_n(\hat{G}_{EWM}) + W_n(\hat{G}_{EWM}) - W(\hat{G}_{EWM}) \\ &\leq W(\tilde{G}) - W_n(\tilde{G}) + \sup_{G \in \mathcal{G}} \left| W_n(G) - W(G) \right| \\ (\because\ & W_n(\hat{G}_{EWM}) \geq W_n(\tilde{G})\ ) \\ &\leq 2 \sup_{G \in \mathcal{G}} \left| W_n(G) - W(G) \right|. \end{aligned} \tag{2.2}$$

Since it applies to $W(\tilde{G})$ for all $\tilde{G}$, it also applies to $W_{\mathcal{G}}^* = \sup W(\tilde{G})$:

$$W_{\mathcal{G}}^* - W(\hat{G}_{EWM}) \leq 2 \sup_{G \in \mathcal{G}} \left| W_n(G) - W(G) \right|. \tag{2.3}$$

Therefore, the expected welfare loss can be bounded uniformly in P by a distribution-free upper bound of $E_{P^n}(\sup_{G \in \mathcal{G}} |W_n(G) - W(G)|)$. Since $W_n(G) - W(G)$ can be seen as the centered empirical process indexed by $G \in \mathcal{G}$, an application of the existing moment inequality for the supremum of centered empirical processes indexed by a VC-class yields the following distribution-free upper bound. A proof, which closely follows the proofs of Theorems 1.16 and 1.17 in Lugosi (2002) in the classification problem, is given in Section A.2 of Kitagawa and Tetenov (2017c).

Theorem 2.1. *Under Assumption 2.1, we have*

$$\sup_{P\in\mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^{*}-W(\hat{G}_{EWM})\right]\leq C_1\frac{M}{\kappa}\sqrt{\frac{v}{n}},$$

where C_1 is a universal constant defined in Lemma A.4 in Kitagawa and Tetenov (2017c).

This theorem shows that the convergence rate of the worst-case welfare loss for the EWM rule is no slower than $n^{-1/2}$. The upper bound is increasing in the VC-dimension of $\mathcal{G}$, implying that, as the candidate treatment assignment rules become more complex in terms of VC-dimension, $\hat{G}_{EWM}$ tends to overfit the data in the sense that the distribution of regret $W_{\mathcal{G}}^{*}-W(\hat{G}_{EWM})$ is more and more dispersed, and, with n fixed, this overfitting results in inflating the average welfare regret.[5]

The next theorem concerns a universal lower bound of the worst-case average welfare loss. It shows that no data-based treatment choice rule can have a uniform convergence rate faster than $n^{-1/2}$.

Theorem 2.2. *Suppose that Assumption 2.1 holds. For any treatment choice rule $\hat{G}$ as a function of $(Z_1,\ldots,Z_n)$, it holds*

$$\sup_{P\in\mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^{*}-W(\hat{G})\right]\geq 2^{-1}\exp\left\{-2\sqrt{2}\right\}M\sqrt{\frac{v}{n}} \quad \textit{for all } n\geq 16v.$$

We derive this lower bound by bounding below the worst-case welfare regret by the risk of a parametric Bayes decision problem (i.e., a prior that only supports a parametric subclass $\mathcal{P}^{*}\subset\mathcal{P}(\mathcal{M},\kappa)$) and maximizing the Bayes risk over $P\in\mathcal{P}^{*}$. A similar proof technique appears in Devroye and Lugosi (1995) in their regret lower bound analysis of ERM classifiers.

This theorem, combined with Theorem 2.1, implies that $\hat{G}_{EWM}$ is *minimax rate optimal* over the class of data generating process $\mathcal{P}(M,\kappa)$, since the rate of the convergence of the upper bound of $\sup_{P\in\mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^{*}-W(\hat{G}_{EWM})\right]$ agrees with the convergence rate of the universal lower bound. Accordingly, we can conclude that no other data-driven procedure for obtaining a treatment choice rule can outperform $\hat{G}_{EWM}$ in terms of the uniform convergence rate over $\mathcal{P}(M,\kappa)$. It is worth noting that the rate lower bound is uniform in P and does not apply pointwise. Theorem 2.3 shows that EWM rules have faster convergence rates for *some* distributions. It is also possible that $E_{P^n}\left[W(\hat{G})\right]>W_{\mathcal{G}}^{*}$ for some pairs of $\hat{G}$ and P, but it can never hold for all distributions in $\mathcal{P}(M,\kappa)$.[6]

[5] Note that $W_{\mathcal{G}}^{*}$ weakly increases if a more complex class $\mathcal{G}$ is chosen. Our welfare loss criterion is defined for a specific class $\mathcal{G}$ and does not capture the potential gain in the maximal welfare from the choice of a more complex $\mathcal{G}$.

[6] For example, if $\hat{G}$ is a nonparametric plug-in rule and the first-best decision rule G_{FB}^{*} for distribution P does not

Remark 2.1. *Capacity or budget constraints that restrict the proportion of the target population that could be assigned to treatment exist in various treatment choice problems. An attractive feature of the empirical welfare maximization method is the simplicity of incorporating these constraints in the estimation of a treatment choice rule.*

Assume that the proportion of the target population that could receive treatment 1 cannot exceed $K \in (0, 1)$. If the population distribution of covariates P_X were known, maximization of the empirical welfare criterion could be simply restricted to sets in class $\mathcal{G}$ that satisfy the capacity constraint $\mathcal{G}^K \equiv \{G \in \mathcal{G} : P_X(G) \leq K\}$. Being a subset of $\mathcal{G}$, the class of sets $\mathcal{G}^K$ has the same complexity as $\mathcal{G}$ (or lower), and Theorem 2.1 could be applied simply by replacing $\mathcal{G}$ with $\mathcal{G}^K$.

When P_X is unknown, it is not guaranteed with certainty that estimated treatment rule $\hat{G}$ satisfies the capacity constraint. To evaluate the welfare in this setting, we assume that if the treatment rule G violates the capacity constraint, $P_X(G) > K$, then the scarce treatment is randomly allocated ("rationed") to a fraction $\frac{K}{P_X(G)}$ of the assigned recipients with $X \in G$ independently of (X, Y_0, Y_1).[7] *If G does not violate the capacity constraint, then there is no rationing and all recipients with covariates $X \in G$ receive treatment 1. This allows us to clearly define the capacity-constrained welfare of the treatment rule indexed by any subset $G \subset \mathcal{X}$ of the covariate space as*

$$W^K(G) \equiv E_P \left[\begin{array}{c} \left[Y_1 \cdot \min\left\{1, \frac{K}{P_X(G)}\right\} + Y_0 \cdot \left(1 - \min\left\{1, \frac{K}{P_X(G)}\right\}\right)\right] \cdot 1\{X \in G\} \\ +Y_0 \cdot 1\{X \notin G\} \end{array} \right].$$

Then the capacity-constrained welfare gain of the treatment rule G relative to the no-treatment policy is given by

$$V^K(G) \equiv W^K(G) - W^K(\emptyset) = \min\left\{1, \frac{K}{P_X(G)}\right\} \cdot E_P[\tau(X)1\{X \in G\}].$$

Observe that rationing dilutes the effect of treatment rules that violate the capacity constraint and we take into account this effect on welfare. We hence propose a treatment rule that maximizes the empirical analog of the capacity-constrained welfare gain $V^K(G)$ (and, hence, welfare):

$$\hat{G}^K \equiv \arg\max_{G \in \mathcal{G}} V_n^K(G), \tag{2.4}$$

where

$$V_n^K(G) \equiv \min\left\{1, \frac{K}{P_{X,n}(G)}\right\} \cdot E_n\left[\left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1 - D_i)}{1 - e(X_i)}\right) \cdot 1\{X_i \in G\}\right],$$

belong to $\mathcal{G}$, then the welfare of $\hat{G}$ will exceed $W^*_{\mathcal{G}}$ in sufficiently large samples. However, the uniform lower bound still applies because there exist other distributions for which $E_{P^n} W(\hat{G}) \leq W^*_{\mathcal{G}} - (n^{-1/2}$ bound) for the same sample size.

[7] In comparison, nonparametric plug-in treatment rules proposed by Bhattacharya and Dupas (2012) are only required to satisfy the capacity constraint on average over repeated data samples.

and $P_{X,n}$ is the empirical probability distribution of $(X_1, \ldots, X_n)$. Theorem D.1 (Kitagawa and Tetenov, 2017b) shows that similarly to Theorem 2.1, the expected welfare of $\hat{G}^K$ converges to the maximum at least at $n^{-1/2}$ rate.

Remark 2.2. *Empirical Welfare Maximization method can be adapted to situations where a target population shares the conditional treatment effect with the sampled population, but differs in the distribution of covariates X. Let $\rho(x) \equiv p_X^T(x)/p_X(x)$ be the density ratio of the marginal distributions of X, where p_X and p_X^T are those of the sampled population and the target population. Assume $\rho(x) \leq \bar{\rho} < \infty$ for all x. The welfare gain of treatment rule G on the target population can be written as*

$$V^T(G) \equiv \int_{\mathcal{X}} \tau(x) 1\{x \in G\} \rho(x) \; dP_X(x).$$

Since the first-best treatment rule $G_{FB}^* = 1\{x : \tau(x) \geq 0\}$ is the same in the sampled and the target populations, if $G_{FB}^* \in \mathcal{G}$, we could directly apply the EWM rule computed for the sampled population to the target population. In contrast, if the first-best policy is not feasible ($G_{FB}^* \notin \mathcal{G}$), the second-best policies for the sampled and the target populations are generally different, and the welfare of treatment rules proposed in the previous sections does not generally converge to the second-best in the target population $\sup_{G \in \mathcal{G}} V^T(G)$.

The second-best in the target population could be obtained by reweighting the argument of the EWM problem by the density ratio $\rho(X_i)$:

$$\hat{G}_{EWM}^T \in \arg\max_{G \in \mathcal{G}} E_n \left[\left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)} \right) \cdot \rho(X_i) \cdot 1\left\{X_i \in G\right\} \right]. \tag{2.5}$$

As an extension of Theorem 2.1, the welfare loss of the reweighted EWM rule in the target population can be shown to converge to zero at least at $n^{-1/2}$ rate.

Remark 2.3. *The EWM rule (1.7) is invariant to multiplying all outcomes Y_i by a positive constant, but is not invariant to adding a constant. If all Y_i's are replaced by $Y_i + c$, the welfare estimate $W_n(G)$ changes by $c \cdot E_n \left[\frac{D_i}{e(X_i)} \cdot 1\left\{X_i \in G\right\} + \frac{1-D_i}{1-e(X_i)} \cdot 1\left\{X_i \notin G\right\} \right]$. This difference converges to c for every G, but its value varies with G in finite samples. This could pose problems for applied work because the researcher has some room to change the treatment assignment rule by changing the coding of the outcome variable. We propose a simple modification of the EWM rule invariant to positive affine transformations of outcomes. Denote by $Y_i^{dm} \equiv Y_i - E_n\left[Y_i\right]$ the outcomes demeaned by their sample mean; they are invariant to changing Y_i's by a constant. Then the demeaned EWM* *rule*

$$\hat{G}^{dm}_{EWM} \in \arg\max_{G\in\mathcal{G}} W^{dm}_n(G), \tag{2.6}$$

$$\textit{where } W^{dm}_n(G) \equiv E_n\left[\frac{Y^{dm}_i D_i}{e(X_i)}\cdot 1\{X_i \in G\} + \frac{Y^{dm}_i(1-D_i)}{1-e(X_i)}\cdot 1\{X_i \notin G\}\right], \tag{2.7}$$

is invariant to positive affine transformations of Y_i.

The rate result of Theorem 2.1 also holds for demeaned EWM rule $\hat{G}^{dm}_{EWM}$, as we show in Section D.2 of Kitagawa and Tetenov (2017b).[8] *In our empirical application we compute the demeaned EWM treatment rules.*

Remark 2.4. *In Section D.3 of Kitagawa and Tetenov (2017b) we discuss how the EWM approach could be applied with more than two treatments. The rate result of Theorem 2.1 holds for multiple treatments if treatment assignment covariate subsets for each treatment belong to a VC-class.*

2.3 Rate Improvement by Margin Assumption

The welfare loss upper bounds obtained in Theorem 2.1 can indeed tighten up and the uniform convergence rate can improve, as we further constrain the class of data generating processes. In this section, we investigate (i) what feature of data generating processes can affect the upper bound on the welfare loss of the EWM rule, and (ii) whether or not the EWM rule remains minimax rate optimal even under the additional constraints. For this goal, we consider imposing the following two assumptions.

Assumption 2.2.
(FB) Correct Specification: The first-best treatment rule G^*_{FB} defined in (1.8) belongs to the class of candidate treatment rules $\mathcal{G}$.
(MA) Margin Assumption: There exist constants $0 < \eta \leq M$ and $0 < \alpha < \infty$ such that

$$P_X(|\tau(X)| \leq t) \leq (t/\eta)^{\alpha}, \quad \forall 0 \leq t \leq \eta,$$

where $M < \infty$ is the constant as defined in Assumption 2.1 (BO).

The assumption of correct specification means that the class of feasible policy rules specified by $\mathcal{G}$ contains an unconstrained first-best treatment rule G^*_{FB}. This assumption is plausible if, for instance, the policy maker's specification of $\mathcal{G}$ is based on a credible assumption about the shape

[8] In our simulations demeaned EWM never performed much worse than standard EWM in terms of welfare. Demeaned EWM performed much better in cases where $E[Y]$ was very far from zero.

of the contour set $\{x : \tau(x) \geq 0\}$. This assumption can be, on the other hand, restrictive if the specification of $\mathcal{G}$ comes from some exogenous constraints for feasible policy rules, as in the case of Example 2.1.

The second assumption (MA) concerns the way in which the distribution of conditional treatment effect $\tau(X)$ behaves in the neighborhood of $\tau(X) = 0$. A similar assumption has been considered in the literature on classification analysis (Mammen and Tsybakov (1999), Tsybakov (2004), among others), and we borrow the term "margin assumption" from Tsybakov (2004). Parameters η and α characterize the size of population with the conditional treatment effect close to the margin $\tau(X) = 0$. *Smaller* η and α imply that *more* individuals can concentrate in a neighborhood of $\tau(X) = 0$. The next examples illustrate this interpretation of η and α.

Example 2.4. *Suppose that X contains a continuously distributed covariate and that the conditional treatment effect $\tau(X)$ is continuously distributed. If the probability density function of $\tau(X)$ is bounded from above by $p_\tau < \infty$, then the margin assumption holds with $\alpha = 1$ and $\eta = (2p_\tau)^{-1}$.*

Example 2.5. *Suppose that X is a scalar and follows the uniform distribution on $[-1/2, 1/2]$. Specify the conditional treatment effect to be $\tau(X) = (-X)^3$. In this specification, $\tau(X)$ "flats out" at $X = 0$, and accordingly, the density function of $\tau(X)$ is unbounded in the neighborhood of $\tau(X) = 0$. This specification leads to $P_X(|\tau(X)| \leq t) = 2t^{1/3}$, so the margin assumption holds with $\alpha = 1/3$ and $\eta = 1/8$.*

Example 2.6. *Suppose that the distribution of X is the same as in Example 2.5. Let $h > 0$ and specify $\tau(X)$ as*

$$\tau(X) = \begin{cases} X - h & \text{for} \quad X \leq 0, \\ X + h & \text{for} \quad X > 0. \end{cases}$$

This $\tau(X)$ is discontinuous at $X = 0$, and the distribution of $\tau(X)$ has zero probability around the margin of $\tau(X) = 0$. It holds

$$P_X(|\tau(X)| \leq t) = \begin{cases} 0 & \text{for} \quad t \leq h, \\ 2(t-h) & \text{for} \quad h < t \leq \frac{1}{2} + h. \end{cases}$$

By setting $\eta = h$, the margin condition holds for arbitrarily large α. In general, if the distribution of $\tau(X)$ has a gap around the margin of $\tau(X) = 0$, the margin condition holds with arbitrarily large α.

From now on, we denote the class of P satisfying Assumptions 2.1 and 2.2 by $\mathcal{P}_{FB}(M, \kappa, \eta, \alpha)$.[9] The next theorem provides the upper bound of the welfare loss of the EWM rule when a class of data distributions is constrained to $\mathcal{P}_{FB}\left(M, \kappa, \eta, \alpha\right)$.

Theorem 2.3. *Under Assumptions 2.1 and 2.2,*

$$\sup_{P \in \mathcal{P}_{FB}(M,\kappa,\eta,\alpha)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G}_{EWM})\right] \leq c\left(\frac{v}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$$

holds for all n, where c is a positive constant that depends only on M, κ, η, and α.

Similarly to Theorem 2.1, the presented welfare loss upper bound is non-asymptotic, and it is valid for every sample size. Our derivation of this theorem can be seen as an extension of the finite sample risk bound for the classification error shown in Theorem 2 of Massart and Nédélec (2006). Our rate upper bound is consistent with the uniform convergence rate of the classification risk of the empirical risk minimizing classifier shown in Theorem 1 of Tsybakov (2004).[10] This coincidence is somewhat expected, given that the empirical welfare criterion that the EWM rule maximizes resembles the empirical classification risk in the classification problem.

The next theorem shows that the uniform convergence rate of $n^{-\frac{1+\alpha}{2+\alpha}}$ obtained in Theorem 2.3 attains the minimax rate lower bound, implying that any treatment choice rule $\hat{G}$ based on data (including $\hat{G}_{EWM}$) cannot attain a uniform convergence rate faster than $n^{-\frac{1+\alpha}{2+\alpha}}$. This means that the EWM rule remains rate optimal even when the class of data generating processes is constrained additionally by Assumption 2.2.[11]

Theorem 2.4. *Suppose Assumptions 2.1 and 2.2 hold. Assume that the VC-dimension of $\mathcal{G}$ satisfies $v \geq 2$. Then, for any treatment choice rule $\hat{G}$ as a function of $(Z_1, \ldots, Z_n)$, it holds*

$$\sup_{P \in \mathcal{P}_{FB}(M,\kappa,\eta,\alpha)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G})\right] \geq 2^{-1}\exp\left\{-2\sqrt{2}\right\} M^{\frac{2(1+\alpha)}{2+\alpha}} \eta^{-\frac{\alpha}{2+\alpha}} \left(\frac{v-1}{n}\right)^{\frac{1+\alpha}{2+\alpha}}$$

for all $n \geq \max\left\{(M/\eta)^2, 4^{2+\alpha}\right\}(v-1)$.

[9] Note that $P_{FB}(M, \kappa, \eta, \alpha)$ depends on the set of feasible treatment rules $\mathcal{G}$ via Assumption 2.2 (FB).

[10] Tsybakov (2004) defines the complexity of the decision sets $\mathcal{G}$ in terms of the growth coefficient ρ of the bracketing number of $\mathcal{G}$. We control complexity of $\mathcal{G}$ in terms of the VC-dimension, which corresponds to Tsybakov's growth coefficient ρ being arbitrarily close to zero.

[11] Assumption 2.2 rules out data generating processes with $P_X(\tau(X) = 0) > 0$, which can constitute focal null hypotheses (often $P_X(\tau(X) = 0) = 1$) in program evaluation studies. A practical implication of the refined minimax rate result shown in this section is that the EWM rule remains a recommended choice even when we know ex ante that there is substantial effect heterogeneity in x and $\tau(x) \neq 0$ for most x.

The following remarks summarize some analytical insights associated with Theorems 2.1 - 2.4.

Remark 2.5. *The convergence rates of the worst-case EWM welfare loss obtained by Theorems 2.1 and 2.3 highlight how margin coefficient α influences the uniform performance of the EWM rule. Higher α improves the welfare loss convergence rate of EWM, and the convergence rate approaches n^{-1} in an extreme case, where the distribution of $\tau(X)$ has a gap around $\tau(X) = 0$. As fewer individuals are around the margin of $\tau(X) = 0$, we can attain the maximal welfare quicker. Conversely, as α approaches zero (more individuals around the margin), the welfare loss convergence rate of EWM approaches $n^{-1/2}$, and it corresponds to the uniform convergence rate of Theorem 2.1.*

Remark 2.6. *The upper bounds of welfare loss convergence rate shown in Theorems 2.1 and 2.3 are increasing in the VC-dimension of $\mathcal{G}$. Since they are valid at every n, we can allow the VC-dimension of the candidate treatment rules to grow with the sample size. For instance, if we consider a sequence of candidate decision sets $\{\mathcal{G}_n : n = 1, 2, \dots\}$, for which the VC-dimension grows with the sample size at rate n^{λ}, $0 < \lambda < 1$, Theorems 2.1 and 2.3 imply that the welfare loss uniform convergence rate of the EWM rule slows down to $n^{-\frac{1-\lambda}{2}}$ for the case without Assumption 2.2 and to $n^{-(1-\lambda)\frac{(1+\alpha)}{2+\alpha}}$ for the case with Assumption 2.2.*[12] *Note that the welfare loss lower bounds shown in Theorems 2.2 and 2.4 have the VC-dimensions of the same order as in the corresponding upper bounds, so we can conclude that the EWM rule is also minimax rate optimal even in the situations where the complexity of $\mathcal{G}$ grows with the sample size.*

Remark 2.7. *Note that the welfare loss lower bounds of Theorems 2.2 and 2.4 are valid for any estimated treatment choice rule $\hat{G}$ irrespective of whether $\hat{G}$ is constrained to $\mathcal{G}$ or not. Therefore, the nonparametric plug-in rule $\hat{G}_{plug-in}$ defined in (1.13) is subject to the same lower bound.*[13]

Remark 2.8. *Let $\mathcal{P}_{FB}(M, \kappa)$ be the class of data generating processes that satisfy Assumption 2.1 and Assumption 2.2 (FB). A close inspection of the proofs of Theorems 2.1 and 2.2 given in Section A.2 of Kitagawa and Tetenov (2017c) shows that the same lower and upper bounds of Theorems 2.1 and 2.2 can be obtained even when $\mathcal{P}(M, \kappa)$ is replaced with $\mathcal{P}_{FB}(M, \kappa)$. In this sense, Assumption 2.2 (MA) plays the main role in improving the welfare loss convergence rate.*

[12] Note that for the case without Assumption 2.2 (FB), the maximal attainable welfare $W^*_{\mathcal{G}}$ increases weakly as the complexity of $\mathcal{G}$ grows. On the other hand, with Assumption 2.2 (FB), the set of data generating processes $\mathcal{P}_{FB}(M, \kappa, \eta, \alpha)$ expands as the complexity of $\mathcal{G}$ grows.

[13] Section D.4 of Kitagawa and Tetenov (2017b) discusses the welfare loss uniform convergence rate of the nonparametric plug-in rule.

2.4 Unknown Propensity Score

We have so far considered situations where the true propensity score is known. This would not be the case if the data were obtained from an observational study in which the assignment of treatment is not generally under the control of the experimenter. To cope with the unknown propensity score, this section considers two hybrids of the EWM approach and the parametric/nonparametric plug-in approach: the m-hybrid rule defined in (1.11) and the e-hybrid rule defined in (1.12). The e-hybrid rule employs the trimming rule $1\left\{\varepsilon_n \leq \hat{e}\left(X_i\right) \leq 1-\varepsilon_n\right\}$ with a deterministic sequence $\left\{\varepsilon_n : n=1,2,\ldots\right\}$, which we assume to converge to zero faster than some polynomial rate, $\epsilon_n \leq O\left(n^{-a}\right)$, $a>0$.[14]

The next condition concerns the convergence rate of the average estimation error of the conditional treatment effect estimators.

Condition 2.1.
(m) (m-hybrid case): Let $\hat{\tau}^m(x)=\hat{m}_1(x)-\hat{m}_0(x)$ *be an estimator for the conditional treatment effect* $\tau(x)=m_1(x)-m_0(x)$. *For a class of data generating processes* $\mathcal{P}_m$, *there exists a sequence* $\psi_n \rightarrow \infty$ *such that*

$$\limsup_{n\rightarrow\infty} \sup_{P\in\mathcal{P}_m} \psi_n E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tau}^m(X_i)-\tau(X_i)\right|\right]<\infty \tag{2.8}$$

holds.
(e) (e-hybrid case): Let $\hat{\tau}_i^e=\left[\frac{Y_iD_i}{\hat{e}(X_i)}-\frac{Y_i(1-D_i)}{1-\hat{e}(X_i)}\right]\cdot 1\left\{\varepsilon_n \leq \hat{e}\left(X_i\right) \leq 1-\varepsilon_n\right\}$ *be an estimator for* $\tau_i=\frac{Y_iD_i}{e(X_i)}-\frac{Y_i(1-D_i)}{1-e(X_i)}$, *where* $\hat{e}(\cdot)$ *is an estimated propensity score. For a class of data generating processes* $\mathcal{P}_e$, *there exists a sequence* $\phi_n \rightarrow \infty$ *such that*

$$\limsup_{n\rightarrow\infty} \sup_{P\in\mathcal{P}_e} \phi_n E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tau}_i^e-\tau_i\right|\right]<\infty. \tag{2.9}$$

In Section E of Kitagawa and Tetenov (2017b), we show that the estimators $\hat{\tau}^m\left(\cdot\right)$ and $\hat{\tau}_i^e$ constructed via local polynomial regressions satisfy this condition for a certain class of data generating processes. Theorems 2.5 and 2.6 below derive the uniform convergence rate bounds of the hybrid rules in two different scenarios. In Theorem 2.5, we constrain the class of data generating processes only by Assumption 2.1 and Condition 2.1, and, importantly, we allow the class of decision rules $\mathcal{G}$ to exclude the first-best rule G_{FB}^*. See Kitagawa and Tetenov (2017c) for proofs of these theorems.

[14] The trimming sequence ε_n is introduced only to simplify the derivation of the rate upper bound of the welfare loss. In practical terms, if the overlap condition is well satisfied in the given data, the trimming is not necessary for computing the e-hybrid rule.

Theorem 2.5. *Suppose Assumption 2.1 holds.*

(m) (m-hybrid case): Given a class of data generating processes $\mathcal{P}_m$, if an estimator for the conditional treatment effect $\hat{\tau}^m(\cdot)$ satisfies Condition 2.1 (m), then,

$$\sup_{P \in \mathcal{P}_m \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{G}_{m-hybrid})\right] \leq O\left(\psi_n^{-1} \vee n^{-1/2}\right).$$

(e) (e-hybrid case): Given a class of data generating processes $\mathcal{P}_e$, if an estimator for the propensity score $\hat{e}(\cdot)$ satisfies Condition 2.1 (e), then,

$$\sup_{P \in \mathcal{P}_e \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{G}_{e-hybrid})\right] \leq O\left(\phi_n^{-1} \vee n^{-1/2}\right).$$

A comparison of Theorem 2.5 with Theorem 2.1 shows that the uniform rate upper bounds for the hybrid EWM rules are no faster than the welfare loss convergence rate of the EWM with known propensity score. Note that if some nonparametric estimator is used to estimate $\tau(\cdot)$ or $e\left(\cdot\right)$, ψ_n or ϕ_n specified in Condition 2.1 is generally slower than $n^{1/2}$. Hence, the welfare loss upper bounds of the hybrid rules are determined by the nonparametric rate ψ_n^{-1} or ϕ_n^{-1}. A special case where the estimation of $\tau\left(\cdot\right)$ or $e\left(\cdot\right)$ does not affect the uniform convergence rate is when $\tau\left(\cdot\right)$ or $e\left(\cdot\right)$ is assumed to belong to a parametric family and it is estimated parametrically, i.e., ψ_n or ϕ_n is equal to $n^{1/2}$.

In the second scenario, we consider the case where $\mathcal{G}$ contains the first-best decision rule G_{FB}^* and the data generating processes are constrained further by the margin assumption (Assumption 2.2) with margin coefficient $\alpha \in (0,1]$.

Theorem 2.6. *Suppose Assumptions 2.1 and 2.2 hold with a margin coefficient $\alpha \in (0,1]$. Assume that a stronger version of Condition 2.1 holds, where (2.8) and (2.9) are replaced by*

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_m} E_{P^n}\left[\left(\tilde{\psi}_n \max_{1\leq i\leq n} \left|\hat{\tau}^m(X_i) - \tau(X_i)\right|\right)^2\right] < \infty \quad \text{and} \tag{2.10}$$

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_e} E_{P^n}\left[\left(\tilde{\phi}_n \max_{1\leq i\leq n} \left|\hat{\tau}_i^e - \tau_i\right|\right)^2\right] < \infty, \tag{2.11}$$

for sequences $\tilde{\psi}_n \to \infty$ and $\tilde{\phi}_n \to \infty$, respectively. Then, we have

$$\begin{aligned}
\sup_{P \in \mathcal{P}_m \cap \mathcal{P}_{FB}(M,\kappa,\alpha,\eta)} E_{P^n}\left[W(G_{FB}^*) - W(\hat{G}_{m-hybrid})\right] &\leq O\left(\tilde{\psi}_n^{-(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}} \log \tilde{\psi}_n\right), \\
\sup_{P \in \mathcal{P}_e \cap \mathcal{P}_{FB}(M,\kappa,\alpha,\eta)} E_{P^n}\left[W(G_{FB}^*) - W(\hat{G}_{e-hybrid})\right] &\leq O\left(\tilde{\phi}_n^{-(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}} \log \tilde{\phi}_n\right).
\end{aligned}$$

Theorem 2.6 shows that even when $\tau(\cdot)$ or $e(\cdot)$ have to be estimated, the margin coefficient α influences the rate upper bound of the welfare loss. A higher α leads to a faster rate of the welfare loss convergence regardless of whether $\tau(\cdot)$ and $e(\cdot)$ are estimated parametrically or nonparametrically. In the situation where $\tau(\cdot)$ or $e(\cdot)$ is estimated parametrically (with a compact support of X), $\tilde{\psi}_n$ or $\tilde{\phi}_n$ is equal to $n^{1/2}$; thus, the uniform welfare loss convergence rate is given by the second argument in $O(\cdot)$, $n^{-\frac{1+\alpha}{2+\alpha}} \log n$. On the other hand, when $\tau(\cdot)$ or $e(\cdot)$ is estimated nonparametrically, which of the two terms in $O(\cdot)$ converges slower depends on the dimension of X and the degree of smoothness of the underlying nonparametric function. For instance, Section E of Kitagawa and Tetenov (2017b) shows for suitably constructed $\mathcal{P}_m$, local polynomial estimation for $\tau(x)$ can attain $\tilde{\psi}_n = n^{\frac{1}{2+d_x/\beta_m}} (\log n)^{-\frac{1}{2+d_x/\beta_m}-2}$, where d_x is the dimension of X and β_m is the degree of Hölder smoothness of $m_1(x)$ and $m_0(x)$. Hence, if $d_x/\beta_m > \alpha$, the rate upper bound of Theorem 2.6 implies that the welfare loss convergence rate of the m-hybrid rule is guaranteed to be $n^{-\frac{1+\alpha}{2+d_x/\beta_m}} (\log n)^{\left(\frac{1}{2+d_x/\beta_m}+2\right)(1+\alpha)}$.

Note that Theorems 2.5 and 2.6 concern the upper bound of the convergence rate. We do not have the universal rate lower bound results for these constrained classes of data generating processes. We leave the investigation of the sharp rate bound of the hybrid-EWM welfare loss for future research.

3 Empirical Application

We illustrate the Empirical Welfare Maximization method by applying it to experimental data from the National Job Training Partnership Act (JTPA) Study. A detailed description of the study and an assessment of average program effects for five large subgroups of the target population is found in Bloom et al. (1997). The study randomized whether applicants would be eligible to receive a mix of training, job-search assistance, and other services provided by the JTPA for a period of 18 months. It collected background information on the applicants prior to random assignment, as well as administrative and survey data on applicants' earnings in the 30-month period following the assignment. Our sample consists of 9,223 observations with available data on years of education and pre-program earnings from the sample of adults (22 years and older) used in the original evaluation of the program and in subsequent studies (Bloom et al., 1997, Heckman et al., 1997, Abadie et al., 2002). The probability of being assigned to the treatment was two thirds in this sample.

We use two welfare outcome measures for our illustration. The first is the total individual earnings in the 30 months after program assignment. The second is the 30-month earnings minus \$774 for individuals who were assigned to treatment. This is the average cost of services per treatment *assignment* (estimated from Table 5 in Bloom et al. (1997)), which takes into account varying take-up of program services. The first outcome measure reflects social preferences that put no weight on the costs of the program incurred by the government. The second outcome measure weighs participants' gains and the government's losses equally.

We consider these outcomes (and costs) from an *intention-to-treat* perspective. We view the policy maker's problem as a choice of the eligibility criteria and not as a choice of the take-up rate (decided by individuals); hence, we are not interested in the treatment effect on compliers. Since we have to compare welfare effects of policies that assign different proportions of the population to the treatment, we report estimates of *the average effect per population member* $E[(Y_1 - Y_0) \cdot 1\{X \in G\}]$, which is proportional to the total welfare effect of the treatment rule G.

We consider conditioning treatment assignment on two pre-treatment variables: the individual's years of education and earnings in the year prior to the assignment. Both variables may plausibly affect how much effect the individual gets from the program services. We do not use race, gender, or age. Though treatment effects may vary with these characteristics, policy makers usually cannot use them to determine treatment assignment. Education and earnings are generally verifiable characteristics. This is an important feature for implementing the proposed treatment assignment because the empirical welfare estimates are not valid for the target population if the individuals could manipulate their characteristics to obtain the desired treatment.

Table 1 reports the estimated welfare gains $W_n^{dm}(G) - W_n^{dm}(\emptyset)$ of alternative treatment rules relative to the benchmark of assigning no-one to treatment, as well as the estimated proportion of individuals assigned to treatment 1 by each rule.[15]

We consider three candidate classes of treatment rules for EWM.[16] The first is the class of *quadrant treatment rules:*

$$\mathcal{G}_Q \equiv \left\{ \begin{array}{c} \{x : s_1(\text{education} - t_1) > 0 \;\&\; s_2(\text{prior earnings} - t_2) > 0\}\,, \\ s_1, s_2 \in \{-1, 1\}, t_1, t_2 \in \mathbb{R}. \end{array} \right\} \tag{3.1}$$

This class of treatment rules is easily implementable and is often used in practice. To be assigned to treatment according to such a rule, an individual's education and pre-program earnings have to be above (or below) some specific thresholds. The EWM method searches over all possible thresholds and directions. Figure 1 demonstrates the quadrant treatment rules selected by the EWM criterion. The full shaded area indicates individuals who would be assigned to treatment if it were costless

[15] We report welfare gain estimates using equation (2.7) with demeaned outcome variable (see Remark 2.3). These estimates are invariant to translation of the outcome variable by a constant.

[16] Specifically, we implement demeaned EWM described in Remark 2.3 with known constant propensity score $e(X_i) = 2/3$. Further details on computing EWM rules are found in Section C of Kitagawa and Tetenov (2017c).

(education ≤ 15, prior earnings $\leq \$19,670$). The EWM treatment rule that takes into account \$774 treatment assignment cost has the same earnings threshold, but lowers the education threshold to 12. The size of black dots indicates the number of individuals with different covariate values.

Second, we consider two classes of *linear treatment rules:*

$$\mathcal{G}_{LES} \equiv \left\{ \left\{ x : \beta_0 + \beta_1 \cdot \text{prior earnings} + \beta_2 \cdot \text{education} > 0 \right\}, \beta_0, \beta_1, \beta_2 \in \mathbb{R} \right\}, \tag{3.2}$$

$$\mathcal{G}^3_{LES} \equiv \left\{ \left\{ x : \begin{pmatrix} \beta_0 + \beta_1 \cdot \text{prior earnings} + \beta_2 \cdot \text{education} \\ +\beta_3 \cdot (\text{education})^2 + \beta_4 \cdot (\text{education})^3 \end{pmatrix} > 0 \right\}, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R} \right\}.$$

Linear treatment rules that maximize empirical welfare could markedly differ from plug-in rules derived from linear regressions. When treatment costs are not considered, the direction of treatment assignment as a function of prior earnings differs between the EWM rule (Figure 2) and the linear regression plug-in rule (Figure 3). When treatment costs are considered, the direction of treatment assignment is similar, but the two treatment rules substantially differ in the proportion of population assigned (69% by the EWM rule vs. 86% by the plug-in rule, Table 1). If a regression equation is correctly specified, the regression plug-in and EWM rules have identical large sample limits. If the regression equation is misspecified, however, only the linear EWM treatment rule converges with sample size to the welfare-maximizing limit, and the welfare level attained by the regression plug-in rule can be suboptimal even in large samples.

When quadratic and cubic terms for education are included, the EWM rule (Figure 4) uses the additional flexibility of this functional form to change the shape of the treatment assignment boundary, setting a higher prior earnings threshold for individuals with 13 years of education. It could be seen from Figure 6 that treatment effects estimated by nonparametric regression for these excluded individuals are also low. In comparison, the linear regression uses the additional terms to improve the global fit of the earnings equation and does not exclude this subset of the population for whom the treatment effect seems to be negative.

Figure 6 shows plug-in treatment rules based on kernel regressions of treatment and control outcomes on the covariates.[17] The class of nonparametric plug-in rules is much richer than the quadrant or the linear class of treatment rules, and it may obtain higher welfare in large samples. It is clear from the figure, however, that this class of patchy decision rules may be difficult to implement in public policy, where clear and transparent treatment rules are required.

[17] The bandwidths were chosen by Silverman's rule of thumb.

4 Conclusion

The EWM approach considered in this paper directly maximizes a sample analog of the welfare criterion of a policy maker. This welfare-function-based statistical procedure for treatment choice differs fundamentally from parametric and nonparametric plug-in approaches, which do not integrate statistical inference and the decision problem at hand. We investigated the statistical performances of the EWM rule in terms of the uniform convergence rate of the welfare loss and demonstrated that with known propensity scores, the EWM rule attains minimax optimal rates over various classes of feasible data distributions. The EWM approach offers a useful framework for the individualized policy assignment problems, as the EWM approach can easily accommodate the constraints that policy makers commonly face in reality. We also presented methods to compute the EWM rule for many practically important classes of treatment assignment rules and demonstrated them using experimental data from the JTPA program.

Several extensions and open questions remain to be answered. First, this paper assumed that the class of candidate policies $\mathcal{G}$ is given exogenously to the policy maker. We did not consider how to select the class $\mathcal{G}$ when the policy maker is free to do so. See Swaminathan and Joachims (2015) and Mbakop and Tabord-Meehan (2017) for recent developments. Second, while EWM attains minimax rate-optimality, it is unclear whether the EWM rule has stronger decision-theoretic optimality properties for the nonparametric class of data generating processes we considered. It remains to be seen whether EWM obtains the lowest asymptotic maximum regret within the class of minimax rate-optimal rules, whether it is admissible, and whether it is Bayes-optimal for some prior over P. It is an open question whether modifications of EWM or other rate-optimal rules could perform better. Third, we ruled out the case in which the data are subject to selection on unobservables or the overlap conditions fails. In these situations, the welfare criterion could be only set-identified, and it is not clear how to extend the EWM idea to this case. Fourth, we restricted our analysis to the additive social welfare criterion, but in some contexts, policy makers have a non-additive social welfare criterion. See Kitagawa and Tetenov (2017a) for an extension of the EWM approach to a class of generalized Gini social welfare functions.

Table 1: Estimated welfare gain of alternative treatment assignment rules that condition on education and pre-program earnings.

Outcome variable:	*30-month post-program earnings,* *no treatment cost*		*30-month post-program earnings,* *$774 cost for each assigned treatment*	
Treatment rule:	Share of population to be treated	Est. welfare gain per population member	Share of population to be treated	Est. welfare gain per population member
Treat everyone:	1	$1,180 ($464, $1,896)	1	$404 (-$313, $1,121)
EWM quadrant rule	0.95	$1,340 ($441, $2,239)	0.8	$643 (-$258, $1,544)
EWM linear rule	0.96	$1,364 ($398, $2,330)	0.69	$792 (-$177, $1,761)
EWM linear rule (with (education)^2 and (education)^3)	0.88	$1,489 ($374, $2,603)	0.75	$897 (-$214, $2,008)
Linear regression plug-in rule	0.98	$1,152	0.86	$527
Linear regression plug-in rule (with (education)^2 and (education)^3)	0.95	$1,263	0.91	$547
Nonparametric plug-in rule	0.91	$1,693	0.78	$996

Two-sided 95% confidence intervals in parentheses.

See Section B in Kitagawa and Tetenov (2017c) for their construction and asymptotic validity.

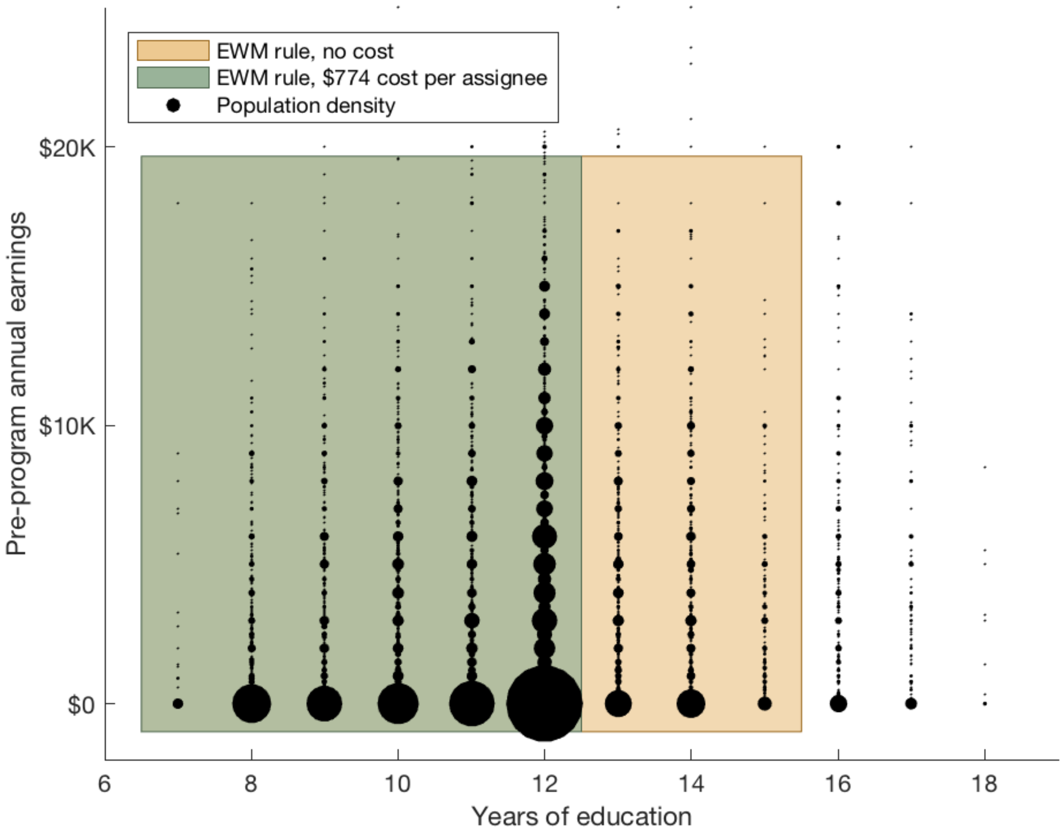

Figure 1: Empirical Welfare-Maximizing treatment rules from the quadrant class conditioning on years of education and pre-program earnings

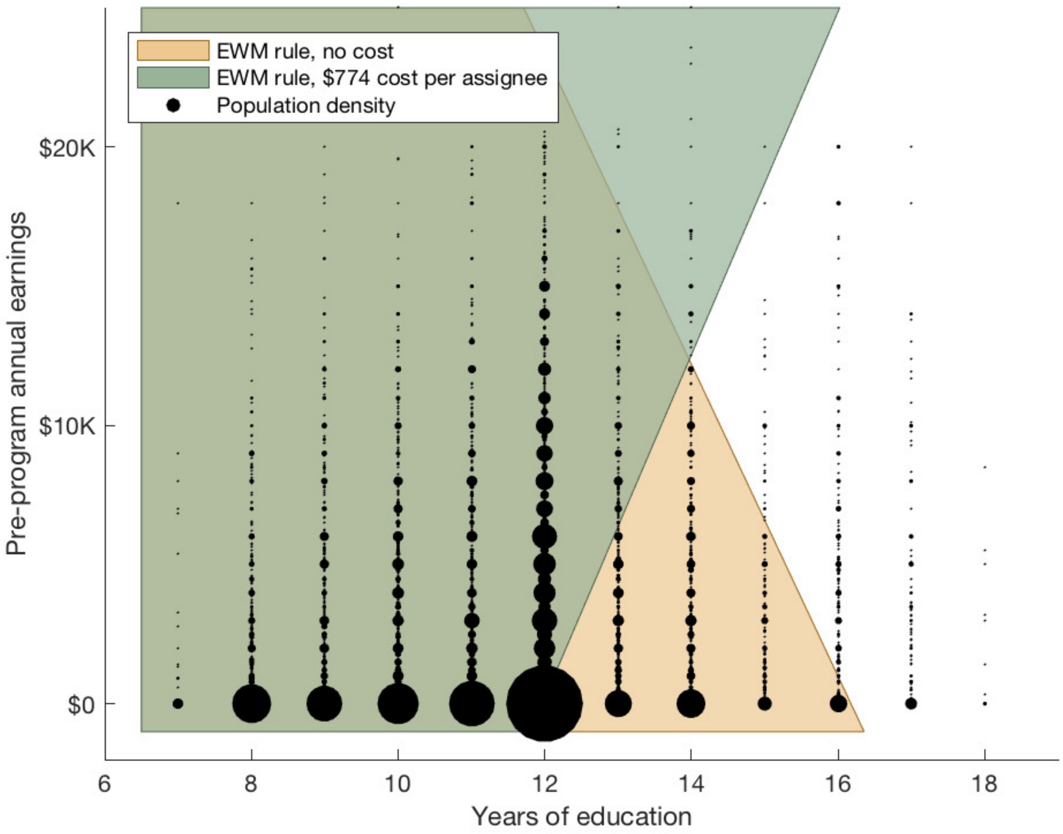

Figure 2: Empirical Welfare-Maximizing treatment rules from the linear class conditioning on years of education and pre-program earnings

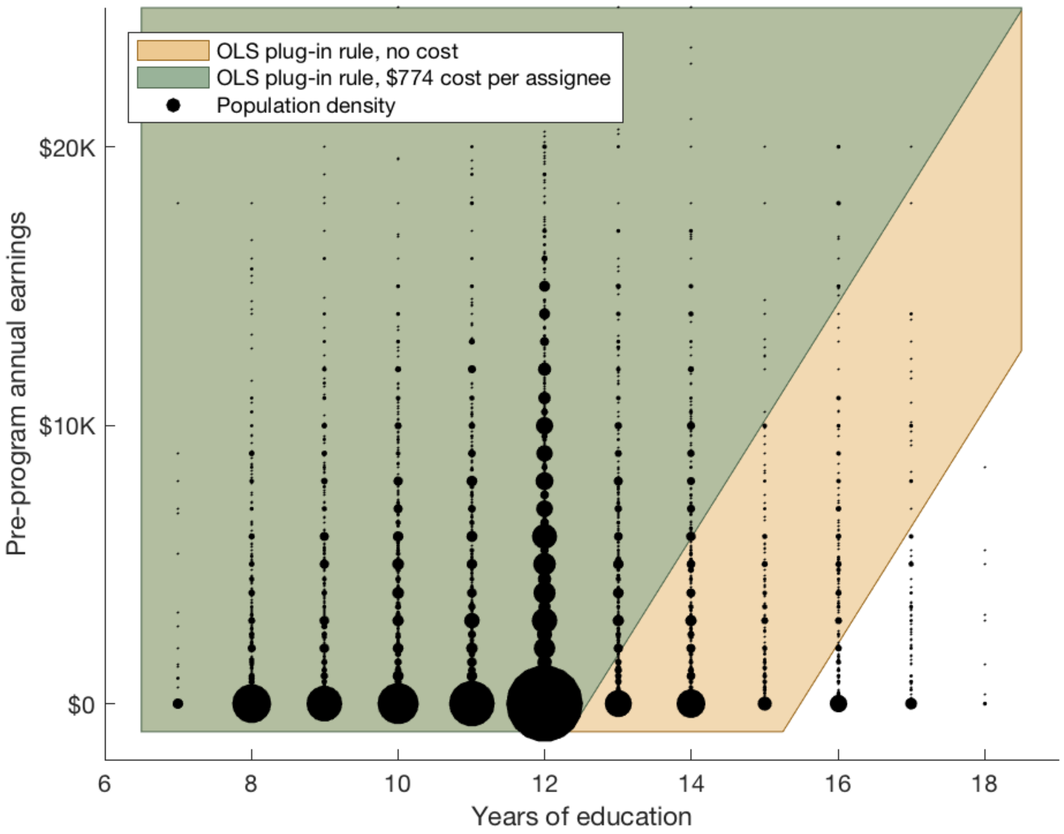

Figure 3: Parametric plug-in treatment rules based on the linear regressions of treatment outcomes on years of education and pre-program earnings

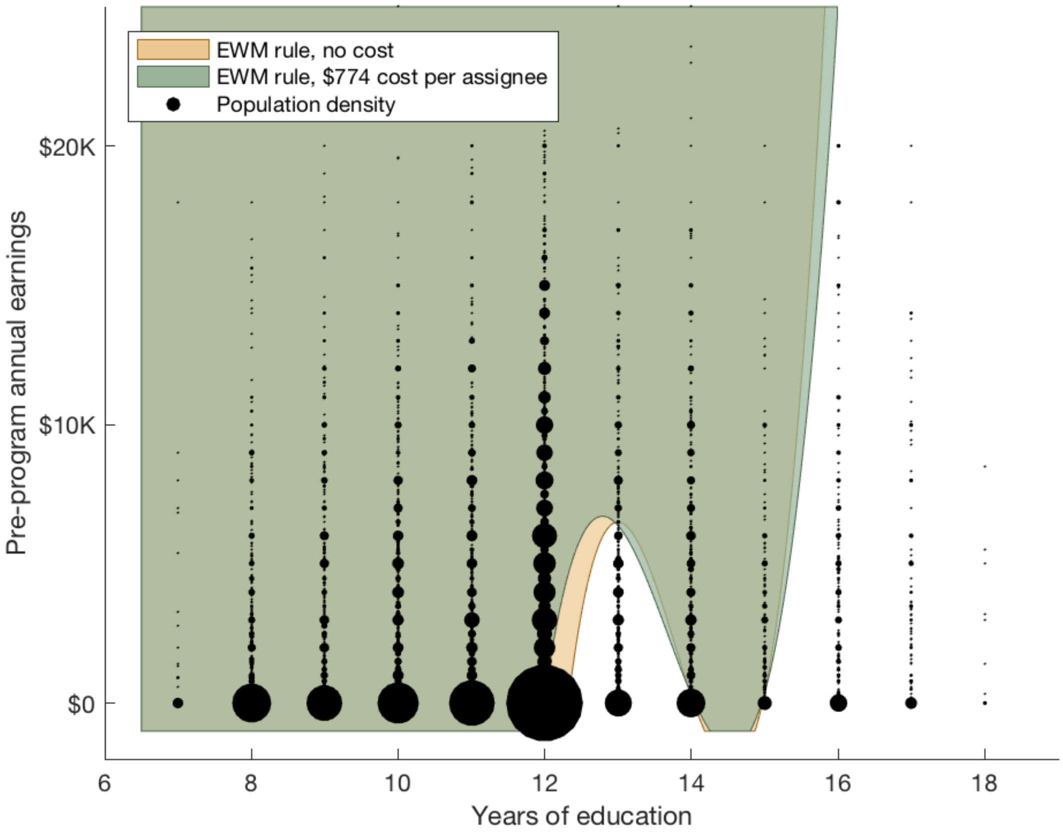

Figure 4: Empirical Welfare-Maximizing treatment rules from the linear class conditioning on years of education, $(\text{education})^2$, $(\text{education})^3$, and pre-program earnings

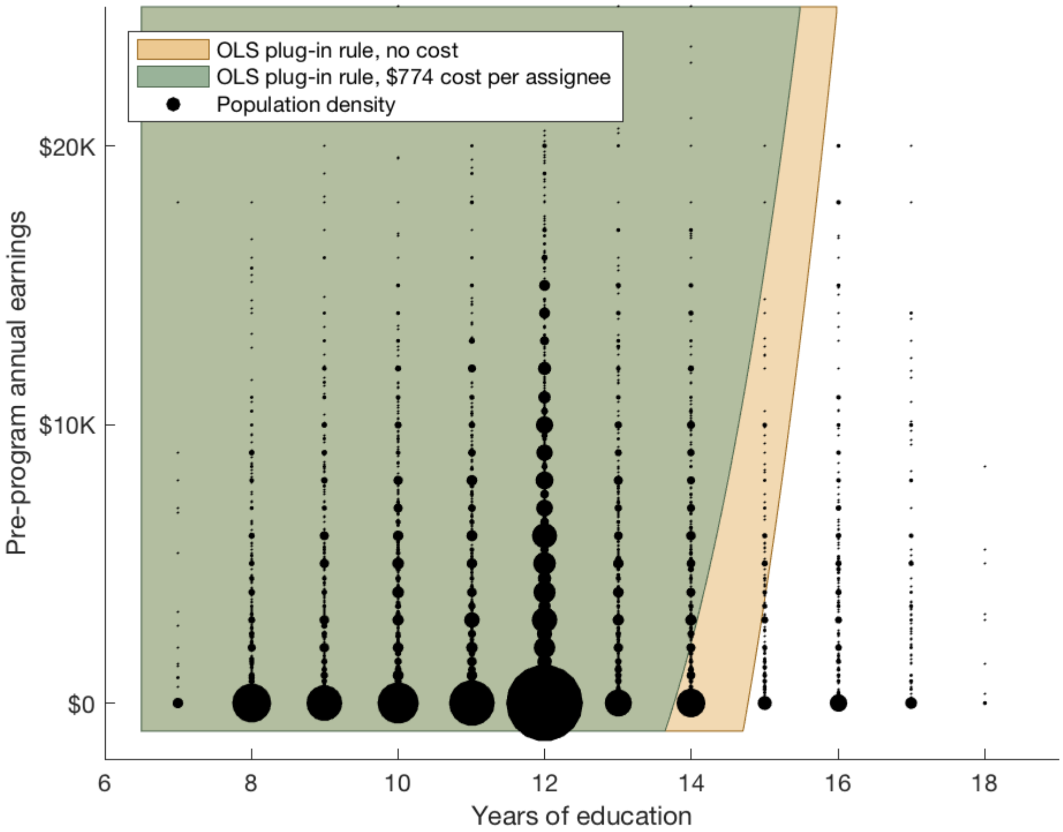

Figure 5: Parametric plug-in treatment rules based on the regressions of treatment outcomes on years of education, (education)^2, (education)^3, and pre-program earnings

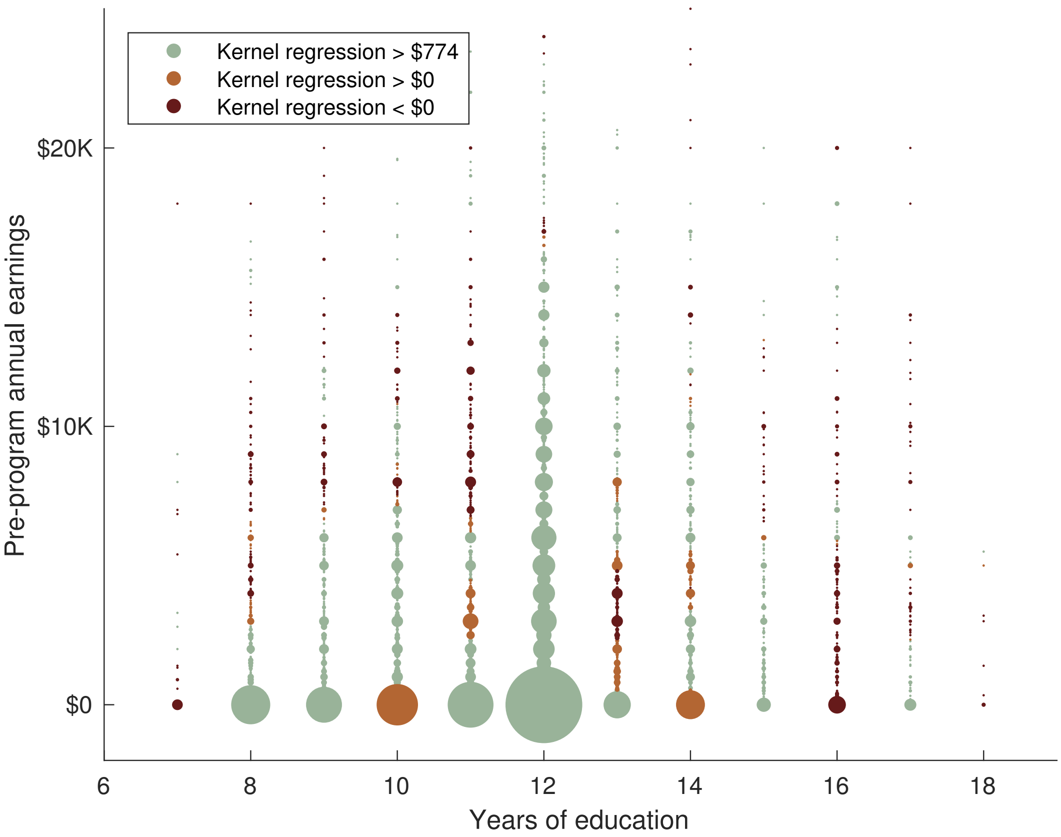

Figure 6: Nonparametric plug-in treatment rules based on the kernel regressions of treatment outcomes on years of education and pre-program earnings

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Abadie, A., M. M. Chingos, and M. R. West (2017): "Endogenous Stratification in Randomized Experiments," *unpublished manuscript.*

Armstrong, T. B. and S. Shen (2015): "Inference on Optimal Treatment Assignments," *unpublished manuscript.*

Athey, S. and S. Wager (2017): "Efficient Policy Learning," *arXiv preprint*, arXiv 1702.0289.

Atkinson, A. B. (1970): "On the measurement of inequality," *Journal of Economic Theory*, 2, 244–263.

Audibert, J.-Y. and A. B. Tsybakov (2007): "Fast Learning Rates for Plug-in Classifiers," *The Annals of Statistics*, 35, 608–633.

Beygelzimer, A. and J. Langford (2009): "The Offset Tree for Learning with Partial Labels," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 129–137.

Bhattacharya, D. and P. Dupas (2012): "Inferring Welfare Maximizing Treatment Assignment under Budget Constraints," *Journal of Econometrics*, 167, 168–196.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos (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, 549–576.

Chamberlain, G. (2011): "Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, Oxford University Press, 11–39.

Dehejia (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173.

Devroye, L., L. Györfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition*, Springer.

Devroye, L. and G. Lugosi (1995): “Lower Bounds in Pattern Recognition and Learning,” *Pattern Recognition*, 28, 1011–1018.

Dudík, M., J. Langford, and L. Li (2011): “Doubly Robust Policy Evaluation and Learning,” *Proceedings of the 28th International Conference on Machine Learning*, 1097–1104.

Dudley, R. (1999): *Uniform Central Limit Theorems*, Cambridge University Press.

Elliott, G. and R. P. Lieli (2013): “Predicting Binary Outcomes,” *Journal of Econometrics*, 174, 15–26.

Florios, K. and S. Skouras (2008): “Exact Computation of Max Weighted Score Estimators,” *Journal of Econometrics*, 146, 86–91.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): “Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *The Review of Economic Studies*, 64, 605–654.

Hirano, K. and J. R. Porter (2009): “Asymptotics for Statistical Treatment Rules,” *Econometrica*, 77, 1683–1701.

Kasy, M. (2016): “Partial identification, distributional preferences, and the welfare ranking of policies,” *Review of Economics and Statistics*, 98, 111–131.

——— (2017): “Optimal Taxation and Insurance Using Machine Learning,” *unpublished manuscript.*

Kerkyacharian, G., A. B. Tsybakov, V. Temlyakov, D. Picard, and V. Koltchinskii (2014): “Optimal Exponential Bounds on the Accuracy of Classification,” *Constructive Approximation*, 39, 421–444.

Kitagawa, T. and A. Tetenov (2017a): “Equality-Minded Treatment Choice,” *Cemmap working paper*, 10/17.

——— (2017b): “Online Appendix to ‘Who Should Be Treated? Empirical Welfare Maximization Method for Treatment Choice’,” *Unpublished manuscript.*

——— (2017c): “Supplement to ‘Who Should Be Treated? Empirical Welfare Maximization Method for Treatment Choice’,” *Unpublished manuscript.*

Lieli, R. P. and H. White (2010): "The Construction of Empirical Credit Scoring Rules Based on Maximization Principles," *Journal of Econometrics*, 157, 110–119.

Lugosi, G. (2002): "Pattern Classification and Learning Theory," in *Principles of Nonparametric Learning*, ed. by L. Györfi, Vienna: Springer, 1–56.

Mammen, E. and A. B. Tsybakov (1999): "Smooth Discrimination Analysis," *Annals of Statistics*, 27, 1808–1829.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

Manski, C. F. and T. Thompson (1989): "Estimation of Best Predictors of Binary Response," *Journal of Econometrics*, 40, 97–123.

Massart, P. and E. Nédélec (2006): "Risk Bounds for Statistical Learning," *The Annals of Statistics*, 34, 2326–2366.

Mbakop, E. and M. Tabord-Meehan (2017): "Model Selection for Treatment Choice: Penalized Welfare Maximization," *arXiv preprint*, arXiv 1609.03167.

Qian, M. and S. A. Murphy (2011): "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180–1210.

Stoye, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151, 70–81.

——— (2012): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 166, 138–156.

Swaminathan, A. and T. Joachims (2015): "Counterfactual Risk Minimization: Learning from Logged Bandit Feedback," *Journal of Machine Learning Research*, 16, 1731–1755.

Tetenov, A. (2012): "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 166, 157–165.

Tsybakov, A. B. (2004): "Optimal Aggregation of Classifiers in Statistical Learning," *The Annals of Statistics*, 32, 135–166.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Vapnik, V. N. (1998): *Statistical Learning Theory*, John Wiley & Sons.

Zadrozny, B. (2003): "Policy Mining: Learning Decision Policies from Fixed Sets of Data," *Ph.D Thesis, University of California, San Diego.*

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012): "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012): "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118.

Supplement to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice"

Toru Kitagawa[*] and Aleksey Tetenov[†]

November 27, 2017

A Lemmas and Proofs

A.1 Notations and Basic Lemmas

Let $Z_i = (Y_i, D_i, X_i) \in \mathcal{Z}$. The *subgraph* of a real-valued function $f : \mathcal{Z} \mapsto \mathbb{R}$ is the set

$$SG(f) \equiv \{(z,t) \in \mathcal{Z} \times \mathbb{R} : 0 \leq t \leq f(z) \text{ or } f(z) \leq t \leq 0\}.$$

The following lemma establishes a link between the VC-dimension of a class of subsets in the covariate space $\mathcal{X}$ and the VC-dimension of a class of subgraphs of functions on $\mathcal{Z} = \mathbb{R} \times \{0,1\} \times \mathcal{X}$ (their subgraphs will be in $\mathcal{Z} \times \mathbb{R}$).

Lemma A.1. *Let $\mathcal{G}$ be a VC-class of subsets of $\mathcal{X}$ with VC-dimension $v < \infty$. Let g and h be two given functions from $\mathcal{Z}$ to $\mathbb{R}$. Then the set of functions from $\mathcal{Z}$ to $\mathbb{R}$*

$$\mathcal{F} = \{f_G(z) = g(z) \cdot 1\{x \in G\} + h(z) 1\{x \notin G\} : G \in \mathcal{G}\}$$

is a VC-subgraph class of functions with VC-dimension less than or equal to v.

Proof. Let $z_i = (y_i, d_i, x_i)$. By the assumption, no set of $(v+1)$ points in $\mathcal{X}$ could be shattered by $\mathcal{G}$. Take an arbitrary set of $(v+1)$ points in $\mathcal{Z} \times \mathbb{R}$, $A = \{(z_1, t_1), ..., (z_{v+1}, t_{v+1})\}$. Denote the collection of subgraphs of $\mathcal{F}$ by $SG(\mathcal{F}) \equiv \{\mathrm{SG}(f_G), G \in \mathcal{G}\}$. We want to show that $SG(\mathcal{F})$ doesn't shatter A.

If for some $i \in \{1, \ldots, (v+1)\}$, $(z_i, t_i) \in \mathrm{SG}(g) \cap \mathrm{SG}(h)$ then $SG(\mathcal{F})$ cannot pick out all of the subsets of A because the i-th point is included in any $S \in SG(\mathcal{F})$. Similarly, if for some $i \in \{1, \ldots, (v+1)\}$, $(z_i, t_i) \in \mathrm{SG}(g)^c \cap \mathrm{SG}(h)^c$, then point i cannot be included in any $S \in SG(\mathcal{F})$.

The remaining case is that, for each i, either $(z_i, t_i) \in \mathrm{SG}(g) \cap \mathrm{SG}(h)^c$ or $(z_i, t_i) \in \mathrm{SG}(g)^c \cap \mathrm{SG}(h)$ holds. Indicate the former case by $\delta_i = 0$ and the latter case by $\delta_i = 1$. The points with $\delta_i = 0$ could be picked by $\mathrm{SG}(f_G)$ if and only if $x_i \notin G$. The points with $\delta_i = 1$ could be picked if and

[*] Cemmap/University College London, Department of Economics. Email: t.kitagawa@ucl.ac.uk

[†] University of Bristol, Email: a.tetenov@bristol.ac.uk

only if $x_i \in G$. Given that $\mathcal{G}$ is a VC-class with VC-dimension v, there exists a subset X_0 of $\{x_1, \ldots, x_{v+1}\}$ such that $X_0 \neq (\{x_1, \ldots, x_{v+1}\} \cap G)$ for any $G \in \mathcal{G}$. Then there could be no set $S \in SG(\mathcal{F})$ that picks out the set (possibly empty)

$$\{(z_i, t_i) : (x_i \in X_0 \text{ and } \delta_i = 1) \text{ or } (x_i \notin X_0 \text{ and } \delta_i = 0)\}, \tag{A.1}$$

because this set of points could only be picked out by $\mathrm{SG}(f_G)$ if $(\{x_1, \ldots, x_{v+1}\} \cap G) = X_0$. Hence, $\mathcal{F}$ is a VC subgraph class of functions with VC-dimension less than or equal to v. □

In addition to the notations introduced in the main text, the following notations are used throughout the supplementary material. The empirical probability distribution based on an iid size n sample of $Z_i = (Y_i, D_i, X_i)$ is denoted by P^n. $L_2(P)$ metric for f is denoted by $\|f\|_{L_2(P)} = \left[\int_{\mathcal{Z}} f^2 dP\right]^{1/2}$, and the sup-metric of f is denoted by $\|f\|_\infty$. Positive constants that only depend on the class of data generating processes, not on the sample size nor the VC-dimension, are denoted by $c_1, c_2, c_3, \ldots$. The universal constants are denoted by the capital letter $C_1, C_2, C_3, \ldots$.

In what follows, we present lemmas that will be used in the proofs of Theorems 2.1 and 2.3. Lemmas A.2 and A.3 are classical inequalities whose proofs can be found, for instance, in Lugosi (2002).

Lemma A.2. *Hoeffding's Lemma: let X be a random variable with $EX = 0$, $a \leq X \leq b$. Then, for $s > 0$,*

$$E\left(e^{sX}\right) \leq e^{s^2(b-a)^2/8}.$$

Lemma A.3. *Let $\lambda > 0$, $n \geq 2$, and let $Y_1, \ldots, Y_n$ be real-valued random variables such that for all $s > 0$ and $1 \leq i \leq n$, $E(e^{sY_i}) \leq e^{s^2\lambda^2/2}$ holds. Then,*

$$\begin{aligned} &(i) \quad E\left(\max_{i \leq n} Y_i\right) &\leq&\quad \lambda\sqrt{2\ln n}, \\ &(ii) \quad E(\max_{i \leq n} |Y_i|) &\leq&\quad \lambda\sqrt{2\ln(2n)}. \end{aligned}$$

The next two lemmas give maximal inequalities that bound the mean of a supremum of centered empirical processes indexed by a VC-subgraph class of functions. The first maximal inequality (Lemma A.4) is standard in the empirical process literature, and it yields our Theorem 2.1 as a corollary. Though its proof can be found elsewhere (e.g., Dudley (1999), van der Vaart and Wellner (1996)), we present it here for the sake of completeness and for later reference in the proof of Lemma A.5. The second maximal inequality (Lemma A.5) concerns the class of functions whose diameter is constrained by the $L_2(P)$-norm. Lemma A.5 will be used in the proof of Theorem 2.3. A lemma similar to our Lemma A.5 appears in Massart and Nédélec (2006, Lemma A.3).

Lemma A.4. *Let $\mathcal{F}$ be a class of uniformly bounded functions, i.e., there exists $\bar{F} < \infty$ such that $\|f\|_\infty \leq \bar{F}$ for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a VC-subgraph class with VC-dimension $v < \infty$. Then, there is a universal constant C_1 such that*

$$E_{P^n}\left[\sup_{f\in\mathcal{F}} |E_n(f) - E_P(f)|\right] \leq C_1\bar{F}\sqrt{\frac{v}{n}}$$

holds for all $n \geq 1$.

Proof. Introduce $(Z_1', \ldots, Z_n')$, an independent copy of $(Z_1, \ldots, Z_n) \sim P^n$. We denote the probability law of $(Z_1', \ldots, Z_n')$ by $P^{n'}$, its expectation by $E_{P^{n'}}(\cdot)$, and the sample average with respect to $(Z_1', \ldots, Z_n')$ by $E_n'(\cdot)$. Define iid Rademacher variables $\sigma^n \equiv (\sigma_1, \ldots, \sigma_n)$ such that $\Pr(\sigma_1 = -1) = \Pr(\sigma_1 = 1) = 1/2$ and they are independent of $Z_1, Z_1', \ldots, Z_n, Z_n'$. Then,

$$\begin{aligned}
E_{P^n}\left[\sup_{f\in\mathcal{F}} |E_n(f) - E_P(f)|\right] &= E_{P^n}\left[\sup_{f\in\mathcal{F}} \left|E_{P^{n'}}\left[E_n(f) - E_n'(f)|Z_1, \ldots, Z_n\right]\right|\right] \\
&\leq E_{P^n}\left[\sup_{f\in\mathcal{F}} E_{P^{n'}}\left[\left|E_n(f) - E_n'(f)\right| |Z_1, \ldots, Z_n\right]\right] \\
(&\because \text{ Jensen's inequality}) \\
&\leq E_{P^n, P^{n'}}\left[\sup_{f\in\mathcal{F}} \left|E_n(f) - E_n'(f)\right|\right] \\
&= \frac{1}{n} E_{P^n, P^{n'}}\left\{\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \left(f(Z_i) - f(Z_i')\right)\right|\right\} \\
&= \frac{1}{n} E_{P^n, P^{n'}, \sigma^n}\left\{\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \sigma_i\left(f(Z_i) - f(Z_i')\right)\right|\right\} \\
(&\because \ f(Z_i) - f(Z_i') \sim \sigma_i\left(f(Z_i) - f(Z_i')\right) \text{ for all } i \ ) \\
&\leq \frac{1}{n} \ E_{P^n, P^{n'}, \sigma^n}\left\{\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \sigma_i f(Z_i)\right| + \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \sigma_i f(Z_i')\right|\right\} \\
&= \frac{2}{n} E_{P^n, \sigma^n}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \sigma_i f(Z_i)\right|\right] \\
&= \frac{2}{n} E_{P^n}\left\{E_{\sigma^n}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \sigma_i f(Z_i)\right| |Z_1, \ldots, Z_n\right]\right\}. \qquad \text{(A.2)}
\end{aligned}$$

Fix $Z_1, \ldots, Z_n$, and define $\mathbf{f} \equiv (f(Z_1), \ldots, f(Z_n)) = (f_1, \ldots, f_n)$, which is a vector of length n collecting the value of $f \in \mathcal{F}$ evaluated at each of $(Z_1, \ldots, Z_n)$. Let $\mathbf{F} \equiv \{\mathbf{f} : f \in \mathcal{F}\} \subset \mathbb{R}^n$, which is a bounded set in $\mathbb{R}^n$ with radius $\bar{F}$, since $\mathcal{F}$ is the set of uniformly bounded functions with

$|f(\cdot)| \leq \bar{F}$. Introduce the Euclidean norm to $\mathbf{F}$,

$$\rho(\mathbf{f},\mathbf{f}') = \left(\frac{1}{n}\sum_{i=1}^{n}\left(f_i - f_i'\right)^2\right)^{1/2}.$$

Let $\mathbf{f}^{(0)} = (0,\ldots,0)$, and $\mathbf{f}^* = (f_1^*,\ldots,f_n^*)$ be a random element in $\mathbf{F}$ maximizing $|\sum_{i=1}^{n}\sigma_i f_i|$. Let $B_0 = \left\{\mathbf{f}^{(0)}\right\}$ and construct $\left\{B_k : k = 1,\ldots,\bar{K}\right\}$ a sequence of covers of $\mathbf{F}$, such that $B_k \subset \mathbf{F}$ is a minimal cover with radius $2^{-k}\bar{F}$ and $B_{\bar{K}} = \mathbf{F}$. Note that such $\bar{K} < \infty$ exists at given n and $(Z_1,\ldots,Z_n)$. Define also $\left\{\mathbf{f}^{(k)} \in B_k : k = 1,\ldots,\bar{K}\right\}$ be a random sequence such that $\mathbf{f}^{(k)} \in \arg\min_{\mathbf{f}\in\mathbf{B}_k}\rho\left(\mathbf{f},\mathbf{f}^*\right)$. Since B_k is a cover with radius $2^{-k}\bar{F}$, $\rho\left(\mathbf{f}^{(k)},\mathbf{f}^*\right) \leq 2^{-k}\bar{F}$ holds. In addition, we have

$$\rho\left(\mathbf{f}^{(k-1)},\mathbf{f}^{(k)}\right) \leq \rho\left(\mathbf{f}^{(k)},\mathbf{f}^*\right) + \rho\left(\mathbf{f}^{(k-1)},\mathbf{f}^*\right) \leq 3\cdot 2^{-k}\bar{F}.$$

By a telescope sum,

$$\sum_{i=1}^{n}\sigma_i f_i^* = \sum_{i=1}^{n}\sigma_i f_i^{(0)} + \sum_{k=1}^{\bar{K}}\sum_{i=1}^{n}\sigma_i\left(f_i^{(k)} - f_i^{(k-1)}\right) = \sum_{k=1}^{\bar{K}}\sum_{i=1}^{n}\sigma_i\left(f_i^{(k)} - f_i^{(k-1)}\right).$$

We hence obtain

$$\begin{aligned} E_{\sigma^n}\left|\sum_{i=1}^{n}\sigma_i f_i^*\right| &\leq \sum_{k=1}^{\bar{K}} E_{\sigma^n}\left|\sum_{i=1}^{n}\sigma_i\left(f_i^{(k)} - f_i^{(k-1)}\right)\right| \\ &\leq \sum_{k=1}^{\bar{K}} E_{\sigma^n}\max_{\mathbf{f}\in B_k,\mathbf{g}\in B_{k-1}:\rho(\mathbf{f},\mathbf{g})\leq 3\cdot 2^{-k}\bar{F}}\left|\sum_{i=1}^{n}\sigma_i\left(f_i - g_i\right)\right|. \qquad \text{(A.3)} \end{aligned}$$

We apply Lemma A.2 to obtain

$$\begin{aligned} E_{\sigma^n}\left(e^{s\sum_{i=1}^{n}\sigma_i(f_i-g_i)}\right) &= \prod_{i=1}^{n} E_{\sigma_i}\left[e^{s\sigma_i(f_i-g_i)}\right] \leq \prod_{i=1}^{n} e^{s^2(f_i-g_i)^2/2} = \exp\left(s^2 n\rho^2(\mathbf{f},\mathbf{g})/2\right) \\ &\leq \exp\left(s^2 n(3\cdot 2^{-k}\bar{F})^2/2\right). \end{aligned}$$

An application of Lemma A.3 (ii) with $\lambda = 3\sqrt{n}\cdot 2^{-k}\bar{F}$ and $n = |B_k|\,|B_{k-1}| \leq |B_k|^2$ then yields

$$\begin{aligned} E_{\sigma^n}\max_{\mathbf{f}\in B_k,\mathbf{g}\in B_{k-1}:\rho(\mathbf{f},\mathbf{g})\leq 3\cdot 2^{-k}\bar{F}}\left|\sum_{i=1}^{n}\sigma_i\left(f_i - g_i\right)\right| &\leq 3\sqrt{n}\cdot 2^{-k}\bar{F}\sqrt{2\ln 2\,|B_k|^2} \\ &= 3\sqrt{n}\cdot 2^{-k}\bar{F}\sqrt{2\ln 2N(2^{-k}\bar{F},\mathbf{F},\rho)^2} \\ &= 6\sqrt{n}\cdot 2^{-k}\bar{F}\sqrt{\ln 2^{1/2}N(2^{-k}\bar{F},\mathbf{F},\rho)}, \end{aligned}$$

where $N(r, \mathbf{F},\rho)$ is the covering number of $\mathbf{F}$ with radius r in terms of norm ρ. Accordingly,

$$\begin{aligned} E_{\sigma^n}\left|\sum_{i=1}^n \sigma_i f_i^*\right| &\leq \sum_{k=1}^{\bar{K}} 6\sqrt{n}\cdot 2^{-k}\bar{F}\sqrt{\ln 2^{1/2}N(2^{-k}\bar{F}, \mathbf{F},\rho)} \\ &\leq 12\sqrt{n}\sum_{k=1}^{\infty} 2^{-(k+1)}\bar{F}\sqrt{\ln 2^{1/2}N(2^{-k}\bar{F}, \mathbf{F},\rho)} \\ &\leq 12\sqrt{n}\int_0^1 \bar{F}\sqrt{\ln 2^{1/2}N(\epsilon\bar{F}, \mathbf{F},\rho)}d\epsilon, \end{aligned} \tag{A.4}$$

where the last line follows from the fact that $N(\epsilon\bar{F}, \mathbf{F},\rho)$ is decreasing in ϵ.

To bound (A.4) from above, we apply a uniform entropy bound for the covering number. In Theorem 2.6.7 of van der Vaart and Wellner (1996), by setting $r = 2$ and Q at the empirical probability measure of $(Z_1, \ldots, Z_n)$, we have,

$$N(\epsilon\bar{F}, \mathbf{F}, \rho) \leq K(v+1)(16e)^{(v+1)}\left(\frac{1}{\epsilon}\right)^{2v}, \tag{A.5}$$

where $K > 0$ is a universal constant. Plugging this into (A.4) leads to

$$\begin{aligned} E_{\sigma}\left|\sum_{i=1}^n \sigma_i f_i^*\right| &\leq 12\bar{F}\sqrt{n}\int_0^1 \sqrt{\ln(2^{1/2}K) + \ln(v+1) + (v+1)\ln(16e) - 2v\ln\epsilon d\epsilon} \\ &\leq 12\bar{F}\sqrt{nv}\int_0^1 \sqrt{\ln(2^{1/2}K) + \ln 2 + 2\ln(16e) - 2\ln\epsilon d\epsilon} \\ &= C'\bar{F}\sqrt{nv}, \end{aligned} \tag{A.6}$$

where $C' = 12\int_0^1 \sqrt{\ln(2^{1/2}K) + \ln 2 + 2\ln(16e) - 2\ln\epsilon d\epsilon} < \infty$. Combining (A.6) with (A.2) and setting $C_1 = 2C'$ lead to the conclusion. □

Lemma A.5. *Let $\mathcal{F}$ be a class of uniformly bounded functions with $\|f\|_\infty \leq \bar{F} < \infty$ for all $f \in \mathcal{F}$. Assume that $\mathcal{F}$ is a VC-subgraph class with VC-dimension $v < \infty$. Assume further that* $\sup_{f\in\mathcal{F}}\|f\|_{L_2(P)} \leq \delta$. *Then, there exists a positive universal constant C_2 such that*

$$E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n(f) - E_P(f)\right)\right] \leq C_2\delta\sqrt{\frac{v}{n}}$$

holds for all $n \geq C_1^2\bar{F}^2 v/\delta^2$, where C_1 is the universal constant defined in Lemma A.4.

Proof. By the same symmetrization argument and the same use of Rademacher variables as in the proof of Lemma A.4, we have

$$E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n(f) - E_P(f)\right)\right] \leq \frac{2}{n}E_{P^n}\left\{E_{\sigma^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(Z_i)|Z_1, \ldots, Z_n\right]\right\}. \tag{A.7}$$

Fix the values of $Z_1,\ldots,Z_n$, and define $\mathbf{f}$, $\mathbf{f}^{(0)}$, $\mathbf{F}$, and norm $\rho(\mathbf{f},\mathbf{f}')$ as in the proof of Lemma A.4. Let $\mathbf{f}^*$ be a maximizer of $\sum_{i=1}^n \sigma_i f(Z_i)$ in $\mathbf{F}$ and let $\delta_n = \sup_{\mathbf{f}\in\mathbf{F}} \rho(\mathbf{f}^{(0)},\mathbf{f}) \leq \bar{F}$. Let $B_0 = \left\{\mathbf{f}^{(0)}\right\}$ and construct $\left\{B_k : k = 1,\ldots,\bar{K}\right\}$ a sequence of covers of $\mathbf{F}$, such that $B_k \subset \mathbf{F}$ is a minimal cover with radius $2^{-k}\delta_n$ and $B_{\bar{K}} = \mathbf{F}$. We define $\left\{\mathbf{f}^{(k)} \in B_k : k = 1,\ldots,\bar{K}\right\}$ to be a random sequence such that $\mathbf{f}^{(k)} \in \arg\min_{\mathbf{f}\in\mathbf{B}_k} \rho\left(\mathbf{f},\mathbf{f}^*\right)$. By applying the chaining argument in the proof of Lemma A.4, Lemma A.3 (i), and the uniform bound of the covering number (A.5), we obtain

$$E_\sigma \sum_{i=1}^n \sigma_i f_i^* \;\leq\; 12\sqrt{n}\int_0^1 \delta_n \sqrt{\log N(\epsilon\delta_n,\mathbf{F},\rho)}d\epsilon \;\leq\; 2^{-1}C_1\delta_n\sqrt{nv}.$$

for the universal constant C_1 defined in the proof of Lemma A.4. Hence, from (A.7), we have

$$\begin{aligned} E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n\left(f\right)-E_P(f)\right)\right] &\leq C_1\sqrt{\frac{v}{n}}E_{P^n}\left(\delta_n\right) \;=\; C_1\sqrt{\frac{v}{n}}E_{P^n}\left(\left[\sup_{f\in\mathcal{F}} E_n\left(f^2\right)\right]^{1/2}\right) \\ &\leq C_1\sqrt{\frac{v}{n}}\left[E_{P^n}\left(\sup_{f\in\mathcal{F}} E_n\left(f^2\right)\right)\right]^{1/2}. \qquad \text{(A.8)} \end{aligned}$$

Note that $E_n\left(f^2\right)$ is bounded by

$$\begin{aligned} E_n\left(f^2\right) &= E_n\left(f^2 - E_P\left(f^2\right)\right) + E_P(f^2) \\ &= E_n\left[\left(f - \|f\|_{L_2(P)}\right)\left(f + \|f\|_{L_2(P)}\right)\right] + \|f\|_{L_2(P)}^2 \\ &\leq 2\bar{F}E_n\left[f - \|f\|_{L_2(P)}\right] + \|f\|_{L_2(P)}^2 \\ &\leq 2\bar{F}E_n\left[f - E_P(f)\right] + \|f\|_{L_2(P)}^2 \\ &(\;\because\;\; \|f\|_{L_2(P)} \geq E_P(f) \text{ by the Cauchy-Schwartz inequality.}) \end{aligned}$$

Combining this inequality with (A.8) yields

$$E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n\left(f\right)-E_P(f)\right)\right] \leq C_1\sqrt{\frac{v}{n}}\sqrt{2\bar{F}E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n\left(f\right)-E_P(f)\right)\right]+\delta^2}.$$

Solving this inequality for $E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n\left(f\right)-E_P(f)\right)\right]$ leads to

$$E_{P^n}\left[\sup_{f\in\mathcal{F}}\left(E_n\left(f\right)-E_P(f)\right)\right] \leq \bar{F}C_1^2\sqrt{\frac{v}{n}}\left(\sqrt{\frac{v}{n}}+\sqrt{\frac{v}{n}+\frac{\delta^2}{\bar{F}^2C_1^2}}\right).$$

For $\frac{v}{n} \leq \frac{\delta^2}{\bar{F}^2C_1^2}$, that is, $n \geq \frac{C_1^2\bar{F}^2v}{\delta^2}$, the upper bound can be further bounded by $(1+\sqrt{2})C_1\delta\sqrt{\frac{v}{n}}$, so the conclusion of the lemma follows with $C_2 = (1+\sqrt{2})C_1$. □

A.2 Proofs of Theorems 2.1 and 2.2

Proof of Theorem 2.1. Define

$$f(Z_i;G)=\left[\frac{Y_iD_i}{e(X_i)}\cdot 1\left\{X_i\in G\right\}+\frac{Y_i(1-D_i)}{1-e(X_i)}\cdot 1\left\{X_i\notin G\right\}\right],$$

and the class of functions on $\mathcal{Z}$

$$\mathcal{F}=\left\{f(\cdot;G):G\in\mathcal{G}\right\}.$$

With these notations, we can express inequality (2.3) in the main text as

$$W^*_{\mathcal{G}}-W(\hat{G}_{EWM})\leq 2\sup_{f\in\mathcal{F}}\left|E_n(f)-E_P(f)\right|. \tag{A.9}$$

Note that Assumption 2.1 (BO) and (SO) imply that $\mathcal{F}$ has uniform envelope $\bar{F}=M/(2\kappa)$. Also, by Assumption 2.1 (VC) and Lemma A.1, $\mathcal{F}$ is a VC-subgraph class of functions with VC-dimension at most v. We apply Lemma A.4 to (A.9) to obtain

$$E_{P^n}\left[W^*_{\mathcal{G}}-W(\hat{G}_{EWM})\right]\leq C_1\frac{M}{\kappa}\sqrt{\frac{v}{n}}.$$

Since this upper bound does not depend on $P\in\mathcal{P}(M,\kappa)$, the upper bound is uniform over $\mathcal{P}(M,\kappa)$.

□

Proof of Theorem 2.2. In obtaining the rate lower bound, we normalize the support of outcomes to $Y_{1,i},Y_{0,i}\in\left[-\frac{1}{2},\frac{1}{2}\right]$. That is, we focus on bounding $\sup_{P\in\mathcal{P}(1,\kappa)}E_{P^n}\left[W^*_{\mathcal{G}}-W(G_n)\right]$. The lower bound of the original welfare loss $\sup_{P\in\mathcal{P}(M,\kappa)}E_{P^n}\left[W^*_{\mathcal{G}}-W(G_n)\right]$ is obtained by multiplying by M the lower bound of $\sup_{P\in\mathcal{P}(1,\kappa)}E_{P^n}\left[W^*_{\mathcal{G}}-W(G_n)\right]$.

We consider a suitable subclass $\mathcal{P}^*\subset\mathcal{P}\left(1,\kappa\right)$, for which the worst case welfare loss can be bounded from below by a distribution-free term that converges at rate $n^{-1/2}$. The construction of $\mathcal{P}^*$ proceeds as follows. First, let $x_1,\dots,x_v\in\mathcal{X}$ be v points that are shattered by $\mathcal{G}$. We constrain P_X (the marginal distribution of X) to being supported only on $(x_1,\dots,x_v)$. We put the equal mass $1/v$ at x_i, $i\leq v$. Thus-constructed marginal distribution of X is common in $\mathcal{P}^*$. Let the distribution of treatment indicator D be independent of (Y_1,Y_0,X), and D follows the Bernoulli distribution with $\Pr(D=1)=1/2$. Let $\mathbf{b}=(b_1,\dots,b_v)\in\{0,1\}^v$ be a bit vector used to index a member of $\mathcal{P}^*$, i.e., $\mathcal{P}^*$ consists of a finite number of DGPs. For each $j=1,\dots,v$, and depending on $\mathbf{b}$, construct the following conditional distribution of Y_1 given $X=x_j$: if $b_j=1$,

$$Y_1=\begin{cases}\frac{1}{2} & \text{with prob. } \frac{1}{2}+\gamma,\\ -\frac{1}{2} & \text{with prob. } \frac{1}{2}-\gamma,\end{cases} \tag{A.10}$$

and, if $b_j = 0$,

$$Y_1 = \begin{cases} \frac{1}{2} & \text{with prob. } \frac{1}{2} - \gamma, \\ -\frac{1}{2} & \text{with prob. } \frac{1}{2} + \gamma, \end{cases} \tag{A.11}$$

where $\gamma \in \left[0, \frac{1}{2}\right]$ is chosen properly in a later step of the proof. As for Y_0's conditional distribution, we consider the degenerate distribution at $Y_0 = 0$ at every $X = x_j$, $j = 1, \ldots, v$. That is, when $b_j = 1$, $\tau(x_j) = \gamma$, and when $b_j = 0$, $\tau(x_j) = -\gamma$. For each $\mathbf{b} \in \{0,1\}^v$, $P_{\mathbf{b}} \in \mathcal{P}(1, \kappa)$ clearly holds. We accordingly define a sublass of $\mathcal{P}(1, \kappa)$ by $\mathcal{P}^* = \{P_{\mathbf{b}} : \mathbf{b} \in \{0,1\}^v\}$.

With knowledge of $P_{\mathbf{b}} \in \mathcal{P}^*$, the optimal treatment assignment rule is

$$G^*_{\mathbf{b}} = \{x_j : b_j = 1, j \leq v\},$$

which is feasible $G^*_{\mathbf{b}} \in \mathcal{G}$ by the construction of the support points of X. The maximized social welfare is

$$W(G^*_{\mathbf{b}}) = v^{-1}\gamma \left(\sum_{j=1}^{v} b_j\right).$$

Let $\hat{G}$ be an arbitrary treatment choice rule depending on sample $(Z_1, \ldots, Z_n)$, and $\hat{\mathbf{b}} \in \{0,1\}^v$ be a binary vector whose j-th element is $\hat{b}_j = 1\left\{x_j \in \hat{G}\right\}$. Consider $\pi(\mathbf{b})$ a prior distribution for $\mathbf{b}$ such that $b_1, \ldots, b_v$ are iid and $b_1 \sim Ber(1/2)$. The welfare loss satisfies the following inequalities:

$$\begin{aligned} \sup_{P \in \mathcal{P}(1,\kappa)} E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G})\right] &\geq \sup_{P_{\mathbf{b}} \in \mathcal{P}^*} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\hat{G})\right] \\ &\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\hat{G})\right] d\pi(\mathbf{b}) \\ &= \gamma \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[P_X\left(G^*_{\mathbf{b}} \triangle \hat{G}\right)\right] d\pi(\mathbf{b}) \\ &= \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\left\{b(X) \neq \hat{b}(X)\right\}\right) dP^n_{\mathbf{b}}(Z_1, \ldots, Z_n)\, d\pi(\mathbf{b}) \\ &\geq \inf_{\hat{G}} \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\left\{b(X) \neq \hat{b}(X)\right\}\right) dP^n_{\mathbf{b}}(Z_1, \ldots, Z_n)\, d\pi(\mathbf{b}) \end{aligned}$$

where $b(X)$ and $\hat{b}(X)$ are elements of $\mathbf{b}$ and $\hat{\mathbf{b}}$, respectively, such that $b(x_j) = b_j$ and $\hat{b}(x_j) = \hat{b}_j$. Note that the infimum over assignment rules $\hat{G}$ can be seen as the minimization problem of the Bayes risk with the loss function corresponding to the classification error for predicting binary random variable $b(X)$. Hence, a minimizer of the Bayes risk is attained by the Bayes classifier,

$$\hat{G}^* = \left\{x_j : \pi(b_j = 1 | Z_1, \ldots, Z_n) \geq \frac{1}{2}, j \leq v\right\},$$

where $\pi\left(b_j = 1 | Z_1, \ldots, Z_n\right)$ is the posterior probability for $b_j = 1$. The minimized Bayes risk is given by

$$\begin{aligned} & \gamma \int_{Z_1,\ldots,Z_n} E_X\left[\min\left\{\pi\left(b(X) = 1 | Z_1, \ldots, Z_n\right), 1 - \pi\left(b(X) = 1 | Z_1, \ldots, Z_n\right)\right\}\right] d\tilde{P}^n \\ = & \ v^{-1}\gamma \int_{Z_1,\ldots,Z_n} \sum_{j=1}^{v} \left[\min\left\{\pi\left(b_j = 1 | Z_1, \ldots, Z_n\right), 1 - \pi(b_j = 1 | Z_1, \ldots, Z_n)\right\}\right] d\tilde{P}^n, \qquad \text{(A.12)} \end{aligned}$$

where $\tilde{P}^n$ is the marginal likelihood of $\{(Y_{1,i}, Y_{0,i}, D_i, X_i) : i = 1, \ldots, n\}$ with prior $\pi\left(\mathbf{b}\right)$. For each $j = 1, \ldots, (v)$, let

$$k_j^+ = \#\left\{i : X_i = x_j,\ Y_i D_i = \frac{1}{2}\right\}, \quad k_j^- = \#\left\{i : X_i = x_j,\ Y_i D_i = -\frac{1}{2}\right\}.$$

The posterior for $b_j = 1$ can be written as

$$\pi\left(b_j = 1 | Z_1, \ldots, Z_n\right) = \begin{cases} \frac{1}{2} \quad \text{if} \ \#\left\{i : X_i = x_j, D_i = 1\right\} = 0, \\ \frac{\left(\frac{1}{2}+\gamma\right)^{k_j^+}\left(\frac{1}{2}-\gamma\right)^{k_j^-}}{\left(\frac{1}{2}+\gamma\right)^{k_j^+}\left(\frac{1}{2}-\gamma\right)^{k_j^-} + \left(\frac{1}{2}+\gamma\right)^{k_j^-}\left(\frac{1}{2}-\gamma\right)^{k_j^+}} \quad \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} & \min\left\{\pi\left(b_j = 1 | Z_1, \ldots, Z_n\right), 1 - \pi(b_j = 1 | Z_1, \ldots, Z_n)\right\} \\ = & \ \frac{\min\left\{\left(\frac{1}{2}+\gamma\right)^{k_j^+}\left(\frac{1}{2}-\gamma\right)^{k_j^-}, \left(\frac{1}{2}+\gamma\right)^{k_j^-}\left(\frac{1}{2}-\gamma\right)^{k_j^+}\right\}}{\left(\frac{1}{2}+\gamma\right)^{k_j^+}\left(\frac{1}{2}-\gamma\right)^{k_j^-} + \left(\frac{1}{2}+\gamma\right)^{k_j^-}\left(\frac{1}{2}-\gamma\right)^{k_j^+}} = \frac{\min\left\{1, \left(\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}\right)^{k_j^+ - k_j^-}\right\}}{1 + \left(\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}\right)^{k_j^+ - k_j^-}} \\ = & \ \frac{1}{1 + a^{\left|k_j^+ - k_j^-\right|}}, \quad \text{where } a = \frac{1+2\gamma}{1-2\gamma} > 1. \qquad \text{(A.13)} \end{aligned}$$

Since $k_j^+ - k_j^- = \sum_{i: X_i = x_j} 2Y_i D_i$, plugging (A.13) into (A.12) yields

$$\begin{aligned} v^{-1}\gamma \sum_{j=1}^{v} E_{\tilde{P}^n}\left[\frac{1}{1 + a^{\left|\sum_{i:X_i = x_j} 2Y_i D_i\right|}}\right] & \geq \frac{\gamma}{2v} \sum_{j=1}^{v} E_{\tilde{P}^n}\left[\frac{1}{a^{\left|\sum_{i:X_i = x_j} 2Y_i D_i\right|}}\right] \\ & \geq \frac{\gamma}{2v} \sum_{j=1}^{v} a^{-E_{\tilde{P}^n}\left|\sum_{i:X_i = x_j} 2Y_i D_i\right|}, \end{aligned}$$

where $E_{\tilde{P}^n}\left(\cdot\right)$ is the expectation with respect to the marginal likelihood of $\{(Y_{1,i}, Y_{0,i}, D_i, X_i),$ $i = 1, \ldots, n\}$. The second line follows by $a > 1$, and the third line follows by Jensen's inequality. Given our prior specification for $\mathbf{b}$, the marginal distribution of $Y_{1,i}$ is $\Pr(Y_{1,i} = 1/2) = \Pr(Y_{1,i} =$

$-1/2) = 1/2$, so

$$\begin{aligned} E_{\tilde{P}^n}\left|\sum_{i:X_i=x_j} 2Y_iD_i\right| &= E_{\tilde{P}_n}\left|\sum_{i=1:X_i=x_j,D_i=1} 2Y_{1,i}\right| \\ &= \sum_{k=0}^{n}\binom{n}{k}\left(\frac{1}{2v}\right)^k\left(1-\frac{1}{2v}\right)^{n-k} E\left|B(k,\frac{1}{2})-\frac{k}{2}\right| \end{aligned}$$

holds, where $B(k, \frac{1}{2})$ is the binomial random variable with parameters k and $\frac{1}{2}$. By noting

$$\begin{aligned} E\left|B(k,\frac{1}{2})-\frac{k}{2}\right| &\leq \sqrt{E\left(B(k,\frac{1}{2})-\frac{k}{2}\right)^2} \quad (\because \text{Cauchy-Schwartz inequality}) \\ &= \sqrt{\frac{k}{4}}, \end{aligned}$$

we obtain

$$\begin{aligned} E_{\tilde{P}^n}\left|\sum_{i:X_i=x_j} 2Y_iD_i\right| &\leq \sum_{k=0}^{n}\binom{n}{k}\left(\frac{1}{2v}\right)^k\left(1-\frac{1}{2v}\right)^{n-k}\sqrt{\frac{k}{4}} = E\sqrt{\frac{B\left(n,\frac{1}{2v}\right)}{4}} \\ &\leq \sqrt{\frac{n}{8v}} \quad (\because \text{Jensen's inequality}). \end{aligned}$$

Hence, the Bayes risk is bounded from below by

$$\begin{aligned} \frac{\gamma}{2}a^{-\sqrt{\frac{n}{8v}}} &\geq \frac{\gamma}{2}\exp\left\{-(a-1)\sqrt{\frac{n}{8v}}\right\} \quad (\because 1+x \leq e^x \ \forall x) \\ &= \frac{\gamma}{2}\exp\left\{-\frac{4\gamma}{1-2\gamma}\sqrt{\frac{n}{8v}}\right\}. \end{aligned} \tag{A.14}$$

This lower bound of the Bayes risk has the slowest convergence rate when γ is set to be proportional to $n^{-1/2}$. Specifically, let $\gamma = \sqrt{\frac{v}{n}}$. Then, we have

$$\frac{\gamma}{2}\exp\left\{-\frac{4\gamma}{1-2\gamma}\sqrt{\frac{n}{8v}}\right\} = \frac{1}{2}\sqrt{\frac{v}{n}}\exp\left\{-\frac{\sqrt{2}}{1-2\gamma}\right\} \geq \frac{1}{2}\sqrt{\frac{v}{n}}\exp\left\{-2\sqrt{2}\right\} \quad \text{if } 1-2\gamma \geq \frac{1}{2}.$$

The condition $1 - 2\gamma \geq \frac{1}{2}$ is equivalent to $n \geq 16v$. Multiplying M to this lower bound completes the proof. □

A.3 Proofs of Theorems 2.3 and 2.4

The next lemma is the concentration inequality of Bousquet (2002).

Lemma A.6. *Let $\mathcal{F}$ be a countable family of measurable functions, such that* $\sup_{f \in \mathcal{F}} E_P(f^2) \leq \delta^2$ *and* $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq \bar{F}$ *for some constants δ and $\bar{F}$. Let* $S = \sup_{f \in \mathcal{F}} (E_n(f) - E_P(f))$. *Then, for every positive t,*

$$P^n\left(S - E_{P^n}(S) \geq \sqrt{\frac{2\left[\delta^2 + 4\bar{F}E_{P^n}(S)\right]t}{n}} + \frac{2\bar{F}t}{3n}\right) \leq \exp(-t).$$

In proving Theorem 2.3, it is convenient to work with the *normalized welfare difference*,

$$d(G, G') \equiv \frac{\kappa}{M}\left[W(G) - W(G')\right],$$

and its sample analogue

$$d_n(G, G') \equiv \frac{\kappa}{M}\left[W_n(G) - W_n(G')\right]. \tag{A.15}$$

By Assumption 2.1 (BO) and (SO), both $d(G, G')$ and $d_n(G, G')$ are bounded in $[-1, 1]$, and the normalized welfare difference relates to the original welfare loss of decision set G as

$$d(G^*_{FB}, G) = \frac{\kappa}{M}\left[W(G^*_{FB}) - W(G)\right] \in [0, 1]. \tag{A.16}$$

Hence, the welfare loss upper bound of $\hat{G}_{EWM}$ can be obtained by multiplying M/κ by the upper bound of $d(G^*_{FB}, \hat{G}_{EWM})$.

Note that $d(G^*_{FB}, G)$ can be bounded from above by $P_X(G^*_{FB} \triangle G)$, since

$$d(G^*_{FB}, G) \;=\; \frac{\kappa}{M}\int_{G^*_{FB}\triangle G} |\tau(X)|\, dP_X \;\leq\; \kappa P_X(G^*_{FB}\triangle G) \;\leq\; P_X(G^*_{FB}\triangle G). \tag{A.17}$$

On the other hand, with Assumption 2.2 (MA) imposed, $P_X(G^*_{FB}\triangle G)$ can be bounded from above by a function of $d(G^*_{FB}, G)$, as the next lemma shows. We borrow this lemma from Tsybakov (2004).

Lemma A.7. *Suppose Assumption 2.2 (MA) holds with margin coefficient $\alpha \in (0, \infty)$. Then*

$$P_X(G^*_{FB}\triangle G) \leq c_1(M, \kappa, \eta, \alpha) d(G^*_{FB}, G)^{\frac{\alpha}{1+\alpha}}$$

holds for all $G \in \mathcal{G}$, where $c_1(M, \kappa, \eta, \alpha) = \left(\frac{M}{\kappa\eta\alpha}\right)^{\frac{\alpha}{1+\alpha}}(1 + \alpha)$.

Proof. Let $A = \{x : |\tau(x)| > t\ \}$ and consider the following inequalities:

$$\begin{aligned} W(G^*_{FB}) - W(G) &= \int_{G^*_{FB}\triangle G} |\tau(X)|\, dP_X \;\geq\; \int_{G^*_{FB}\triangle G} |\tau(X)|\, 1\{X \in A\}\, dP_X \\ &\geq\; tP_X\left((G^*_{FB}\triangle G) \cap A\right) \;\geq\; t\left[P_X\left(G^*_{FB}\triangle G\right) - P_X(A^c)\right] \\ &\geq\; t\left[P_X\left(G^*_{FB}\triangle G\right) - \left(\frac{t}{\eta}\right)^\alpha\right], \end{aligned}$$

where the final line uses the margin condition. The right-hand side is maximized at $t = \eta(1+\alpha)^{-\frac{1}{\alpha}} \left[P_X\left(G^*_{FB} \triangle G\right)\right]^{\frac{1}{\alpha}} \leq \eta$, so it holds

$$W(G^*_{FB}) - W(G) \geq \eta\alpha \left(\frac{1}{1+\alpha}\right)^{\frac{1+\alpha}{\alpha}} \left[P_X\left(G^*_{FB} \triangle G\right)\right]^{\frac{1+\alpha}{\alpha}}.$$

This, in turn, implies

$$P_X\left(G^*_{FB} \triangle G\right) \leq \left(\frac{M}{\kappa\eta\alpha}\right)^{\frac{\alpha}{1+\alpha}} (1+\alpha) d(G^*_{FB}, G)^{\frac{\alpha}{1+\alpha}}.$$

□

Proof of Theorem 2.3. Let $a = \sqrt{kt}\epsilon_n$ with $k \geq 1$, $t \geq 1$, and $\epsilon_n > 0$, where $t \geq 1$ is arbitrary, k is a constant that we choose later, and ϵ_n is a sequence indexed by sample size n whose proper choice will be discussed in a later step. The normalized welfare loss can be bounded by

$$d(G^*_{FB}, \hat{G}_{EWM}) \leq d(G^*_{FB}, \hat{G}_{EWM}) - d_n\left(G^*_{FB}, \hat{G}_{EWM}\right),$$

as $d_n\left(G^*_{FB}, \hat{G}_{EWM}\right) \leq 0$ by Assumption 2.2 (FB). Define a class of functions induced by $G \in \mathcal{G}$.

$$\begin{aligned} \mathcal{H} &\equiv \left\{h(Z_i; G) : G \in \mathcal{G}\right\}, \\ h(Z_i; G) &\equiv \frac{\kappa}{M}\left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right)\left[1\left\{X_i \in G\right\} - 1\left\{X_i \in G^*_{FB}\right\}\right]. \end{aligned}$$

By Assumption 2.1 (VC) and Lemma A.1, $\mathcal{H}$ is a VC-subgraph-class with VC-dimension at most $v < \infty$ with envelope $\bar{H} = 1$. Using $h(Z_i; G)$, we can write $d(G^*_{FB}, G) = -E_P\left(h(Z_i; G)\right)$. Since $d(G^*_{FB}, G) \geq 0$ for all $G \in \mathcal{G}$, it holds $-E_P(h) \geq 0$ for all $h \in \mathcal{H}$.

Since we have

$$d(G^*_{FB}, \hat{G}_{EWM}) - d_n\left(G^*_{FB}, \hat{G}_{EWM}\right) = E_n\left(h(Z_i; \hat{G}_{EWM})\right) - E_P\left(h(Z_i; \hat{G}_{EWM})\right)$$

and $d_n\left(G^*_{FB}, \hat{G}_{EWM}\right) \leq 0$, the normalized welfare loss can be bounded by

$$\begin{aligned} d(G^*_{FB}, \hat{G}_{EWM}) &\leq E_n\left(h(Z_i; \hat{G}_{EWM})\right) - E_P\left(h(Z_i; \hat{G}_{EWM})\right) \\ &\leq V_a\left[d(G^*_{FB}, \hat{G}_{EWM}) + a^2\right], \end{aligned}$$

where

$$V_a = \sup_{h \in \mathcal{H}}\left\{\frac{E_n(h) - E_P(h)}{-E_P(h) + a^2}\right\} = \sup_{h \in \mathcal{H}}\left\{E_n\left(\frac{h}{-E_P(h) + a^2}\right) - E_P\left(\frac{h}{-E_P(h) + a^2}\right)\right\}.$$

On event $V_a < \frac{1}{2}$, $d(G^*_{FB}, \hat{G}_{EWM}) \leq a^2$ holds, so this implies

$$P^n \left(d(G^*_{FB}, \hat{G}_{EWM}) \geq a^2 \right) \leq P^n \left(V_a \geq \frac{1}{2} \right). \tag{A.18}$$

In what follows, our aim is to construct an exponential inequality for $P^n \left(V_a \geq \frac{1}{2} \right)$ involving only t, and we make use of such exponential tail bound to bound $E_{P^n} \left(d(G^*_{FB}, \hat{G}_{EWM}) \right)$.

To apply Bousquet's inequality (Lemma A.6) to V_a, note first that

$$\begin{aligned}
E_P \left(\left(\frac{h}{-E_P(h) + a^2} \right)^2 \right) &\leq \frac{P_X(G^*_{FB} \triangle G)}{(-E_P(h) + a^2)^2} \leq c_1 \frac{[-E_P(h)]^{\frac{\alpha}{1+\alpha}}}{(-E_P(h) + a^2)^2} \\
& \qquad (\because \text{ by Lemma A.7 and } d(G^*_{FB}, G) = -E_P(h(Z_i; G))) \\
&\leq c_1 \sup_{\epsilon \geq 0} \frac{\epsilon^{\frac{2\alpha}{1+\alpha}}}{(\epsilon^2 + a^2)^2} \leq c_1 \frac{1}{a^2} \sup_{\epsilon \geq 0} \frac{\epsilon^{\frac{2\alpha}{1+\alpha}}}{\epsilon^2 + a^2} \leq c_1 \frac{1}{a^2} \sup_{\epsilon \geq 0} \left(\frac{\epsilon^{\frac{\alpha}{1+\alpha}}}{\epsilon \vee a} \right)^2 \\
&\leq c_1 \frac{1}{a^4} a^{\frac{2\alpha}{1+\alpha}},
\end{aligned}$$

where c_1 is a constant that depends only on $(M, \kappa, \eta, \alpha)$ as defined in Lemma A.7. We, on the other hand, have

$$\sup_{h \in \mathcal{H}} \left| \sup_Z \frac{h}{-E_P(h) + a^2} \right| \leq \frac{1}{a^2}.$$

Hence, Lemma A.6 gives, with probability larger than $1 - \exp(-t)$,

$$V_a \leq E_{P^n}(V_a) + \sqrt{\frac{2 \left[c_1 a^{\frac{2\alpha}{1+\alpha} - 2} + 4 E_{P^n}(V_a) \right] t}{n a^2}} + \frac{2t}{3na^2}. \tag{A.19}$$

Next, we derive an upper bound of $E_{P^n}(V_a)$ by applying the maximal inequality of Lemma A.5. Let $r > 1$ be arbitrary and consider partitioning $\mathcal{H}$ by $\mathcal{H}_0, \mathcal{H}_1, \ldots$, where $\mathcal{H}_0 = \left\{ h \in \mathcal{H} : -E_P(h) \leq a^2 \right\}$ and $\mathcal{H}_j = \left\{ h \in \mathcal{H} : r^{2(j-1)} a^2 < -E_P(h) \leq r^{2j} a^2 \right\}$, $j = 1, 2, \ldots$. Then,

$$\begin{aligned}
V_a &\leq \sup_{h \in \mathcal{H}_0} \left\{ \frac{E_n(h) - E_P(h)}{-E_P(h) + a^2} \right\} + \sum_{j \geq 1} \sup_{h \in \mathcal{H}_j} \left\{ \frac{E_n(h) - E_P(h)}{-E_P(h) + a^2} \right\} \\
&\leq \frac{1}{a^2} \left[\sup_{h \in \mathcal{H}_0} (E_n(h) - E_P(h)) + \sum_{j \geq 1} (1 + r^{2(j-1)})^{-1} \sup_{h \in \mathcal{H}_j} (E_n(h) - E_P(h)) \right] \\
&\leq \frac{1}{a^2} \left[\begin{array}{c} \sup_{-E_P(h) \leq a^2} (E_n(h) - E_P(h)) \\ + \sum_{j \geq 1} (1 + r^{2(j-1)})^{-1} \sup_{-E_P(h) \leq r^{2j} a^2} (E_n(h) - E_P(h)) \end{array} \right].
\end{aligned} \tag{A.20}$$

Since it holds $\|h\|^2_{L_2(P)} \leq P_X(G^*_{FB} \triangle G) \leq c_1(M, \kappa, \eta, \alpha) [-E_P(h)]^{\frac{\alpha}{1+\alpha}}$, where the latter inequality follows from Lemma A.7, $-E_P(h) \leq r^{2j} a^2$ implies $\|h\|_{L_2(P)} \leq c_1^{1/2} r^{\frac{\alpha}{1+\alpha} j} a^{\frac{\alpha}{1+\alpha}}$. Hence, (A.20) can be further bounded by

$$V_a \le \frac{1}{a^2}\left[\begin{array}{c} \sup_{\|h\|_{L_2(P)}\le c_1^{1/2}a^{\frac{\alpha}{1+\alpha}}}\left(E_n\left(h\right)-E_P(h)\right) \\ +\sum_{j\ge 1}(1+r^{2(j-1)})^{-1}\sup_{\|h\|_{L_2(P)}\le c_1^{1/2}r^{\frac{\alpha}{1+\alpha}j}a^{\frac{\alpha}{1+\alpha}}}\left(E_n\left(h\right)-E_P(h)\right)\end{array}\right].$$

We apply Lemma A.5 to each supremum term, and obtain

$$E_{P^n}(V_a) \;\le\; C_2\frac{c_1^{\frac{1}{2}}}{a^2}\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}}\sum_{j\ge 0}\frac{r^{\frac{\alpha}{1+\alpha}j}}{1+r^{2(j-1)}} \;\le\; C_2 c_1^{\frac{1}{2}}\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}\left(\frac{r^2}{1-r^{-\frac{2+\alpha}{1+\alpha}}}\right) \;\le\; c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}$$

for

$$n\ge\frac{C_1 v}{c_1 a^{\frac{2\alpha}{1+\alpha}}} \quad\Longleftrightarrow\quad a\ge\left(\frac{C_1}{c_1}\right)^{\frac{1+\alpha}{2\alpha}}\left(\frac{v}{n}\right)^{\frac{1+\alpha}{2\alpha}} \tag{A.21}$$

where C_1 and C_2 are universal constants defined in Lemmas A.4 and A.5, and $c_2 = C_2 c_1^{\frac{1}{2}}\left(\frac{r^2}{1-r^{-\frac{2+\alpha}{1+\alpha}}}\right)\vee 1$ is a constant greater than or equal to one and depends only on (M,κ,η,α), as $r>1$ is fixed. We plug in this upper bound into (A.19) to obtain

$$V_a \le c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}+\sqrt{\frac{2\left[c_1 a^{\frac{2\alpha}{1+\alpha}-2}+4c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}\right]t}{na^2}}+\frac{2t}{3na^2}. \tag{A.22}$$

Choose ϵ_n as the root of $c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}=1$, i.e.,

$$\epsilon_n=\left(c_2\sqrt{\frac{v}{n}}\right)^{\frac{1+\alpha}{2+\alpha}}. \tag{A.23}$$

Note that the right hand side of (A.22) is decreasing in a, and $a\ge\epsilon_n$ by the construction. Hence, if ϵ_n satisfies inequality (A.21), i.e.,

$$n\ge c_2^{-\alpha}\left(\frac{C_1}{c_1}\right)^{1+\frac{\alpha}{2}}v,$$

which can be reduced to an innocuous restriction $n\ge 1$ by inflating, if necessary, c_1 large enough, we can substitute ϵ_n for a to bound the right hand side of (A.22). In particular, by noting

$$\begin{aligned} c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2} &\le \frac{\epsilon_n}{a}=\frac{1}{\sqrt{kt}}\le\frac{1}{\sqrt{k}} \quad\text{and} \\ a^{\frac{2\alpha}{1+\alpha}-2} &= a^{2\left(\frac{\alpha}{1+\alpha}-2\right)}a^2\le\left[\epsilon_n^{\frac{\alpha}{1+\alpha}-2}\right]^2\epsilon_n^2=c_2^{-2}v^{-1}n\epsilon_n^2, \end{aligned}$$

the right-hand side of (A.22) can be bounded by

$$\begin{aligned} V_a &\leq \frac{1}{\sqrt{k}} + \sqrt{2\frac{c_1 c_2^{-2} v^{-1} n\epsilon_n^2 + 8}{nk\epsilon_n^2}} + \frac{2}{3nk\epsilon_n^2} = \frac{1}{\sqrt{k}} + \sqrt{\frac{2c_1 c_2^{-2} v^{-1}}{k} + \frac{8}{nk\epsilon_n^2}} + \frac{2}{3nk\epsilon_n^2} \\ &\leq \frac{1}{\sqrt{k}} + \sqrt{\frac{2c_1 c_2^{-2} v^{-1}}{k} + \frac{8}{k}} + \frac{2}{3k} \quad \text{for } n\epsilon_n^2 \geq 1. \end{aligned} \tag{A.24}$$

Note that condition $n\epsilon_n^2 \geq 1$ used to derive the last line is valid for all n, since it is equivalent to $n \geq c_2^{-2(1+\alpha)} v^{-(1+\alpha)}$, which holds for all $n \geq 1$ since $c_2 \geq 1$ and $v \geq 1$. By choosing k large enough so that the right-hand side of (A.24) is less than $\frac{1}{2}$, we can conclude

$$\Pr(V_a < \frac{1}{2}) \geq 1 - \exp(-t). \tag{A.25}$$

Hence, (A.18) yields

$$P^n\left(d(G_{FB}^*, \hat{G}_{EWM}) \geq kt\epsilon_n^2\right) \leq \exp(-t)$$

for all $t \geq 1$. From this exponential bound, we obtain

$$\begin{aligned} E_{P^n}\left(d(G_{FB}^*, \hat{G}_{EWM})\right) &= \int_0^\infty P^n\left(d(G_{FB}^*, \hat{G}_{EWM}) > t'\right) dt' \\ \leq \int_0^{k\epsilon_n^2} P^n\left(d(G_{FB}^*, \hat{G}_{EWM}) \geq t'\right) dt' &+ \int_{k\epsilon_n^2}^\infty P^n\left(d(G_{FB}^*, \hat{G}_{EWM}) \geq t'\right) dt' \leq k\epsilon_n^2 + k\epsilon_n^2 e^{-1} \\ = (1+e^{-1}) k c_2^{\frac{2(1+\alpha)}{2+\alpha}} \left(\frac{v}{n}\right)^{\frac{1+\alpha}{2+\alpha}}. \end{aligned}$$

So, setting $c = \frac{M}{\kappa}(1+e^{-1}) k c_2^{\frac{2(1+\alpha)}{2+\alpha}}$ leads to the conclusion. □

Proof of Theorem 2.4. As in the proof of Theorem 2.2, we work with the normalized outcome support, $Y_{1,i}, Y_{0,i} \in \left[-\frac{1}{2}, \frac{1}{2}\right]$. With the normalized outcome, we can assume without loss of generality that constant η of the margin assumption satisfies $\eta \leq 1$.

Let $\alpha \in (0, \infty)$ and $\eta \in (0, 1]$ be given. Similarly to the proof of Theorem 2.2, we consider constructing a suitable subclass $\mathcal{P}^* \subset \mathcal{P}(1, \kappa, \eta, \alpha)$. Let $x_1, \dots, x_v \in \mathcal{X}$ be v points that are shattered by $\mathcal{G}$, and let γ be a positive number satisfying $\gamma \leq \min\left\{\eta, \frac{1}{2}\right\}$, whose proper choice will be given later. We fix the marginal distribution of X at the one supported only on $(x_1, \dots, x_v)$ and having the probability mass function,

$$\begin{aligned} P_X(X_i &= x_j) = \frac{1}{v-1}\left(\frac{\gamma}{\eta}\right)^\alpha, \quad \text{for } j = 1, \dots, (v-1), \text{ and} \\ P_X(X_i &= x_v) = 1 - \left(\frac{\gamma}{\eta}\right)^\alpha. \end{aligned}$$

Thus-constructed marginal distribution of X is common in $\mathcal{P}^*$. As in the proof of Theorem 2.2, we specify D to be independent of (Y_1, Y_0, X) and follow the Bernoulli distribution with $\Pr(D = 1) = 1/2$. Let $\mathbf{b} = (b_1, \ldots, b_{v-1}) \in \{0,1\}^{v-1}$ be a binary vector that uniquely indexes a member of $\mathcal{P}^*$, and, accordingly, write $\mathcal{P}^* = \left\{P_{\mathbf{b}} : \mathbf{b} \in \{0,1\}^{v-1}\right\}$. For each $j = 1, \ldots, (v-1)$, we specify the conditional distribution of Y_1 given $X = x_j$ to be (A.10) if $b_j = 1$ and (A.11) if $b_j = 0$. For $j = v$, the conditional distribution of Y_1 given $X = x_v$ is degenerate at $Y_1 = \frac{1}{2}$. As for the conditional distribution of Y_0 given $X = x_j$, we consider the degenerate distribution at $Y_0 = 0$ for $j = 1, \ldots, (v-1)$, and the degenerate distribution at $Y_0 = -\frac{1}{2}$ for $X = x_v$. In this specification of $\mathcal{P}^*$, it holds

$$P_X(|\tau(X)| \leq t) = \begin{cases} 0 & \text{for } t \in [0, \gamma), \\ \left(\frac{\gamma}{\eta}\right)^{\alpha} & \text{for } t \in [\gamma, 1), \\ 1 & \text{for } t \geq 1. \end{cases},$$

for every $P_{\mathbf{b}} \in \mathcal{P}^*$. Since $\gamma \leq \eta$, $P_X(|\tau(X)| \leq t) \leq (t/\eta)^{\alpha}$ holds for all $t \in [0, \eta]$. Furthermore, by the construction of the support points, for every $P_{\mathbf{b}} \in \mathcal{P}^*$, the first-best decision rule $G^*_{\mathbf{b}} = \{x_j : j < v, b_j = 1\} \cup \{x_v\}$ is contained in $\mathcal{G}$. Hence, $\mathcal{P}^* \subset \mathcal{P}_{FB}(1, \kappa, \eta, \alpha)$ holds.

Let $\pi(\mathbf{b})$ be a prior distribution for $\mathbf{b}$ such that $b_1, \ldots, b_{v-1}$ are iid and $b_1 \sim Ber(1/2)$. The maximized social welfare is

$$W(G^*_{\mathbf{b}}) = \frac{\gamma}{v-1}\left(\frac{\gamma}{\eta}\right)^{\alpha}\left(\sum_{j=1}^{v-1} b_j\right) + \left[1 - \left(\frac{\gamma}{\eta}\right)^{\alpha}\right].$$

Let $\hat{G}$ be an arbitrary treatment choice rule as a function of $(Z_1, \ldots, Z_n)$, and $\hat{\mathbf{b}} \in \{0,1\}^v$ be a binary vector whose j-th element is $\hat{b}_j = 1\left\{x_j \in \hat{G}\right\}$.

The welfare loss can be bounded from below as follows:

$$\begin{aligned} &\sup_{P \in \mathcal{P}(1,\kappa,\eta,\alpha)} E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G})\right] \geq \sup_{P_{\mathbf{b}} \in \mathcal{P}^*} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\hat{G})\right] \\ \geq\ & \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\hat{G})\right] d\pi(\mathbf{b}) \geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W(G^*_{\mathbf{b}}) - W(\hat{G} \cup \{x_v\})\right] d\pi(\mathbf{b}) \\ =\ & \gamma \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[P_X\left((G^*_{\mathbf{b}} \triangle \hat{G}) \cap \{x_1, \ldots, x_{v-1}\}\right)\right] d\pi(\mathbf{b}) \\ =\ & \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\left\{b(X) \neq \hat{b}(X)\right\} \cap \{x_1, \ldots, x_{v-1}\}\right) dP^n_{\mathbf{b}}(Z_1, \ldots, Z_n)\, d\pi(\mathbf{b}) \\ \geq\ & \inf_{G_n} \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\{b(X) \neq b_n(X)\}\right) dP^n_{\mathbf{b}}(Z_1, \ldots, Z_n)\, d\pi(\mathbf{b}), \end{aligned}$$

where the second line follows since $W(G^*_{\mathbf{b}}) - W(\hat{G}) \geq W(G^*_{\mathbf{b}}) - W(\hat{G} \cup \{x_v\})$ holds for every $\mathbf{b}$ and

$\hat{G}$ and $G_n = \{x_j : b_n(x_j) = 1\}$ for which the infimum taken in the last line is an estimator for the decision set that is constrained to deterministically containing $\{x_v\}$, i.e., $b_n(x_v) = 1$.

By the same reasonings as in obtaining (A.12), the lower bound of the welfare loss as viewed as the Bayes risk can be expressed as

$$\begin{aligned} &\sup_{P\in\mathcal{P}(1,\kappa,\eta,\alpha)} E_{P^n}\left[W(G^*) - W(\hat{G})\right] \\ \geq\ & \frac{\gamma}{v-1}\left(\frac{\gamma}{\eta}\right)^{\alpha} \int\limits_{Z_1,\ldots,Z_n} \sum_{j=1}^{v-1} \left[\min\left\{\pi\left(b_j = 1|Z_1,\ldots,Z_n\right), 1-\pi(b_j = 1|Z_1,\ldots,Z_n)\right\}\right] d\tilde{P}^n. \end{aligned}$$

Repeating the same bounding arguments as in the proof of Theorem 2.2, a lower bound of the Bayes risk analogous to (A.14) is obtained by

$$\sup_{P\in\mathcal{P}(1,\kappa,\eta,\alpha)} E_{P^n}\left[W(G^*) - W(\hat{G})\right] \geq \frac{\gamma}{2}\left(\frac{\gamma}{\eta}\right)^{\alpha} \exp\left\{-\frac{4\gamma}{1-2\gamma}\sqrt{\frac{n}{8(v-1)}\left(\frac{\gamma}{\eta}\right)^{\alpha}}\right\}.$$

The slowest convergence rate of this lower bound can be obtained by tuning γ to be converging at the rate of $n^{-\frac{1}{2+\alpha}}$. In particular, by choosing $\gamma = \eta^{\frac{\alpha}{2+\alpha}}\left(\frac{v-1}{n}\right)^{\frac{1}{2+\alpha}}$ assuming $\gamma \leq \frac{1}{4}$, the exponential term can be bounded from below by $\exp\left\{-2\sqrt{2}\right\}$, so we obtain the following lower bound,

$$\frac{1}{2}\eta^{-\frac{\alpha}{2+\alpha}}\left(\frac{v-1}{n}\right)^{\frac{1+\alpha}{2+\alpha}} \exp\left\{-2\sqrt{2}\right\}. \tag{A.26}$$

Recall that γ is constrained to $\gamma \leq \min\left\{\eta, \frac{1}{4}\right\}$. This implies that the obtained bound is valid for

$$n \geq \left(\max\left\{\eta^{-1}, 4\right\}\right)^{2+\alpha}\eta^{\alpha}(v-1),$$

whose stronger but simpler form is given by

$$n \geq \max\left\{\eta^{-2}, 4^{2+\alpha}\right\}(v-1). \tag{A.27}$$

The lower bound presented in this theorem follows by denormalizing the outcomes, i.e., multiply M to (A.26) and substitute η/M for η appearing in (A.26) and (A.27). □

A.4 Proof of Theorems 2.5 and 2.6

Proof of Theorem 2.5. Let $W_n^{\tau}(G)$ be the sample analogue of the welfare criterion (1.2) in the main text that one would construct if the true regression equations were known, $W_n^{\tau}(G) \equiv E_n(m_0(X_i)) + E_n(\tau(X_i) \cdot 1\{X_i \in G\})$, and $\hat{W}_n^{\tau}(G)$ be the empirical welfare with the conditional treatment effect estimators $\hat{\tau}^m(\cdot)$ plugged in,

$$\hat{W}_n^{\tau}(G) \equiv E_n\left[m_0(X_i) + \hat{\tau}^m(X_i)\,1\{X_i \in G\}\right]. \tag{A.28}$$

Since the m-hybrid rule maximizes $\hat{W}_n^{\tau}(\cdot)$, it holds $\hat{W}_n^{\tau}(\hat{G}_{m-hybrid}) - \hat{W}_n^{\tau}(\tilde{G}) \geq 0$ for any $\tilde{G} \in \mathcal{G}$. The following inequalities therefore follow:

$$\begin{aligned} W(\tilde{G}) - W(\hat{G}_{m-hybrid}) &\leq W_n^{\tau}(\tilde{G}) - \hat{W}_n^{\tau}\left(\tilde{G}\right) - W_n^{\tau}(\hat{G}_{m-hybrid}) + \hat{W}_n^{\tau}\left(\hat{G}_{m-hybrid}\right) \\ &\quad + W(\tilde{G}) - W(\hat{G}_{m-hybrid}) - W_n^{\tau}(\tilde{G}) + W_n^{\tau}(\hat{G}_{m-hybrid}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\tau\left(X_i\right) - \hat{\tau}^m(X_i)\right]\left[1\left\{X_i \in \tilde{G}\right\} - 1\left\{X_i \in \hat{G}_{m-hybrid}\right\}\right] \\ &\quad + W(\tilde{G}) - W_n^{\tau}(\tilde{G}) + W_n^{\tau}(\hat{G}_{m-hybrid}) - W(\hat{G}_{m-hybrid}) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tau}^m\left(X_i\right) - \tau(X_i)\right| + 2\sup_{G\in\mathcal{G}}\left|W_n^{\tau}(G) - W(G)\right|. \end{aligned} \tag{A.29}$$

This implies that the average welfare loss of the m-hybrid rule can be bounded by

$$E_{P^n}\left[W_{\mathcal{G}}^{*} - W(\hat{G}_{m-hybrid})\right] \leq E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tau}^m(X_i) - \tau(X_i)\right|\right] + 2E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W_n^{\tau}(G) - W(G)\right|\right]. \tag{A.30}$$

For the e-hybrid rule, replacing $W_n^{\tau}(\cdot)$ and $\hat{W}_n^{\tau}(\cdot)$ in (A.29) with the empirical welfare $W_n(\cdot)$ defined in (1.7) and $\hat{W}_n(G) \equiv E_n\left[\frac{Y_i(1-D_i)}{1-e(X_i)} + \hat{\tau}_i^e \cdot 1\{X_i \in G\}\right]$, respectively, yields a similar upper bound

$$E_{P^n}\left[W_{\mathcal{G}}^{*} - W(\hat{G}_{e-hybrid})\right] \leq E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tau}_i^e - \tau_i\right|\right] + 2E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W_n(G) - W(G)\right|\right], \tag{A.31}$$

where $\tau_i = \frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}$. Note that the uniform convergence rate of $E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W_n^{\tau}(G) - W(G)\right|\right]$ is $n^{-1/2}$, same as that of $E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W_n(G) - W(G)\right|\right]$, since the proof of Theorem 2.1 can be applied to the following class of functions:

$$\mathcal{F}^{\tau} \equiv \left\{f(X_i;G) = m_0(X_i) + \tau(X_i)\cdot 1\{X_i \in G\} : G \in \mathcal{G}\right\},$$

which is the VC-subgraph class with the VC-dimension at most v by Lemma A.1. Combined with Condition 2.1 (m), (A.30) implies the uniform convergence rate of the m-hybrid rule given in the current theorem. Similarly, combined with Condition 2.1 (e) and $n^{-1/2}$-convergence rate of $E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W_n(G) - W(G)\right|\right]$, (A.31) leads to the uniform convergence rate of $\phi_n^{-1} \vee n^{-1/2}$ for the e-hybrid rule. □

The next lemma gives a linearized solution of a certain polynomial inequality. We owe this lemma to Shin Kanaya (2014, personal communication). The technique of applying the mean value expansion to an implicit function defined as the root of a polynomial equation has been used in the context of bandwidth choice in Kanaya and Kristensen (2014).

Lemma A.8. *Let $A \geq 0$, $B \geq 0$, and $X \geq 0$. For any $\alpha \geq 0$, $X \leq AX^{\frac{\alpha}{1+\alpha}} + B$ implies*

$$X \leq A^{1+\alpha} + (1+\alpha)B.$$

Proof. When $A = B = 0$, the conclusion trivially holds. When $B > 0$, $X = AX^{\frac{\alpha}{1+\alpha}} + B$ has a unique root, and we denote it by $X^* = g(A, B)$. When $A > 0$ and $B = 0$, we mean by $g(A, 0)$ the nonzero root of $X = AX^{\frac{\alpha}{1+\alpha}}$. Let $f(X, A, B) = X - AX^{\frac{\alpha}{1+\alpha}} - B$. By the form of the inequality, the original inequality can be equivalently written as $X \leq X^* = g(A, B)$, so we aim to verify that X^* is bounded from above by $A^{1+\alpha} + (1+\alpha)B$. Consider the mean value expansion of $g(A, B)$ in B at $B = 0$,

$$X^* = g(A, 0) + \frac{\partial g}{\partial B}\left(A, \tilde{B}\right) \times B \quad \text{for some } 0 \leq \tilde{B} \leq B.$$

Note $g(A, 0) = A^{1+\alpha}$. In addition, by the implicit function theorem, we have, with $\tilde{X} = g(A, \tilde{B})$,

$$\begin{aligned}\frac{\partial g}{\partial B}\left(A, \tilde{B}\right) &= -\frac{\frac{\partial f}{\partial B}(\tilde{X}, A, \tilde{B})}{\frac{\partial f}{\partial X}(\tilde{X}, A, \tilde{B})} = \frac{1}{1 - \frac{\alpha}{1+\alpha} A \tilde{X}^{-\frac{1}{1+\alpha}}} = \frac{\tilde{X}}{\frac{\tilde{X}}{1+\alpha} + \frac{\alpha}{1+\alpha}\left(\tilde{X} - A\tilde{X}^{\frac{\alpha}{1+\alpha}}\right)} \\ &= \frac{\tilde{X}}{\frac{\tilde{X}}{1+\alpha} + \frac{\alpha}{1+\alpha}\tilde{B}} \leq 1 + \alpha.\end{aligned}$$

Hence, $X^* \leq A^{1+\alpha} + (1+\alpha)B$ holds. □

The next lemma provides an exponential tail probability bound of the supremum of the centered empirical processes. This lemma follows from Theorem 2.14.9 in van der Vaart and Wellner (1996) combined with their Theorem 2.6.4.

Lemma A.9. *Assume $\mathcal{G}$ is a VC-class of subsets in $\mathcal{X}$ with VC-dimension $v < \infty$. Let $P_{X,n}(\cdot)$ be the empirical probability distribution on $\mathcal{X}$ constructed upon $(X_1, \ldots, X_n)$ generated iid from $P_X(\cdot)$. Then,*

$$P^n\left(\sup_{G \in \mathcal{G}} |P_{X,n}(G) - P_X(G)| > t\right) \leq \left(\frac{C_4 t}{\sqrt{2v}}\right)^{2v} n^v \exp\left(-nt^2\right)$$

holds for every $t > 0$, where C_4 is a universal constant.

Proof of Theorem 2.6. We first consider the m-hybrid case. Set $\tilde{G} = G^*_{FB}$ in (A.29) and rewrite

(A.29) in terms of the normalized welfare loss for $\hat{G}_{m-hybrid}$,

$$\begin{aligned} d(G^*_{FB}, \hat{G}_{m-hybrid}) &\leq \frac{\kappa}{M}\left[W^\tau_n(G^*_{FB}) - \hat{W}^\tau_n(G^*_{FB}) - W^\tau_n(\hat{G}_{m-hybrid}) + \hat{W}^\tau_n\left(\hat{G}_{m-hybrid}\right)\right] \\ &\quad + d(G^*_{FB}, \hat{G}_{m-hybrid}) - d^\tau_n\left(G^*_{FB}, \hat{G}_{m-hybrid}\right) \\ &\leq \frac{1}{n}\sum_{i=1}^n \frac{\kappa}{M}\left[\tau(X_i) - \hat{\tau}^m(X_i)\right]\left[1\{X_i \in G^*_{FB}\} - 1\left\{X_i \in \hat{G}_{m-hybrid}\right\}\right] \\ &\quad + d(G^*_{FB}, \hat{G}_{m-hybrid}) - d^\tau_n\left(G^*_{FB}, \hat{G}_{m-hybrid}\right) \\ &\leq \rho_n + d(G^*_{FB}, \hat{G}_{m-hybrid}) - d^\tau_n\left(G^*_{FB}, \hat{G}_{m-hybrid}\right) \end{aligned} \tag{A.32}$$

where $d(G^*_{FB}, \hat{G}_{m-hybrid})$ is as defined in equation (A.16), $d^\tau_n\left(G^*_{FB}, \hat{G}_{m-hybrid}\right) = W^\tau_n(G^*_{FB}) - W^\tau_n(\hat{G}_{m-hybrid})$,

$$\rho_n \equiv \frac{\kappa}{M}\max_{1\leq i\leq n}\left|\hat{\tau}^m(X_i) - \tau(X_i)\right| P_{X,n}\left(G^*_{FB}\triangle\hat{G}_{m-hybrid}\right),$$

and $P_{X,n}$ is the empirical distribution on $\mathcal{X}$ constructed upon $(X_1, \ldots, X_n)$. Define a class of functions generated by $G \in \mathcal{G}$,

$$\begin{aligned} \mathcal{H}^\tau &\equiv \{h(Z_i; G) : G \in \mathcal{G}\}, \\ h(Z_i; G) &\equiv \frac{\kappa}{M}\tau(X_i)\cdot\left[1\{X_i \in G\} - 1\{X_i \in G^*_{FB}\}\right], \end{aligned}$$

which is a VC-subgraph class with the VC-dimension at most v with envelope $\bar{H} = 1$ by Lemma A.1. Let $a = \sqrt{kt}\epsilon_n$ be as defined in the proof of Theorem 2.3 and $V^\tau_a \equiv \sup_{h\in\mathcal{H}^\tau}\left\{\frac{E_n(h)-E_P(h)}{-E_P(h)+a^2}\right\}$. By noting

$$d(G^*_{FB}, \hat{G}_{m-hybrid}) - d^\tau_n\left(G^*_{FB}, \hat{G}_{m-hybrid}\right) \leq V^\tau_a(d(G^*_{FB}, \hat{G}_{m-hybrid}) + a^2),$$

inequality (A.32) implies

$$d(G^*_{FB}, \hat{G}_{m-hybrid}) \leq \rho_n + V^\tau_a(d(G^*_{FB}, \hat{G}_{m-hybrid}) + a^2). \tag{A.33}$$

Denote event $\left\{V^\tau_a < \frac{1}{2}\right\}$ by Ω_t, which is equivalent to event $\left\{d(G^*_{FB}, \hat{G}_{m-hybrid}) \leq 2\rho_n + k\epsilon^2_n t\right\}$. The same line of argument that leads to (A.25) in the proof of Theorem 2.3 leads to, for $t \geq 1$,

$$P^n(\Omega_t) = P^n\left(d(G^*_{FB}, \hat{G}_{m-hybrid}) \leq 2\rho_n + k\epsilon^2_n t\right) \geq 1 - \exp(-t), \tag{A.34}$$

where ϵ_n is given in (A.23). We bound ρ_n from above by

$$\rho_n \leq \frac{\kappa}{M}\left[\max_{1\leq i\leq n}\left|\hat{\tau}^m(X_i) - \tau(X_i)\right| P_X\left(G^*_{FB}\triangle\hat{G}_{m-hybrid}\right) + \mathcal{V}_{0,n}\max_{1\leq i\leq n}\left|\hat{\tau}^m(X_i) - \tau(X_i)\right|\right],$$

where

$$\mathcal{V}_{0,n} = \sup_{G\in\mathcal{G}:} \left| P_{X,n}(G^*_{FB} \triangle G) - P_X \left(G^*_{FB} \triangle G\right) \right| .$$

Let $\lambda > 0$, that will be chosen properly later. Define events

$$\Lambda_1 = \left\{ \mathcal{V}_{0,n} \leq n^{-\lambda} \right\}, \quad \Lambda_2 = \left\{ P_X \left(G^*_{FB} \triangle \hat{G}_{m-hybrid} \right) \geq n^{-\lambda} \right\}.$$

Then, on $\Lambda_1 \cap \Lambda_2$, it holds $\mathcal{V}_{0,n} \leq P_X \left(G^*_{FB} \triangle \hat{G}_{m-hybrid} \right)$. Therefore, on $\Lambda_1 \cap \Lambda_2 \cap \Omega_t$, $d(G^*_{FB}, \hat{G}_{m-hybrid})$ can be bounded by

$$\begin{aligned} d(G^*_{FB}, \hat{G}_{m-hybrid}) &\leq 4 \frac{\kappa}{M} \max_{1\leq i\leq n} \left| \hat{\tau}^m\left(X_i\right) - \tau(X_i) \right| P_X \left(G^*_{FB} \triangle \hat{G}_{m-hybrid} \right) + k\epsilon_n^2 t \\ &\leq 4c_1 \frac{\kappa}{M} \max_{1\leq i\leq n} \left| \hat{\tau}^m\left(X_i\right) - \tau(X_i) \right| d(G^*_{FB}, \hat{G}_{m-hybrid})^{\frac{\alpha}{1+\alpha}} + k\epsilon_n^2 t, \end{aligned}$$

where the second line follows from Lemma A.7 with the same definition of c_1 given there. By Lemma A.8 and substituting (A.23) to ϵ_n, we obtain, on event $\Lambda_1 \cap \Lambda_2 \cap \Omega_t$,

$$d(G^*_{FB}, \hat{G}_{m-hybrid}) \leq c_6 \left[\max_{1\leq i\leq n} \left| \hat{\tau}^m\left(X_i\right) - \tau(X_i) \right| \right]^{1+\alpha} + c_7 \left(\frac{v}{n} \right)^{\frac{1+\alpha}{2+\alpha}} t, \tag{A.35}$$

where constants c_6 and c_7 depend only on $(M, \kappa, \eta, \alpha)$.

Using the upper bound derived in (A.35), we obtain, for $t \geq 1$,

$$\begin{aligned} & E_{P^n} \left(d(G^*_{FB}, \hat{G}_{m-hybrid}) \right) \\ = \; & E_{P^n} \left(d(G^*_{FB}, \hat{G}_{m-hybrid}) 1\left\{ \Lambda_1 \cap \Lambda_2 \cap \Omega_t \right\} \right) + E_{P^n} \left(d(G^*_{FB}, \hat{G}_{m-hybrid}) 1\left\{ \Lambda_1^c \cup \Lambda_2^c \cup \Omega_t^c \right\} \right) \\ \leq \; & c_6 E_{P^n} \left(\left[\max_{1\leq i\leq n} \left| \hat{\tau}^m\left(X_i\right) - \tau(X_i) \right| \right]^{1+\alpha} \right) + c_7 \left(\frac{v}{n} \right)^{\frac{1+\alpha}{2+\alpha}} t + P^n\left(\Lambda_1^c\right) \\ & + E_{P^n} \left(d(G^*_{FB}, \hat{G}_{m-hybrid}) 1\{\Lambda_2^c\} \right) + P^n\left(\Omega_t^c\right) \\ \leq \; & \underbrace{c_6 \tilde{\psi}_n^{-(1+\alpha)} E_{P^n} \left(\left[\tilde{\psi}_n \max_{1\leq i\leq n} \left| \hat{\tau}^m\left(X_i\right) - \tau(X_i) \right| \right]^{1+\alpha} \right)}_{A_{1,n}} + \underbrace{c_7 \left(\frac{v}{n} \right)^{\frac{1+\alpha}{2+\alpha}} t}_{A_{2,n}} \\ & + \underbrace{\left(\frac{C_4}{\sqrt{2v}} \right)^{2v} n^{-2v\left(\lambda - \frac{1}{2}\right)} \exp\left(-n^{-2\left(\lambda - \frac{1}{2}\right)} \right)}_{A_{3,n}} + \underbrace{n^{-\lambda}}_{A_{4,n}} + \underbrace{\exp(-t)}_{A_{5,n}}, \end{aligned}$$

where $\tilde{\psi}_n$ is a sequence as specified in equation (2.10) in the main text. In these inequalities, the third line uses (A.35) and $d(G^*_{FB}, \hat{G}_{m-hybrid}) \leq 1$. In the fourth line, $A_{3,n}$ follows from Lemma A.9, $A_{4,n}$ follows from $d(G^*_{FB}, \hat{G}_{m-hybrid}) \leq P_X \left(G^*_{FB} \triangle \hat{G}_{m-hybrid} \right)$ and $P_X \left(G^*_{FB} \triangle \hat{G}_{m-hybrid} \right) < n^{-\lambda}$ on Λ_2^c, and $A_{5,n}$ follows from (A.34).

We now discuss convergence rates of $A_{j,n}$, $j = 1, \ldots, 5$, individually with suitable choices of t and λ. Equation (2.10) assumed in this theorem implies

$$\begin{aligned} & \sup_{P \in \mathcal{P}_m} E_{P^n}\left(\left(\tilde{\psi}_n \max_{1 \leq i \leq n} \left|\hat{\tau}^m(X_i) - \tau(X_i)\right|\right)^{1+\alpha}\right) \\ = & \sup_{P \in \mathcal{P}_m} E_{P^n}\left(\left[\left(\tilde{\psi}_n \max_{1 \leq i \leq n} \left|\hat{\tau}^m(X_i) - \tau(X_i)\right|\right)^2\right]^{\frac{1+\alpha}{2}}\right) \\ \leq & \left(\left[\sup_{P \in \mathcal{P}_m} E_{P^n}\left(\tilde{\psi}_n \max_{1 \leq i \leq n} \left|\hat{\tau}^m(X_i) - \tau(X_i)\right|\right)^2\right]^{\frac{1+\alpha}{2}}\right) \\ = & O(1), \end{aligned}$$

where the third line follows from Jensen's inequality since $(1+\alpha)/2 \leq 1$. Hence, $A_{1,n}$ satisfies $\sup_{P \in \mathcal{P}_m} A_{1,n} = O\left(\tilde{\psi}_n^{-(1+\alpha)}\right)$. By setting $t = (1+\alpha)\log \psi_n$, we can make the convergence rate of $A_{5,n}$ equal to that of $A_{1,n}$. At the same time, by choosing $\lambda > \frac{1+\alpha}{2+\alpha} \geq \frac{1}{2}$, we can make $A_{3,n}$ and $A_{4,n}$ converge faster than $A_{2,n}$. Hence, the uniform convergence rate of $E_{P^n}\left(d(G^*_{FB}, \hat{G}_{m-hybrid})\right)$ over $P \in \mathcal{P}_m \cap \mathcal{P}_{FB}(M, \kappa, \eta, \alpha)$ is bounded by the convergence rates of the $A_{1,n}$ and $A_{2,n}$,

$$O\left(\sup_{P \in \mathcal{P}_m} A_{1,n} \vee \sup_{P \in \mathcal{P}_{FB}(M,\kappa,\eta,\alpha)} A_{2,n}\right) = O\left(\tilde{\psi}_n^{-(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}} \log \tilde{\psi}_n\right).$$

This completes the proof for the m-hybrid case.

A proof for the e-hybrid case follows almost identically to the proof of the m-hybrid case. The differences are that ρ_n in inequality (A.32) is given by

$$\rho_n = \frac{\kappa}{M} \max_{1 \leq i \leq n} \left|\hat{\tau}^e_i - \tau_i\right| P_{X,n}\left(G^*_{FB} \triangle \hat{G}_{e-hybrid}\right).$$

and that inequality (A.33) is replaced by

$$d(G^*_{FB}, \hat{G}_{e-hybrid}) \leq \rho_n + V_a(d(G^*_{FB}, \hat{G}_{e-hybrid}) + a^2), \tag{A.36}$$

where V_a is as defined in the proof of Theorem 2.3. The rest of the proof goes similarly to the proof of the first claim except that the rate $\tilde{\phi}_n$ given in equation (2.11) replaces $\tilde{\psi}_n$ in the first claim. □

B Inference for Welfare Gain

In the proposed EWM procedure, the maximized empirical welfare $W_n(\hat{G}_{EWM})$ can be seen as an estimate of $W(\hat{G}_{EWM})$, the welfare level attained by implementing the estimated treatment rule.[1]

[1] It is important to note that in finite samples, $W_n\left(\hat{G}_{EWM}\right)$ estimates $W(\hat{G}_{EWM})$ with an upward bias. With fixed n, the size of the bias becomes bigger as $\mathcal{G}$ becomes more complex.

In situations where propensity scores are known, this section provides a procedure for constructing asymptotically valid confidence intervals for the population welfare gain of implementing the estimated rule.

Let $\hat{G} \in \mathcal{G}$ be an estimated treatment rule such as $\hat{G}_{EWM}$ or other data-driven way of selecting G from the set of candidate policies. Define the welfare gain of implementing the estimated treatment rule $\hat{G} \in \mathcal{G}$ by

$$V(\hat{G}) \equiv W(\hat{G}) - W(G_0),$$

where G_0 is a benchmark treatment assignment rule with which the estimated treatment rule $\hat{G}$ is compared in terms of the social welfare. For instance, if the estimated treatment rule $\hat{G}$ is compared with the "no treatment" case, G_0 is the empty set $\emptyset$. Alternatively, if a benchmark policy is the non-individualized uniform adoption of the treatment, G_0 is set at $G_0 = \mathcal{X}$, and $V(\hat{G})$ is interpreted as the welfare gain of implementing individualized treatment assignment instead of the non-individualized implementation of the treatment.

A construction of one-sided confidence intervals for $V(\hat{G})$ proceeds as follows. Let $\nu_n(G) = \sqrt{n}\left(V_n(G) - V(G)\right)$, where $V_n(G) \equiv W_n(G) - W_n(G_0)$. If there is a random variable $\bar{\nu}_n$ such that $\nu_n\left(\hat{G}\right) \leq \bar{\nu}_n$ holds P^n-almost surely, and if $\bar{\nu}_n$ converges in distribution to a non-degenerate random variable $\bar{\nu}$, then, with $q_{\bar{\nu}}(1-\bar{\alpha})$, the $(1-\bar{\alpha})$-th quantile of $\bar{\nu}$, it holds

$$P^n\left(\nu_n\left(\hat{G}\right) \leq q_{\bar{\nu}}(1-\bar{\alpha})\right) \geq P^n\left(\bar{\nu}_n \leq q_{\bar{\nu}}(1-\bar{\alpha})\right) \to \Pr\left(\bar{\nu} \leq q_{\bar{\nu}}(1-\bar{\alpha})\right) = 1-\bar{\alpha}, \text{ as } n \to \infty.$$

Hence, if $\hat{q}_{\bar{\nu}}(1-\bar{\alpha})$, a consistent estimator of $q_{\bar{\nu}}(1-\bar{\alpha})$, is available, an asymptotically valid one-sided confidence interval for $V(\hat{G})$ with coverage probability $(1-\bar{\alpha})$ can be given by

$$\left[V_n\left(\hat{G}\right) - \frac{\hat{q}_{\bar{\nu}}(1-\bar{\alpha})}{\sqrt{n}}, \infty\right). \tag{B.1}$$

Two-sided confidence intervals for $V(\hat{G})$ can be constructed similarly by considering a random variable $\tilde{\nu}_n$ that satisfies $|\nu_n(\hat{G})| \leq \tilde{\nu}_n$, P^n-almost surely, and converges to a nondegenerate random variable $\tilde{\nu}$. With $\hat{q}_{\tilde{\nu}}(1-\bar{\alpha})$ a consistent estimator for the $(1-\bar{\alpha})$-th quantile of $\tilde{\nu}$, two sided confidence interval for $V(\hat{G})$ can be given by

$$\left[V_n\left(\hat{G}\right) - \frac{\hat{q}_{\tilde{\nu}}(1-\bar{\alpha})}{\sqrt{n}}, V_n\left(\hat{G}\right) + \frac{\hat{q}_{\tilde{\nu}}(1-\bar{\alpha})}{\sqrt{n}}\right]. \tag{B.2}$$

In the algorithm summarized below, we specify $\bar{\nu}_n$ to be $\bar{\nu}_n = \sqrt{n}\sup_{G \in \mathcal{G}}\left(V_n(G) - V(G)\right)$ and $\tilde{\nu}_n$ to be $\tilde{\nu}_n = \sqrt{n}\sup_{G \in \mathcal{G}}\left|V_n(G) - V(G)\right|$, and estimate the $(1-\bar{\alpha})$-quantiles of their asymptotic distributions by bootstrapping the centered empirical processes.[2]

Algorithm B.1. *1. Let $\hat{G} \in \mathcal{G}$ be an estimated treatment assignment rule (e.g., EWM rule), and $V_n(\cdot) = W_n(\cdot) - W_n(G_0)$ be the empirical welfare gain obtained from the original sample.*

2. Resample n-observations of $Z_i = (Y_i, D_i, X_i)$ randomly with replacement from the original sample and construct the bootstrap analogue of the welfare gain, $V_n^*(\cdot) = W_n^*(\cdot) - W_n^*(G_0)$, where $W_n^*(\cdot)$ is the empirical welfare of the bootstrap sample.

3. For one-sided confidence intervals, compute $\bar{\nu}_n^* = \sqrt{n}\sup_{G\in\mathcal{G}}(V_n^*(G) - V_n(G))$. For two-sided confidence intervals, compute $\tilde{\nu}_n^* = \sqrt{n}\sup_{G\in\mathcal{G}}|V_n^*(G) - V_n(G)|$.

4. Let $\bar{\alpha} \in (0, 1/2)$. Repeat step 2 and 3 many times. For one-sided (two-sided) confidence intervals, obtain $\hat{q}_{\bar{\nu}}(1-\bar{\alpha})$ $(\hat{q}_{\tilde{\nu}}(1-\bar{\alpha}))$ by the empirical $(1-\bar{\alpha})$-th quantile of the bootstrap realizations of $\bar{\nu}_n^*$ $(\tilde{\nu}_n^*)$.

Given Assumption 2.1, the uniform central limit theorem for empirical processes assures that $\bar{\nu}_n$ and $\tilde{\nu}_n$ converge in distribution to the supremum of mean zero Brownian bridge processes and the supremum of their absolute values, respectively. Furthermore, by the well-known result on the asymptotic validity of the bootstrap empirical processes (see, e.g., Section 3.6 of van der Vaart and Wellner (1996)), the bootstrap critical values $\hat{q}_{\bar{\nu}}(1-\bar{\alpha})$ and $\hat{q}_{\tilde{\nu}}(1-\bar{\alpha})$ consistently estimate the corresponding quantiles of the limiting distributions of $\bar{\nu}_n$ and $\tilde{\nu}_n$, respectively. We can therefore assure that the confidence intervals constructed in (B.1) and (B.2) have the desired asymptotic coverage probability.

The same inference procedure is valid for the welfare gain estimated with demeaned outcomes $V_n^{dm}(\hat{G}) \equiv W_n^{dm}(\hat{G}) - W_n^{dm}(G_0)$. Resampling in this case is from observations $Z_i^{dm} = \left(Y_i^{dm}, D_i, X_i\right)$, with outcomes $Y_i^{dm} = Y_i - E_n[Y_i]$ demeaned by the outcome mean in the original sample.

[2] The current choices of $\bar{\nu}_n$ and $\tilde{\nu}_n$ are likely to yield conservative confidence intervals. Keeping the same nominal coverage probability, it is feasible to tighten up the confidence intervals with more sophisticated choices of $\bar{\nu}_n$ and $\tilde{\nu}_n$, such as $\bar{\nu}_n = \sqrt{n}\sup_{G\in\hat{\mathcal{G}}}(V_n(G) - V(G))$ and $\tilde{\nu}_n = \sqrt{n}\sup_{G\in\hat{\mathcal{G}}}|V_n(G) - V(G)|$, where $\hat{\mathcal{G}}$ is a data-dependent subclass of $\mathcal{G}$ that contains $\hat{G}$ with probability approaching one. Such $\hat{\mathcal{G}}$ can be obtained by applying the technique of contact set estimation in the context of stochastic dominance testing. See Linton et al. (2010) and Donald and Hsu (2016), as well as the literature on moment inequalities with moment selection (Andrews and Shi (2013), among others).

C Computing EWM Treatment Rules

The Empirical Welfare Maximization rule $\hat{G}_{EWM}$, as well as hybrid rules $\hat{G}_{m-hybrid}$, and $\hat{G}_{e-hybrid}$, share the same structure

$$\hat{G} \in \arg\max_{G \in \mathcal{G}} \sum_{i=1}^{n} g_i \cdot 1\left\{X_i \in G\right\}, \tag{C.1}$$

where each g_i is a function of the data, i.e., for the EWM rule $\hat{G}_{EWM}$, $g_i = \frac{1}{n}\left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right)$, for the e-hybrid rule $\hat{G}_{e-hybrid}$, $g_i = \hat{\tau}_i^e/n$, and for the m-hybrid rule $\hat{G}_{m-hybrid}$, $g_i = \hat{\tau}^m(X_i)/n$. The objective function in (C.1) is non-convex and discontinuous in G, thus finding $\hat{G}$ could be computationally challenging. In this section, we propose a set of convenient tools that permit solving this optimization problem and performing inference using widely available software for practically important classes of sets $\mathcal{G}$ defined by linear eligibility scores.[3]

C.1 Single Linear Index Rules

We start with the problem of computing optimal treatment rules that assign treatments based on a linear index (linear eligibility score; LES, see Examples 2.1 and 2.2). To reduce notational complexity, we include a constant in the covariate vector X throughout the exposition of this section. An LES rule can be expressed as $1\{X^T\beta \geq 0\}$. This type of treatment rule is commonly used in practice because it offers a simple way to reduce the dimension of observable characteristics. Furthermore, it is easy to enforce monotonicity of treatment assignment in specific covariates by imposing sign restrictions on the components of β.

Let $\mathcal{G}_{LES}$ be a collection of half-spaces of the covariate space $\mathcal{X}$, which are the upper contour sets of linear functions:

$$\begin{aligned} \mathcal{G}_{LES} &= \left\{G_\beta : \beta \in \mathbf{B} \subset \mathbb{R}^{d_x+1}\right\}, \\ G_\beta &= \left\{x : x^T\beta \geq 0\right\}. \end{aligned}$$

Then the optimization problem (C.1) becomes:

$$\max_{\beta \in \mathbf{B}} \sum_{i=1}^{n} g_i \cdot 1\left\{X_i^T\beta \geq 0\right\}. \tag{C.2}$$

This problem is similar to the maximum weighted score problem analyzed in Florios and Skouras (2008). They observe that the maximum score objective function could be rewritten as a Mixed

[3] For the empirical illustration we used IBM ILOG CPLEX Optimization Studio, which is available free for academic use through the IBM Academic Initiative.

Integer Linear Programming problem with additional binary parameters $(z_1, ..., z_n)$ that replace the indicator functions $1\left\{X_i^T\beta \geq 0\right\}$. The equality $z_i = 1\left\{X_i^T\beta \geq 0\right\}$ is imposed by a combination of linear inequality constraints and the restriction that z_i's are binary. The advantage of a MILP representation is that it is a standard optimization problem that could be solved by multiple commercial and open-source solvers. The branch-and-cut algorithms implemented in these solvers are faster than brute force combinatorial optimization.

We propose replacing (C.2) by its equivalent problem:

$$\max_{\substack{\beta\in\mathbf{B},\\ z_1,\dots,z_n\in\mathbb{R}}} \quad \sum_{i=1}^{n} g_i \cdot z_i \tag{C.3}$$

$$\text{s.t.} \quad \frac{X_i^T\beta}{C_i} < z_i \leq 1 + \frac{X_i^T\beta}{C_i} \text{ for } i = 1, \dots, n, \tag{C.4}$$

$$z_i \in \{0, 1\},$$

where constants C_i should satisfy $C_i > \sup_{\beta\in\mathbf{B}} |X_i^T\beta|$. Then the inequality constraints (C.4) and the restriction that z_i's are binary imply that $z_i = 1$ if and only if $X_i^T\beta \geq 0$. It follows that the maximum value of (C.4) for each value of β is the same as the value of (C.2).

The problem (C.3) is a linear optimization problem with linear inequality constraints and integer constraints on z_i's if the set $\mathbf{B}$ is defined by linear inequalities that could be passed to any MILP solver. Florios and Skouras (2008) impose only one side of the inequality constraint (C.4) for each i. For $g_i > 0$, it is sufficient to impose only the upper bound on z_i and for $g_i < 0$ only the lower bound. The other side of the bound is always satisfied by the solution due to the direction of the objective function.

Our formulation has significant advantages. Despite a larger number of inequalities, it reduces the computation time in our applications by a factor of 10-40. Furthermore, it is not sufficient to impose only one side of the inequalities on z_i's for optimization with a capacity constraint considered further below.

Our data contains large sets of observations that differ from each other in only one covariate. Suppose that m observations $i_1, \dots, i_m$ differ only in the value of the last covariate: $X_{i_1} = (1, \tilde{x}_1, \dots, \tilde{x}_{d_x-1}, x_{d_x,i_1}), \dots, X_{i_m} = (1, \tilde{x}_1, \dots, \tilde{x}_{d_x-1}, x_{d_x,i_m})$, and are ordered with $x_{d_x,i_1} \leq x_{d_x,i_2} \leq \cdots \leq x_{d_x,i_m}$. Then the solution must satisfy either $z_{i_1} \leq z_{i_2} \leq \cdots \leq z_{i_m}$ or $z_{i_1} \geq z_{i_2} \geq \cdots \geq z_{i_m}$. We found it advantageous to split the optimization problem in our empirical application into two: one explicitly imposing $z_{i_1} \leq \cdots \leq z_{i_m}$ and one explicitly imposing $z_{i_1} \geq \cdots \geq z_{i_m}$ for sets of observations that have the same values of education, but different values of prior earnings.

Inference on the welfare gain $V(\hat{G}_{EWM})$ of the empirical welfare maximizing policy requires computing $\bar{\nu}_n^* = \sup_{G\in\mathcal{G}} \sqrt{n}\left(V_n^*(G) - V_n(G)\right)$ in each bootstrap sample. Denoting the bootstrap weights by $\{w_i^*\}$, $\sum_{i=1}^n w_i^* = n$, $\bar{\nu}_n^*$ could be expressed as

$$\bar{\nu}_n^* = \sqrt{n} \sup_{G \in \mathcal{G}} \sum_{i=1}^n (w_i^* - 1) g_i \cdot 1\left\{X_i^T \beta \geq 0\right\} \tag{C.5}$$

The optimization problem for $\bar{\nu}_n^*$ is analogous to the optimization problem for $\hat{G}_{EWM}$. Furthermore, solving it does not require the knowledge of $\hat{G}_{EWM}$, hence all bootstrap computations could be performed in parallel with the main EWM problem.

C.2 Multiple Linear Index Rules

We extend this method to compute treatment rules based on multiple linear scores. These rules construct J scores that are linear in covariates (or in their functions) and assign an individual to treatment if each score exceeds a specific threshold. An example of a multiple index treatment rule with three indices is when an individual is assigned to a job training program if (25 $\leq$ age $\leq$ 35) AND (wage at the previous job < \$15). The results are easily extended to treatment rules that apply if any of the indices exceeds its threshold, for example, (age $\geq$ 40) OR (length of unemployment $\geq$ 2 years).

Let the treatment assignment set G be defined as an intersection of upper contour sets of J linear functions:

$$\begin{aligned} \mathcal{G} &= \left\{G_{\beta^1,\ldots,\beta^J}, \beta^1, \ldots, \beta^J \in \mathbf{B}\right\}, \\ G_{\beta^1,\ldots,\beta^J} &= \left\{x : x^T \beta^1 \geq 0, \ldots, x^T \beta^J \geq 0\right\}. \end{aligned}$$

Then the optimization problem (C.1) becomes

$$\max_{\beta^1,\ldots,\beta^J \in \mathbf{B}} \sum_{i=1}^n g_i \cdot 1\{X_i^T \beta^1 \geq 0, \ldots, X_i^T \beta^J \geq 0\}. \tag{C.6}$$

We propose its equivalent formulation as a MILP problem with auxiliary binary variables $\left\{(z_i^1, \ldots, z_i^J, z_i^*), i = 1, \ldots, n\right\}$:

$$\max_{\substack{\beta^1,\ldots,\beta^J \in \mathbf{B}, \\ z_i^1,\ldots,z_i^J, z_i^* \in \mathbb{R}}} \quad \sum_{i=1}^n g_i \cdot z_i^* \tag{C.7}$$

$$\text{s.t.} \quad \frac{X_i^T \beta^j}{C_i} < z_i^j \leq 1 + \frac{X_i^T \beta^j}{C_i} \text{ for } 1 \leq i \leq n, 1 \leq j \leq J, \tag{C.8}$$

$$1 - J + \sum_{j=1}^J z_i^j \leq z_i^* \leq J^{-1} \sum_{j=1}^J z_i^j \text{ for } 1 \leq i \leq n, \tag{C.9}$$

$$z_i^1, \ldots, z_i^J, z_i^* \in \{0, 1\} \text{ for } 1 \leq i \leq n.$$

Similarly to the single index problem, the inequalities (C.8) and the constraint that z_i^j's are binary imply together that $z_i^j = 1\{X_i^T\beta^j \geq 0\}$. Linear inequalities (C.9) and the binary constraints imply together that

$$z_i^* = z_i^1 \cdot ... \cdot z_i^J = 1\{X_i^T\beta^1 \geq 0\} \cdot ... \cdot 1\{X_i^T\beta^J \geq 0\}.$$

The problem for a collection of sets defined by the union of linear inequalities

$$G_{\beta^1,\ldots,\beta^J} = \left\{X : X^T\beta^1 \geq 0 \text{ or } \ldots \text{or } X^T\beta^J \geq 0\right\}$$

could also be written as a MILP problem with the inequality constraint (C.9) replaced by

$$J^{-1}\sum_{j=1}^{J} z_i^j \leq z_i^* \leq \sum_{j=1}^{J} z_i^j \text{ for } i = 1, \ldots, n. \tag{C.10}$$

C.3 Optimization with a Capacity Constraint

When there is a capacity constraint K on the proportion of population that could be assigned to treatment 1, Empirical Welfare Maximization problem (2.4) on a set $\mathcal{G}$ of half-spaces becomes

$$\max_{\beta\in\mathbf{B}} \left[\min\left\{1, \frac{Kn}{\sum_{i=1}^n 1\{X_i^T\beta \geq 0\}}\right\} \sum_{i=1}^{n} g_i \cdot 1\left\{X_i^T\beta \geq 0\right\}\right]. \tag{C.11}$$

This problem cannot be rewritten as a linear optimization problem in the same way as (C.3) because the factor $\min\left\{1, \frac{Kn}{\sum_{i=1}^n 1\{X_i^T\beta\geq 0\}}\right\}$ varies with β. This factor could take fewer than n different values and the maximum of (C.11) could be obtained by solving a sequence of optimization problems each of which holds this factor constant.

$$\begin{array}{ll}
\text{For} & k = \lfloor Kn \rfloor, \ldots, n \\
\max\limits_{\substack{\beta\in\mathbf{B},\\ z_1,\ldots,z_n\in\mathbb{R}}} & \min\left\{1, \dfrac{Kn}{k}\right\} \displaystyle\sum_{i=1}^{n} g_i \cdot z_i \\
\text{s.t.} & \dfrac{X_i^T\beta}{C_i} < z_i \leq 1 + \dfrac{X_i^T\beta}{C_i} \text{ for } 1 \leq i \leq n, \\
& z_i \in \{0, 1\}, \\
& \displaystyle\sum_{i=1}^{n} z_i \leq k.
\end{array}$$

The capacity constrained problem with multiple indexes could be solved similarly.

References

Andrews, D. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Bousquet, O. (2002): "A Bennet Concentration Inequality and its Application to Suprema of Empirical Processes," *Comptes Rendus de l'Académie des Sciences - Series I*, 334, 495–500.

Donald, S. G. and Y.-C. Hsu (2016): "Improving the Power of Tests of Stochastic Dominance," *Econometric Reviews*, 35, 553–585.

Dudley, R. (1999): *Uniform Central Limit Theorems*, Cambridge University Press.

Florios, K. and S. Skouras (2008): "Exact Computation of Max Weighted Score Estimators," *Journal of Econometrics*, 146, 86–91.

Kanaya, S. and D. Kristensen (2014): "Optimal Sampling and Bandwidth Selection for Kernel Estimators of Diffusion Processes," *unpublished manuscript.*

Linton, O., K. Song, and Y. Whang (2010): "An Improved Bootstrap Test of Stochastic Dominance," *Journal of Econometrics*, 154, 186–202.

Lugosi, G. (2002): "Pattern Classification and Learning Theory," in *Principles of Nonparametric Learning*, ed. by L. Györfi, Vienna: Springer, 1–56.

Massart, P. and E. Nédélec (2006): "Risk Bounds for Statistical Learning," *The Annals of Statistics*, 34, 2326–2366.

Tsybakov, A. B. (2004): "Optimal Aggregation of Classifiers in Statistical Learning," *The Annals of Statistics*, 32, 135–166.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Online Appendix to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice"

Toru Kitagawa[*] and Aleksey Tetenov[†]

November 27, 2017

D Extensions

D.1 Empirical Welfare Maximization with a Capacity Constraint

This section shows a proof of the claim given in Remark 2.1 of the main text that says the expected welfare of $\hat{G}^K$ converges to the maximum at least at $n^{-1/2}$ rate. The result is analogous to Theorem 2.1, with the additional term corresponding to potential welfare losses due to estimation errors of $P_X(G)$.

Theorem D.1. *Under Assumption 2.1,*

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} W^K(G) - W^K(\hat{G}^K)\right] \leq C_1 \frac{M}{\kappa}\sqrt{\frac{v}{n}} + C_1 \frac{M}{K}\sqrt{\frac{v}{n}},$$

where C_1 is the universal constant in Lemma A.4.

Proof. Since $W^K(G) - W^K(G') = V^K(G) - V^K(G')$ for all G, G',

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} W^K(G) - W^K(\hat{G}^K)\right] = \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} V^K(G) - V^K(\hat{G}^K)\right], \tag{D.1}$$

and we focus on bounding the latter expression.

Since $\hat{G}^K$ maximizes $V_n^K(G)$, $V_n^K(\tilde{G}) \leq V_n^K(\hat{G}^K)$ for any $\tilde{G} \in \mathcal{G}$ and

$$\begin{aligned} V^K(\tilde{G}) &\leq V_n^K(\tilde{G}) + \sup_{G \in \mathcal{G}} \left|V_n^K(G) - V^K(G)\right| \\ &\leq V_n^K(\hat{G}^K) + \sup_{G \in \mathcal{G}} \left|V_n^K(G) - V^K(G)\right| \\ &\leq V^K(\hat{G}^K) + 2\sup_{G \in \mathcal{G}} \left|V_n^K(G) - V^K(G)\right|. \end{aligned}$$

Applying the inequality for all $\tilde{G} \in \mathcal{G}$, we obtain

$$\sup_{G \in \mathcal{G}} V^K(G) - V^K(\hat{G}^K) \leq 2\sup_{G \in \mathcal{G}} \left|V_n^K(G) - V^K(G)\right|,$$

[*] Cemmap/University College London, Department of Economics. Email: t.kitagawa@ucl.ac.uk

[†] University of Bristol, Email: a.tetenov@bristol.ac.uk

which is also true in expectation over P^n.

The welfare gain estimation error for any treatment rule G could be bounded from above by:

$$\begin{aligned}
&\left|V_n^K(G) - V^K(G)\right| = \left|\frac{K}{\max\{K, P_{X,n}(G)\}} \cdot V_n(G) - \frac{K}{\max\{K, P_X(G)\}} \cdot V(G)\right| \\
&\leq \frac{K}{\max\{K, P_{X,n}(G)\}} \cdot |V_n(G) - V(G)| + V(G) \cdot \left|\frac{K}{\max\{K, P_{X,n}(G)\}} - \frac{K}{\max\{K, P_X(G)\}}\right| \\
&\leq |V_n(G) - V(G)| + \frac{M}{K} \cdot |P_{X,n}(G) - P_X(G)|\,.
\end{aligned}$$

The second line comes from subtracting and adding $\frac{K}{\max\{K,P_{X,n}(G)\}}V(G)$ and then applying the triangle inequality. The third line uses inequalities $\frac{K}{\max\{K,P_{X,n}(G)\}} \leq 1$ and $V(G) \leq M$ (from Assumption 2.1 (BO)), and the observation that for any $a, b \in \mathbb{R}$ and $c > 0$,

$$\left|\frac{c}{\max\{c,a\}} - \frac{c}{\max\{c,b\}}\right| = \left|\frac{c(\max\{c,b\} - \max\{c,a\})}{\max\{c,a\} \cdot \max\{c,b\}}\right| \leq \frac{|\max\{c,b\} - \max\{c,a\}|}{c} \leq \frac{|b-a|}{c}.$$

Then

$$\begin{aligned}
&\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} V^K(G) - V^K(\hat{G}^K)\right] \leq 2 \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} \left|V_n^K(G) - V^K(G)\right|\right] \\
&\leq 2 \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} |V_n(G) - V(G)|\right] + 2\frac{M}{K} \sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} |P_{X,n}(G) - P_X(G)|\right]
\end{aligned}$$

Note that since the class $\mathcal{G}$ has VC-dimension $v < \infty$, the classes of functions

$$\begin{aligned}
f_G(Y, D, X) &\equiv \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right) \cdot 1\{X \in G\}, \\
h_G(Y, D, X) &\equiv 1\{X \in G\} - 1/2,
\end{aligned}$$

are VC-subgraph classes with VC-dimension no greater than v by Lemma A.1. These classes of functions are uniformly bounded by $M/(2\kappa)$ and $1/2$. Since $V_n(G) = E_n(f_G)$, $V(G) = E_P(f_G)$, $P_{X,n}(G) = E_n(h_G) + 1/2$ and $P_X(G) = E_P(h_G) + 1/2$, we could apply Lemma A.4 and obtain

$$\sup_{P \in \mathcal{P}(M,\kappa)} E_{P^n}\left[\sup_{G \in \mathcal{G}} V^K(G) - V^K(\hat{G}^K)\right] \leq C_1 \frac{M}{\kappa}\sqrt{\frac{v}{n}} + C_1 \frac{M}{K}\sqrt{\frac{v}{n}}.$$

The theorem's result follows from (D.1). □

D.2 Demeaned EWM

Define the demeaned population welfare as

$$W^{dm}(G) \equiv W(G) - E_P[Y],$$

then $\sup\limits_{G\in\mathcal{G}} W^{dm}(G) = \sup\limits_{G\in\mathcal{G}} W(G) - E_P[Y] = W^*_{\mathcal{G}} - E_P[Y]$. Analogously to (2.2), for any $\tilde{G}\in\mathcal{G}$,

$$W^{dm}(\tilde{G}) - W^{dm}(\hat{G}^{dm}_{EWM}) \le 2\sup_{G\in\mathcal{G}} \left| W^{dm}_n(G) - W^{dm}(G)\right|,$$

therefore

$$W^*_{\mathcal{G}} - W(\hat{G}^{dm}_{EWM}) \le 2\sup_{G\in\mathcal{G}} \left| W^{dm}_n(G) - W^{dm}(G)\right|.$$

Note that since $Y^{dm}_i = Y_i - E_n[Y_i]$,

$$\begin{aligned}
W^{dm}_n(G) &= E_n\left[\frac{Y^{dm}_i D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{Y^{dm}_i(1-D_i)}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right] \\
&= W_n(G) - E_n[Y_i]\cdot E_n\left[\frac{D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{1-D_i}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right],
\end{aligned}$$

and since $|E_n(Y_i)|\le M/2$,

$$\begin{aligned}
\left|W^{dm}_n(G) - W^{dm}(G)\right| &\le |W_n(G) - W(G)| \\
&+ \left|E_n[Y_i]\cdot E_n\left[\frac{D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{1-D_i}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right] - E_P[Y]\right| \\
&\le |W_n(G) - W(G)| \\
&+ |E_n(Y_i) - E_P[Y]| \\
&+ \frac{M}{2}\cdot\left|E_n\left[\frac{D_i}{e(X_i)}\cdot 1\{X_i\in G\} + \frac{1-D_i}{1-e(X_i)}\cdot 1\{X_i\notin G\}\right] - 1\right|.
\end{aligned}$$

Similarly to the proof of Theorem 2.1, Lemma A.4 applies to all three terms with envelopes $M/(2\kappa)$, $M/2$, and $M/(2\kappa)$, thus

$$E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G}^{dm}_{EWM})\right] \le 2E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|W^{dm}_n(G) - W^{dm}(G)\right|\right] \le C_1 M\left(\frac{2}{\kappa}+1\right)\sqrt{\frac{v}{n}}.$$

D.3 Multiple Treatments

It is feasible to extend the current approach to situations with multiple treatments. Suppose there are K treatments denoted by $D\in\{1,\dots,K\}$. Let $e_k(x) = P(D=k|X=x)$, $k=1,\dots,K$, be the propensity scores in the experimental data, and $\{Y_k : k=1,\dots,K\}$ be the potential outcomes for each treatment. Define a treatment assignment policy by a K-partition of the covariate space $\mathcal{X}$, $\mathbf{G} = (G_1,\dots,G_K)$, where $G_1,\cdots,G_K\subset\mathcal{X}$ are non-intersecting subsets that partition $\mathcal{X}$ into K regions. For each $k=1,\dots,K$, G_k specifies a subpopulation to which treatment $D=k$ is assigned.

Under unconfoundedness, $(Y_1, \ldots, Y_K) \perp D|X$, consider the following empirical welfare criterion;

$$W_n(\mathbf{G}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{Y_i \cdot 1\{D_i = k\}}{e_k(X_i)} \cdot 1\{X_i \in G_k\},$$

which unbiasedly estimates the population welfare attained by policy $\mathbf{G}$,

$$W(\mathbf{G}) = \sum_{k=1}^{K} E[Y_k \cdot 1\{X \in G_k\}].$$

Consider setting the space of policies to $\mathbb{G} = \{\mathbf{G} : G_1 \in \mathcal{G}, \ldots, G_K \in \mathcal{G}, \mathbf{G} \text{ partitions } \mathcal{X}\}$, where $\mathcal{G}$ is a VC-class of subsets in $\mathcal{X}$ including $\emptyset$ such that K distinct subsets in $\mathcal{G}$ can form a partition of $\mathcal{X}$. For instance, when $\mathcal{X} = \mathbb{R}$, a class of connected intervals of the form $\mathcal{G} = \{(x, x'] : -\infty \leq x \leq x' \leq \infty\} \cup \emptyset$ is a VC-class that allows us to pick K-distinct subsets partitioning $\mathbb{R}$. The EWM rule can be then obtained as $\hat{\mathbf{G}}_{EWM} \in \arg\max_{\mathbf{G} \in \mathbb{G}} W(\mathbf{G})$.

Analogous to derivation of inequality (2.3) in the paper, we can bound the welfare loss of the EWM rule as

$$\sup_{\mathbf{G} \in \mathbb{G}} W(\mathbf{G}) - W(\hat{\mathbf{G}}_{EWM}) \leq \sum_{k=1}^{K} 2 \sup_{G_k \in \mathcal{G}} \left| W_n^k(G_k) - W^k(G_k) \right|,$$

where $W_n^k(G_k) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \cdot 1\{D_i = k\}}{e_k(X_i)} \cdot 1\{X_i \in G_k\}$ and $W^k(G_k) \equiv E[Y_k \cdot 1\{X \in G_k\}]$. Assuming bounded outcomes $Y \in [-M/2, M/2]$ and strict overlap, in the sense that $e_k(x) \in [\kappa, 1 - \kappa]$ for all x and $k = 1, \ldots, K$ for some $\kappa > 0$, we apply Lemmas A.1 and A.4 to obtain the mean of $\sup_{G_k \in \mathcal{G}} \left| W_n^k(G_k) - W^k(G_k) \right|$ bounded from above by $C_1 M \sqrt{v/n}/\kappa$. Hence, the whole welfare loss can be bounded from above by that of Theorem 2.1 multiplied by the number of treatments K.

Computing $\hat{\mathbf{G}}_{EWM}$ presents additional challenges when the EWM framework is extended from binary to multiple treatment case. We leave an investigation of computational procedures in this setting for future research.

D.4 Comparison with the Nonparametric Plug-in Rule

The plug-in treatment choice rule (1.13) with parametrically or nonparametrically estimated $m_1(x)$ and $m_0(x)$ is intuitive and simple to implement. In situations where flexible treatment assignment rules are allowed and the dimension of conditioning covariates is small, the nonparametric plug-in rule would be a competing alternative to the EWM approach. In this section, we review the welfare loss convergence rate results of the nonparametric plug-in rule and discuss potential advantages and disadvantages of these two approaches.

We denote the class of data generating processes that satisfy Assumptions 2.1 (UCF), (BO), (SO), Assumption 2.2 (MA), and Assumption E.1 by $\mathcal{P}_{smooth}\left(M, \kappa, \alpha, \eta, \beta_m\right)$. Given the smoothness assumption of the regression equations, we consider estimating m_1 and m_0 by local polynomial estimators of degree $(\beta_m - 1)$. The convergence rate results of the nonparametric plug-in classifiers shown in Theorem 3.3 of Audibert and Tsybakov (2007) can be straightforwardly extended to the treatment choice context, resulting in

$$\sup_{P \in \mathcal{P}_{smooth}(M,\kappa,\alpha,\eta,\beta_m)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G}_{plug-in})\right] \leq O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_m}}\right). \tag{D.2}$$

Furthermore, if $\alpha\beta_m \leq d_x$, Theorem 3.5 of Audibert and Tsybakov (2007) applied to the current treatment choice setup shows that the nonparametric plug-in rule attains the rate lower bound i.e., for any treatment rule $\hat{G}$,

$$\sup_{P \in \mathcal{P}_{smooth}(M,\kappa,\alpha,\eta,\beta_m)} E_{P^n}\left[W(G^*_{FB}) - W(\hat{G})\right] \geq O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_m}}\right)$$

holds.

In practically relevant situations where $\alpha\beta_m \leq d_x$,[1] a naive comparison of the welfare loss convergence rate of the plug-in rule presented here with that of EWM (Theorems 2.3 and 2.4) would suggest that in terms of the welfare loss converge rate, the EWM rule would outperform the nonparametric plug-in rule. It is, however, important to notice that the classes of data generating processes over which the uniform rates are ensured differ between the two cases. $\mathcal{P}_{smooth}\left(M, \kappa, \alpha, \eta, \beta_m\right)$ is constrained by smooth regression equations and continuously distributed X, whereas $\mathcal{P}_{FB}\left(M, \kappa, \alpha, \eta\right)$ considered in Theorems 2.3 and 2.4 allows for discontinuous regression equations and no restriction on the marginal distribution of X's. Assumption 2.2 (FB) on $\mathcal{P}_{FB}\left(M, \kappa, \alpha, \eta\right)$ requires that $\{x : \tau(x) \geq 0\}$ belongs to the pre-specified VC-class $\mathcal{G}$, whereas $\mathcal{P}_{smooth}\left(M, \kappa, \alpha, \eta, \beta_m\right)$ is free from such assumption. This non-nested relationship between $\mathcal{P}_{FB}\left(M, \kappa, \alpha, \eta\right)$ and $\mathcal{P}_{smooth}\left(M, \kappa, \alpha, \eta, \beta_m\right)$ makes the naive rate comparison between (D.2) and Theorem 2.3 less meaningful because a data generating process in $\mathcal{P}_{smooth}\left(M, \kappa, \alpha, \eta, \beta_m\right)$ that yields the slowest convergence rate for the nonparametric plug-in rule is in fact excluded from $\mathcal{P}_{FB}\left(M, \kappa, \alpha, \eta\right)$. Accordingly, unless we can assess which one of $\mathcal{P}_{smooth}\left(M, \kappa, \alpha, \eta, \beta_m\right)$ and $\mathcal{P}_{FB}\left(M, \kappa, \alpha, \eta\right)$ is more

[1] In an analogy to the Proposition 3.4 of Audibert and Tsybakov (2007), when the class of data generating processes is assumed to have $\alpha\beta_m > d_x$, no data generating process in this class can have the conditional treatment effect $\tau(x) = 0$ in an interior of the support of P_X. In the practice of causal inference, we a priori would not restrict the plausible data generating processes only to these extreme cases; therefore, the class of data generating processes with $\alpha\beta > d_x$ would be less relevant in practice.

likely to contain the true data generating process, these rate results offer us limited guidance on the procedure that should be used in a given application.

In practical terms, we consider these two distinct approaches as complementary, and our choice between them should be based on available assumptions and the dimension of covariates in a given application. With knowledge of the propensity score, a practical advantage of the EWM rule is that the welfare loss convergence rate does not directly depend on the dimension of X, so when an available credible assumption on the level set $\{x : \tau(x) \geq 0\}$ implies a certain class of decision sets with a finite VC-dimension, the EWM approach offers a practical solution to get around the curse of dimensionality of X. A potential drawback of using the EWM rule is the risk of misspecification of $\mathcal{G}$, i.e., if Assumption 2.2 (FB) is not valid, the EWM rule only attains the second-best welfare, whereas the nonparametric plug-in rule is guaranteed to yield the first-best welfare in the limit. Another aspect of comparison is that the performance of the EWM rule is stable regardless of whether the underlying data generating processes, including the marginal distribution of X and the regression equations $m_1(X)$ and $m_0(X)$, are smooth or not. In terms of implementation, the EWM approach becomes particularly attractive when the class of candidate decision sets $\mathcal{G}$ is given exogenously, since the user does not have to specify any smoothing parameter in this case. In contrast, when the user can freely choose $\mathcal{G}$, the welfare performance of the EWM rule can be sensitive to how to choose $\mathcal{G}$, similarly to that the performance of nonparametric plug-in rule can be sensitive to the choice of the smoothing parameter.

E Hybrid EWM with Local Polynomial Estimators

This section focuses on the hybrid EWM approaches with local polynomial estimators for $\tau(x)$ and $e(x)$,. We spell out classes of data generating processes $\mathcal{P}_m$ and $\mathcal{P}_e$ as well as ψ_n, $\tilde{\psi}_n$, ϕ_n, and $\tilde{\phi}_n$ that satisfy Condition 2.1 and the assumption of Theorem 2.6.

E.1 Assumptions, Estimators, and Welfare Convergence Rates

Consider the m-hybrid approach in which the leave-one-out local polynomial estimators are used to estimate $m_1(X_i)$ and $m_0(X_i)$, i.e., $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ are constructed by fitting the local polynomials excluding the i-th observation.[2] For any multi-index $s = (s_1, \ldots, s_{d_x}) \in \mathbb{N}^{d_x}$ and any $(x_1, \ldots, x_{d_x}) \in \mathbb{R}^{d_x}$, we define $|s| \equiv \sum_{i=1}^{d_x} s_i$, $s! \equiv s_1! \cdots s_{d_x}!$, $x^s \equiv x_1^{s_1} \cdots x_{d_x}^{s_{d_x}}$, and $\|x\| \equiv$

[2] The reason to consider the leave-one-out fitted values is to simplify analytical verification of Condition 2.1. We believe that the welfare loss convergence rates of the hybrid approaches will not be affected even when the i-th observation is included in estimating $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$.

$\left(x_1^2 + \cdots + x_{d_x}^2\right)^{1/2}$. Let $K(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}$ be a kernel function and $h > 0$ be a bandwidth. At each X_i, $i = 1, \ldots, n$, we define the leave-one-out local polynomial coefficient estimators with degree $l \geq 0$ as

$$\begin{aligned} \hat{\theta}_1(X_i) &= \arg\min_{\theta} \sum_{j \neq i, D_j = 1} \left[Y_j - \theta^T U\left(\frac{X_j - X_i}{h}\right)\right]^2 K\left(\frac{X_j - X_i}{h}\right), \\ \hat{\theta}_0(X_i) &= \arg\min_{\theta} \sum_{j \neq i, D_j = 0} \left[Y_j - \theta^T U\left(\frac{X_j - X_i}{h}\right)\right]^2 K\left(\frac{X_j - X_i}{h}\right), \end{aligned}$$

where $U\left(\frac{X_j - X_i}{h}\right)$ is the vector with elements indexed by the multi-index s, i.e., $U\left(\frac{X_j - X_i}{h}\right) \equiv \left(\left(\frac{X_j - X_i}{h}\right)^s\right)_{0 \leq |s| \leq l}$.[3] Note that $U(0)$ gives vector $(1, 0, \ldots, 0)^T$. Let $\lambda_{n,1}(X_i)$ be the smallest eigenvalue of $B_1(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \neq i, D_j = 1} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$ and $\lambda_{n,0}(X_i)$ be the smallest eigenvalue of $B_0(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \neq i, D_j = 0} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$. Accordingly, we construct leave-one-out local polynomial fits for $m_1(X_i)$ and $m_0\left(X_i\right)$ by

$$\begin{aligned} \hat{m}_1(X_i) &= U^T(0)\hat{\theta}_1(X_i) \cdot 1\left\{\lambda_{n,1}(X_i) \geq t_n\right\}, \\ \hat{m}_0\left(X_i\right) &= U^T(0)\hat{\theta}_0(X_i) \cdot 1\left\{\lambda_{n,0}(X_i) \geq t_n\right\}, \end{aligned}$$

where t_n is a positive sequence that slowly converges to zero, such as $t_n \propto (\log n)^{-1}$. These trimming rules regularize the regressor matrices of the local polynomial regressions and simplify the proof of the uniform consistency of the local polynomial estimators.

To characterize $\mathcal{P}_m$ in Condition 2.1, we impose the following restrictions.

Assumption E.1.

(Smooth-m) Smoothness of the Regressions: The regression equations $m_1(\cdot)$ and $m_0(\cdot)$ belong to a Hölder class of functions with degree $\beta_m \geq 1$ and constant $L_m < \infty$.[4]

(PX) Support and Density Restrictions on P_X: Let $\mathcal{X} \subset \mathbb{R}^{d_x}$ be the support of P_X. Let *Leb*$(\cdot)$ be the Lebesgue measure on $\mathbb{R}^{d_x}$. There exist constants $\underline{c}$ and r_0 such that

$$Leb\left(\mathcal{X} \cap B(x, r)\right) \geq \underline{c} Leb(B(x, r)) \quad \forall 0 < r \leq r_0, \forall x \in \mathcal{X}, \tag{E.1}$$

[3] We specify the same degree of polynomial and bandwidth for these two local polynomial regressions only to suppress notational burden.

[4] Let D^s denote the differential operator $D^s \equiv \frac{\partial^{s_1 + \cdots + s_{d_x}}}{\partial x_1^{s_1} \cdots x_{d_x}^{s_{d_x}}}$. Let $\beta \geq 1$ be an integer. For any $x \in \mathbb{R}^{d_x}$ and any $(\beta - 1)$ times continuously differentiable function $f : \mathbb{R}^{d_x} \to \mathbb{R}$, we denote the Taylor expansion polynomial of degree $(\beta - 1)$ at point x by $f_x(x') \equiv \sum_{|s| \leq \beta - 1} \frac{\left(x' - x\right)^s}{s!} D^s f(x)$. Let $L > 0$. The Hölder class of functions in $\mathbb{R}^{d_x}$ with degree β and constant $0 < L < \infty$ is defined as the set of function $f : \mathbb{R}^{d_x} \to \mathbb{R}$ that are $(\beta - 1)$ times continuously differentiable and satisfy, for any x and $x' \in \mathbb{R}^{d_x}$, the inequality $|f_x(x') - f(x)| \leq L \left\|x - x'\right\|^{\beta}$.

and P_X has the density function $\frac{dP_X}{dx}(\cdot)$ with respect to the Lebesgue measure of $\mathbb{R}^{d_x}$ that is bounded from above and bounded away from zero, $0 < \underline{p}_X \leq \frac{dP_X}{dx}(x) \leq \bar{p}_X < \infty$ for all $x \in \mathcal{X}$.
(Ker) Bounded Kernel with Compact Support: The kernel function $K(\cdot)$ have support $[-1,1]^{d_x}$, $\int_{\mathbb{R}^{d_x}} K(u)du = 1$, and $\sup_u K(u) \leq K_{\max} < \infty$.

Smoothness of the regression equations, Assumption E.1 (Smooth-m), is a standard assumption in the context of nonparametric regressions. Assumption E.1 (PX) is borrowed from Audibert and Tsybakov (2007), and it provides regularity conditions on the marginal distribution of X. Inequality condition (E.1) constrains the shape of the support of X, and it essentially rules out the case where $\mathcal{X}$ has "sharp" spikes, i.e., $\mathcal{X} \cap B(x,r)$ has an empty interior or $Leb(\mathcal{X} \cap B(x,r))$ converges to zero as $r \to 0$ faster than the rate of r^2 for some x in the boundary of $\mathcal{X}$.

Lemma E.4 below shows that when $\mathcal{P}_m$ consists of the data generating processes satisfying Assumption E.1 (Smooth-m) and (PX), Condition 2.1 (m) holds with $\psi_n = n^{\frac{1}{2+d_x/\beta_m}}$, and equation (2.10) in Theorem 2.6 holds with $\tilde{\psi}_n = n^{\frac{1}{2+d_x/\beta_m}}(\log n)^{-\frac{1}{2+d_x/\beta_m}-2}$. The following corollary therefore follows.

Corollary E.1. *Let $\mathcal{P}_m$ consist of data generating processes that satisfy Assumption E.1 (Smooth-m) and (PX). Let $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ be the leave-one-out local polynomial estimators with degree $l = (\beta_m - 1)$, whose kernels satisfy Assumption E.1 (Ker).*

(i) Suppose Assumption 2.1 holds and a bandwidth satisfies $h \propto n^{-\frac{1}{2\beta_m + d_x}}$. Then, it holds

$$\sup_{P \in \mathcal{P}_m \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{G}_{m-hybrid})\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_m}}\right).$$

(ii) Suppose Assumptions 2.1 and 2.2 hold with margin coefficient $\alpha \in (0,1]$, and a bandwidth satisfies $h \propto \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta_m + d_x}}$. Then, it holds

$$\begin{aligned}
&\sup_{P \in \mathcal{P}_m \cap \mathcal{P}_{FB}(M,\kappa,\alpha,\eta)} E_{P^n}\left[W(G_{FB}^*) - W(\hat{G}_{m-hybrid})\right] \\
\leq\ & O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_m}}(\log n)^{\left(\frac{1}{2+d_x/\beta_m}+2\right)(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}}\log n\right).
\end{aligned}$$

Next, consider the e-hybrid approach. For each $i = 1, \ldots, n$, define a leave-one-out local polynomial propensity score estimator as

$$\begin{aligned}
\hat{e}(X_i) &= U^T(0)\hat{\theta}_e(X_i) \cdot 1\{\lambda_n(X_i) \geq t_n\}, \\
\hat{\theta}_e(X_i) &= \arg\min_{\theta} \sum_{j \neq i}\left[D_j - \theta^T U\left(\frac{X_j - X_i}{h}\right)\right]^2 K\left(\frac{X_j - X_i}{h}\right).
\end{aligned}$$

We then construct an estimate of individual treatment effect as

$$\hat{\tau}_i = \left[\frac{Y_i D_i}{\hat{e}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{e}(X_i)}\right] \cdot 1\left\{\varepsilon_n \leq \hat{e}(X_i) \leq 1-\varepsilon_n\right\}, \quad 0 < \epsilon_n \leq O\left(n^{-a}\right), \, a > 0,$$

To ensure Condition 2.1 (e), we assume smoothness of the propensity score function $e(\cdot)$.

Assumption E.2. This assumption is the same as Assumption E.1 except that E.1 *(Smooth-m)* is replaced by
(Smooth-e) Smoothness of the Propensity Score: The propensity score $e(\cdot)$ belongs to a Hölder class of functions with degree $\beta_e \geq 1$ and constant $L_e < \infty$.

Again, Lemma E.4 below shows that $\mathcal{P}_e$ formed by the data generating processes satisfying Assumption E.2, Condition 2.1 (e) holds with $\phi_n = n^{-\frac{1}{2+d_x/\beta_e}}$ and (2.11) with $\tilde{\phi}_n = n^{\frac{1}{2+d_x/\beta_e}} (\log n)^{-\frac{1}{2+d_x/\beta_e}-2}$.

Corollary E.2. *Let $\mathcal{P}_e$ consist of data generating processes that satisfy Assumption E.2 (Smooth-e) and (PX). Let $\hat{e}(X_i)$ be the leave-one-out local polynomial estimator with degree $l = (\beta_e - 1)$, whose kernel satisfy Assumption E.1 (Ker).*

(i) Suppose Assumption 2.1 holds and a bandwidth satisfies $h \propto n^{-\frac{1}{2\beta_e + d_x}}$. Then, it holds

$$\sup_{P \in \mathcal{P}_e \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W_{\mathcal{G}}^* - W(\hat{G}_{e-hybrid})\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right).$$

(ii) Suppose Assumptions 2.1 and 2.2 hold with margin coefficient $\alpha \in (0,1]$, and a bandwidth satisfies $h \propto \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta_e + d_x}}$. Then, it holds

$$\begin{aligned} & \sup_{P \in \mathcal{P}_e \cap \mathcal{P}_{FB}(M,\kappa,\alpha,\eta)} E_{P^n}\left[W(G_{FB}^*) - W(\hat{G}_{e-hybrid})\right] \\ \leq \quad & O\left(n^{-\frac{1+\alpha}{2+d_x/\beta_e}} (\log n)^{\left(\frac{1}{2+d_x/\beta_e}+2\right)(1+\alpha)} \vee n^{-\frac{1+\alpha}{2+\alpha}} \log n\right). \end{aligned}$$

A comparison of Corollaries E.1 and E.2 shows that the rate upper bound of welfare loss differs between the m-hybrid EWM and the e-hybrid EWM approaches when the degree of Hölder smoothness of the regression equations β_m and that of the propensity score β_e are different. For instance, if the propensity score $e(\cdot)$ is smoother than the regression equations of outcome $m_1(\cdot)$ and $m_0(\cdot)$ in the sense of $\beta_e > \beta_m$ and the degree of local polynomial regressions is chosen accordingly, then the rate upper bound of the e-hybrid EWM rule converges faster than that of the m-hybrid EWM rule.

The rest of this section provides formal proofs for validity of Condition 2.1 (m) and (e) for the local polynomial estimators constructed above, when the class of data generating processes $\mathcal{P}_m$ or $\mathcal{P}_e$ is constrained by Assumptions E.1 or E.2. Lemma E.4 shown in Section C.3 proves the main claim. Appendix C.2 collects the preparatory lemmas to prove Lemma E.4.

E.2 Preparatory Lemmas

Let $\mu : \mathbb{R}^{d_x} \to \mathbb{R}$ be a generic notation for a regression equation onto a vector of covariates $X \in \mathbb{R}^{d_x}$. In case of m-hybrid EWM, $\mu(\cdot)$ corresponds to either of $m_1(\cdot)$ or $m_0(\cdot)$. In case of e-hybrid EWM, $\mu(\cdot)$ corresponds to propensity score $e(\cdot)$. We use n to denote the size of the entire sample indexed by $i = 1, \ldots, n$, and denote by $J_i \subset \{1, \ldots, n\}$ a subsample used to estimate $\mu(X_i)$ nonparametrically. Since we consider throughout the leave-one-out regression fits of $\mu(X_i)$, J_i does not include i-th observation. In case of m-hybrid EWM, J_i is either the leave-one-out treated sample $\{j \in \{1, \ldots, n\} : D_j = 1, j \neq i\}$ or the leave-one-out control sample $\{j \in \{1, \ldots, n\} : D_j = 0, j \neq i\}$ depending on $\mu(\cdot)$ corresponds to $m_1(\cdot)$ or $m_0(\cdot)$. Note that, in the m-hybrid case, J_i is random as it depends on a realization of $(D_1, \ldots, D_n)$. When the e-hybrid EWM is considered, J_i is non-stochastic and it is given by $J_i = \{1, \ldots, n\} \setminus \{i\}$. The size of J_i is denoted by n_{J_i}, which is equal to $n_1 - 1$ or $n_0 - 1$ in the m-hybrid case, and is equal to $n - 1$ in the e-hybrid case. With abuse of notations, we use Y_i, $i = 1, \ldots, n$, to denote dependent variable observations and use ξ_i to denote a regression residual, i.e., $Y_i = \mu(X_i) + \xi_i$, $E(\xi_i | X_i) = 0$, holds for all $i = 1, \ldots, n$. For e-hybrid rule, Y_i should be read as the treatment status indicator $D_i \in \{1, 0\}$.

We assume that $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$. Our generic notation for the leave-one-out local polynomial regression fir for $\mu(X_i)$ with degree $l = (\beta - 1)$ is

$$\begin{aligned}
\hat{\mu}_{-i}(X_i) &= U^T(0)\hat{\theta}(X_i) \cdot 1\{\lambda(X_i) \geq t_n\}, \qquad \text{(E.2)}\\
\hat{\theta}_{-i}(X_i) &= \arg\min_{\theta} \sum_{j \in J_i} \left[Y_j - \theta^T U\left(\frac{X_j - X_i}{h}\right)\right]^2 K\left(\frac{X_j - X_i}{h}\right),
\end{aligned}$$

where $U\left(\frac{X_j - X_i}{h}\right)$ is a regressor vector as defined above, $\lambda(X_i)$ is a smallest eigenvalue of $B_{-i}(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \in J_i} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$, and t_n is a sequence of trimming constant converging to zero, whose choice is discussed later. The standard least squares calculus shows

$$\hat{\theta}_{-i}(X_i) = B_{-i}(X_i)^{-1} \left(\frac{1}{nh^{d_x}} \sum_{j \in J_i} U\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_j - X_i}{h}\right)\right),$$

so that $\hat{\mu}(X_i)$ can be written as

$$\begin{aligned}
\hat{\mu}_{-i}(X_i) &= \left[\sum_{j \in J_i} Y_j \omega_j(X_i)\right] \cdot 1\{\lambda(X_i) \geq t_n\}, \qquad \text{(E.3)} \\
\text{where } \omega_j(X_i) &= \frac{1}{nh^{d_x}} U^T(0) \left[B_{-i}(X_i)\right]^{-1} U\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_j - X_i}{h}\right).
\end{aligned}$$

Lemma E.1. *Suppose Assumptions E.1 (PX) and (Ker).*

(i) Conditional on $(X_1, \ldots, X_n)$ such that $\lambda(X_i) > 0$,

$$\begin{aligned}
\max_{j \neq i} |\omega_j(X_i)| &\leq c_5 \frac{1}{nh^{d_x}\lambda(X_i)}, \\
\sum_{j \in J_i} |\omega_j(X_i)| &\leq \frac{c_5}{nh^{d_x}\lambda(X_i)} \sum_{j \in J_i} 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\},
\end{aligned}$$

where c_5 is a constant that depends only on β, d_x and $K_{\max}$.

(ii) For any multi-index s such that $|s| \leq (\beta - 1)$, $\sum_{j \in J_i} \left(\frac{X_j - X_i}{h}\right)^s \omega_j(X_i) = 0$.

(iii) Let $\tilde{\lambda}(x)$ be a smallest eigenvalue of $B(x) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j=1}^n U\left(\frac{X_j - x}{h}\right) U^T\left(\frac{X_j - x}{h}\right) K\left(\frac{X_j - x}{h}\right)$ there exist positive constants c_6 and c_7 that depend only on $\underline{c}$, r_0, $\underline{p}_X$, and $K(\cdot)$ such that

$$P^n\left(\left\{\tilde{\lambda}(x) \leq c_6\right\}\right) \leq 2\left[\dim U\right]^2 \exp\left(-c_7 nh^{d_x}\right)$$

holds for all x, P_X-almost surely, at every $n \geq 1$.

Proof. (i) Since $\|U(0)\| = 1$, it holds

$$\begin{aligned}
|\omega_j(X_i)| &\leq \frac{1}{nh^{d_x}} \left\| \left[B_{-i}(X_i)\right]^{-1} U\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_j - X_i}{h}\right) \right\| \\
&\leq \frac{K_{\max}}{nh^{d_x}\lambda(X_i)} \left\| U\left(\frac{X_j - X_i}{h}\right) 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\} \right\| \\
&\leq \frac{K_{\max} \dim(U)^{1/2}}{nh^{d_x}\lambda(X_i)} \\
&\equiv \frac{c_5}{nh^{d_x}\lambda(X_i)},
\end{aligned}$$

for every $1 \leq j \leq n$. Similarly,

$$\begin{aligned}
\sum_{j \in J_i} |\omega_j(X_i)| &\leq \frac{K_{\max}}{nh^{d_x}\lambda(X_i)} \sum_{j \in J_i} \left\| U\left(\frac{X_j - X_i}{h}\right) \right\| 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\} \\
&= \frac{c_5}{nh^{d_x}\lambda(X_i)} \sum_{j \in J_i} 1\left\{(X_j - X_i) \in [-h, h]^{d_x}\right\}.
\end{aligned}$$

(ii) This claim follows from the first order condition for θ in the least square minimization problem in (E.2).

(iii) This lemma is from Equation (6.3, pp. 626) in the proof of Theorem 3.2 in Audibert and Tsybakov (2007), where suitable choices of constant c_6 and c_7 are given in Equation (6.2, pp.625) in Audibert and Tsybakov (2007). □

The next lemma provides an exponential tail bound for the local polynomial estimators. The first statement is borrowed from Theorem 3.2 in Audibert and Tsybakov (2007), and the second statement is its immediate extension.

Lemma E.2. *(i) Suppose Assumption E.1 (PX) and (Ker) hold, and $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$. Assume J_i is non-stochastic with $n_{J_i} = n - 1$ (e-hybrid case). Then, there exist positive constants c_8, c_9, and c_{10} that depend only on β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$, such that, for any $0 < h < r_0/\underline{c}$, any $c_8 h^\beta < \delta$, and any $n \geq 2$,*

$$P^{n-1}\left(\left|\hat{\mu}_{-n}(x) - \mu(x)\right| > \delta\right) \leq c_9 \exp\left(-c_{10} n h^{d_x} \delta^2\right),$$

holds for almost all x with respect to P_X, where $P^{n-1}(\cdot)$ is the distribution of $\left\{(Y_i, X_i)_{i=1}^{n-1}\right\}$.

(ii) Suppose Assumptions 2.1 (SO), E.1 (PX), and (Ker) hold, and $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$. Assume J_i is stochastic (m-hybrid case) with $J_i = \{j \neq i : D_j = d\}$, $d \in \{1, 0\}$. There exist positive constants c_{11}, c_{12}, and c_{13} that depend only on κ, β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$, such that for any $0 < h < r_0/\underline{c}$, any $c_{11} h^\beta < \delta$, and any $n_{J_n} \geq 1$,

$$P^{n-1}\left(\left|\hat{\mu}_{-n}(x) - \mu(x)\right| > \delta | n_{J_n}\right) \leq c_{12} \exp\left(-c_{13} n_{J_n} h^{d_x} \delta^2\right)$$

holds for almost all x with respect to P_X, where $P^{n-1}(\cdot | n_{J_n})$ is the conditional distribution of $\left\{(Y_i, X_i)_{i=1}^{n-1}\right\}$ given $\sum_{j=1}^{n-1} 1\{D_j = d\}$.

Proof. (i) See Theorem 3.2 in Audibert and Tsybakov (2007).

(ii) Under Assumption 2.1 (SO), the conditional distribution of covariates X given $D = d$, $d \in \{1, 0\}$, has the support $\mathcal{X}$ same as the unconditional distribution P_X, and has bounded density on $\mathcal{X}$, since

$$\frac{\kappa}{1-\kappa}\frac{dP_X}{dx} < \frac{dP_{X|D=d}}{dx} < \frac{1-\kappa}{\kappa}\frac{dP_X}{dx}$$

holds for all $x \in \mathcal{X}$. Therefore, when P_X satisfies Assumption E.1 (PX), the conditional distributions $P_{X|D=d}$, $d \in \{1, 0\}$ also satisfy the support and density conditions analogous to Assumption

E.1 (PX). This implies that, even when we condition on $n_{J_n} = \sum_{j=1}^{n-1} 1\{D_j = d\} \geq 1$, the exponential inequality of (i) in the current lemma is applicable with different constant terms. □

The next lemma concerns an upper bound of the variance of the supremum of centered empirical processes indexed by a class of sets.

Lemma E.3. *Let $\mathcal{B}$ be a countable class of sets in $\mathcal{X}$, and let $\{P_{X,n}(B) : B \in \mathcal{B}\}$ be the empirical distribution based on iid observations, $(X_1, \ldots, X_n)$, $X_i \sim P_X$.*

$$Var\left(\sup_{B\in\mathcal{B}} \{P_{X,n}(B) - P_X(B)\}\right) \leq \frac{2}{n} E\left[\sup_{B\in\mathcal{B}} \{P_{X,n}(B) - P_X(B)\}\right] + \frac{1}{4n}.$$

Proof. In Theorem 11.10 of Boucheron et al. (2013), setting $X_{i,s}$ at the centered indicator function $1\{X_i \in B\} - P_X(B)$, and dividing the inequality of Theorem 11.10 of Boucheron et al. (2013) by n^2 lead to

$$\begin{aligned} Var\left(\sup_{B\in\mathcal{B}} \{P_{X,n}(B) - P_X(B)\}\right) &\leq \frac{2}{n} E\left[\sup_{B\in\mathcal{B}} \{P_{X,n}(B) - P_X(B)\}\right] \\ &\quad + \frac{1}{n} \sup_{B\in\mathcal{B}} \{P_X(B)[1 - P_X(B)]\} \\ &\leq \frac{2}{n} E\left[\sup_{B\in\mathcal{B}} \{P_{X,n}(B) - P_X(B)\}\right] + \frac{1}{4n}. \end{aligned}$$

□

E.3 Main Lemmas and Proofs of Corollaries E.1 and E.2

The next lemma yields Corollaries E.1 and E.2.

Lemma E.4. *Let $\mathcal{P}_\mu$ be a class of joint distributions of (Y, X) such that $\mu(\cdot)$ belongs to a Hölder class of functions with degree $\beta \geq 1$ and constant $0 < L < \infty$, and Assumption E.1 (PX) holds. Let $\hat{\mu}_{-i}(\cdot)$ be the leave-one-out local polynomial fit for $\mu(X_i)$ defined in (E.2), whose kernel function satisfies Assumption E.1 (Ker).*

(i) Then,

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n} \left|\hat{\mu}_{-i}(X_i) - \mu(X_i)\right|\right] \leq O(h^\beta) + O\left(\frac{1}{\sqrt{nh^{d_x}}}\right) \qquad \text{(E.4)}$$

holds. Hence, an optimal choice of bandwidth that leads to the fastest convergence rate of the uniform upper bound is $h \propto n^{-\frac{1}{2\beta+d_x}}$ *and the resulting uniform convergence rate is*

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta}}\right).$$

(ii) Let $t_n \propto (\log n)^{-1}$*. Then,*

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right)^2\right] \leq O\left(\frac{h^{2\beta}}{t_n^2}\right)+O\left(\frac{\log n}{nh^{d_x}t_n^2}\right) \tag{E.5}$$

holds. Hence, an optimal choice of bandwidth that leads to the fastest convergence rate of the uniform upper bound is $h \propto \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+d_x}}$ *and the resulting uniform convergence rate is*

$$\sup_{P\in\mathcal{P}_\mu} E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right)^2\right] \leq O\left((t_n)^{-2}\left(\frac{\log n}{n}\right)^{\frac{2}{2+d_x/\beta}}\right).$$

Proof. (i) First, consider the non-stochastic J_i case with $n_{J_i}=(n-1)$ (e-hybrid case). Since observations are iid (hence exchangeable) and the probability law of $\hat{\mu}_{-i}(\cdot)$ does not depend on X_i, it holds

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|\right] &= E_{P^n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right| \\ &= E_{P_X}\left[E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(X_n)-\mu\left(X_n\right)\right|\left|X_n\right.\right]\right] \\ &= \int_{\mathcal{X}} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|\right]dP_X(x) \\ &= \int_{\mathcal{X}}\left[\int_0^\infty P^{n-1}\left(\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|>\delta\right)d\delta\right]dP_X(x), \end{aligned} \tag{E.6}$$

where $E_{P^{n-1}}[\cdot]$ is the expectation with respect to the first $(n-1)$-observations of (Y_i, X_i). By Lemma E.2 (i), there exist positive constants c_8, c_9, and c_{10} that depend only on β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$ such that, for any $0<h<r_0/\underline{c}$, any $c_8h^\beta<\delta$, and any $n\geq 2$,

$$P^{n-1}\left(\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|>\delta\right) \leq c_9\exp\left(-c_{10}nh^{d_x}\delta^2\right) \tag{E.7}$$

holds for almost all x with respect to P_X. Hence,

$$\begin{aligned} \int_{\mathcal{X}}\left[\int_0^\infty P^{n-1}\left(\left|\hat{\mu}_{-n}(x)-\mu\left(x\right)\right|>\delta\right)d\delta\right]dP_X(x) &\leq c_8h^\beta + c_9\int_0^\infty \exp\left(-c_{10}nh^{d_x}\delta^2\right)d\delta \\ &= c_8h^\beta+\frac{c_{14}}{\sqrt{nh^{d_x}}} \qquad \text{(E.8)} \\ &= O(h^\beta)+O\left(\frac{1}{\sqrt{nh^{d_x}}}\right) \end{aligned}$$

where $c_{14} = c_9(2c_{10})^{-1/2} \int_0^\infty \left(\delta'\right)^{-1/2} \exp\left(-c_{10}\delta'\right) d\delta' < \infty$. Since the upper bound (E.8) does not depend upon $P \in \mathcal{P}_\mu$, this upper bound is uniform over $P \in \mathcal{P}_\mu$, so the conclusion holds.

Next, consider the stochastic J_i case with $n_{J_i} = \sum_{j \neq i} 1\{D_j = d\}$, where $d \in \{1, 0\}$. we can interpret n_{J_i} as a binomial random variable with parameters $(n-1)$ and π, where $\pi = P(D_i = 1)$ when $\mu(\cdot)$ corresponds to $m_1(\cdot)$ and $\pi = P(D_i = 0)$ when $\mu(\cdot)$ corresponds to $m_0(\cdot)$. In either case, $\kappa < \pi < 1 - \kappa$ by Assumption 2.1 (SO). Let $n \geq 1 + \frac{2}{\pi}$ and $\Omega_{\pi,n} \equiv \left\{\left|\frac{n_{J_n}}{n-1} - \pi\right| \leq \frac{1}{2}\pi\right\} = \left\{\frac{(n-1)\pi}{2} \leq n_{J_n} \leq \frac{3(n-1)\pi}{2}\right\}$. Consider

$$\begin{aligned} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| \cdot 1\{\Omega_{\pi,n}\}\right] &= \sum_{n_{J_n} \in \Omega_{\pi,n}} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| |n_{J_n}\right] P^{n-1}(n_{J_n}) \\ &\leq \max_{n_{J_n} \in \Omega_{\pi,n}} \left\{E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| |n_{J_n}\right]\right\} P^{n-1}(\Omega_{\pi,n}) \\ &\leq \max_{n_{J_n} \in \Omega_{\pi,n}} \left\{E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| |n_{J_n}\right]\right\}. \end{aligned}$$

Since $n_{J_n} \geq \frac{(n-1)\pi}{2} \geq 1$ on $\Omega_{\pi,n}$, Lemma E.2 (ii) implies

$$\begin{aligned} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| |n_{J_n}\right] &\leq \int_{\mathcal{X}} \left[\int_0^\infty P^{n-1}\left(\left|\hat{\mu}_{-n}(x) - \mu(x)\right| > \delta |n_{J_n}\right) d\delta\right] dP_X(x) \\ &\leq c_{11}h^\beta + \frac{c_{15}}{\sqrt{n_{J_n}h^{d_x}}}, \end{aligned}$$

where c_{11} and c_{15} are positive constants that depend only on κ, β, d_x, L, $\underline{c}$, r_0, $\underline{p}_X$, and $\bar{p}_X$. Since $n_{J_n} \geq \frac{(n-1)\pi}{2} \geq \frac{n\pi}{4}$ on $\Omega_{\pi,n}$ for $n \geq 2$, it holds

$$\max_{n_{J_n} \in \Omega_{\pi,n}} \left\{E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| |n_{J_n}\right]\right\} \leq c_{11}h^\beta + \frac{2c_{15}}{\sqrt{\pi n h^{d_x}}}.$$

Accordingly, combined with the Hoeffding's inequality $P^{n-1}\left(\Omega_{\pi,n}^c\right) \leq 2\exp\left(-\frac{\pi^2}{4}n\right)$, we obtain

$$\begin{aligned} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right|\right] &\leq E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right| \cdot 1\{\Omega_{\pi,n}\}\right] + MP^{n-1}\left(\Omega_{\pi,n}^c\right) \\ &\leq c_{11}h^\beta + \frac{2c_{15}}{\sqrt{\pi n h^{d_x}}} + 2M\exp\left(-\frac{\pi^2}{4}n\right). \end{aligned}$$

The third term in the right hand side converges faster than the second term, so we have shown

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left|\hat{\mu}_{-i}(X_i) - \mu(X_i)\right|\right] &= \int_{\mathcal{X}} E_{P^{n-1}}\left[\left|\hat{\mu}_{-n}(x) - \mu(x)\right|\right] dP_X(x) \\ &\leq O(h^\beta) + O\left(\frac{1}{\sqrt{nh^{d_x}}}\right) \end{aligned}$$

holds for the stochastic J_i case as well.

(ii) Let $\Omega_{\lambda,n}$ be an event defined by $\{\lambda\left(X_i\right) \geq t_n,\ \forall i = 1, \ldots, n\}$. On $\Omega_{\lambda,n}$, (E.3) implies

$$\begin{aligned}
\left|\hat{\mu}_{-i}(X_i) - \mu\left(X_i\right)\right|^2 &\leq \left|\sum_{j\in J_i} Y_j \omega_j\left(X_i\right) - \mu\left(X_i\right)\right|^2 \\
&= \left|\sum_{j\in J_i}\left(\mu\left(X_j\right) - \mu\left(X_i\right)\right)\omega_j\left(X_i\right) + \sum_{j\in J_i}\xi_j\omega_j\left(X_i\right)\right|^2 \\
&\leq 2\left|\sum_{j\in J_i}\left(\mu\left(X_j\right) - \mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2 + 2\left|\sum_{j\in J_i}\xi_j\omega_j\left(X_i\right)\right|^2, \qquad \text{(E.9)}
\end{aligned}$$

where the second line follows from $Y_j = \mu\left(X_j\right) + \xi_j$ and $\sum_{j\neq i}\omega_j\left(X_i\right) = 0$ as implied by Lemma E.1 (ii). Since $\mu\left(\cdot\right)$ is assumed to belong to the Hölder class, Lemma E.1 (ii) and Assumption E.1 (Ker) imply

$$\begin{aligned}
&\left|\sum\nolimits_{j\in J_i}\left(\mu\left(X_j\right) - \mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2 = \left|\sum\nolimits_{j\in J_i}\left\|X_j - X_i\right\|^\beta \omega_j\left(X_i\right)\right|^2 \\
&= \left|\sum\nolimits_{j\in J_i}\left\|X_j - X_i\right\|^\beta \omega_j\left(X_i\right)\cdot 1\left\{\left(X_j - X_i\right) \in [-h,h]^{d_x}\right\}\right|^2 \\
&\leq d_x^\beta h^{2\beta}\left|\sum\nolimits_{j\in J_i}\left|\omega_j\left(X_i\right)\right|\right|^2 \\
&\leq d_x^\beta h^{2\beta}\left(\frac{c_5}{\lambda(X_i)}\right)^2\left(\frac{1}{nh^{d_x}}\sum\nolimits_{j\in J_i} 1\left\{\left(X_j - X_i\right)\in[-h,h]^{d_x}\right\}\right)^2 \\
&\leq c_{16}\frac{h^{2\beta}}{t_n^2}\left(\frac{1}{nh^{d_x}}\sum\nolimits_{j\in J_i} 1\left\{\left(X_j - X_i\right)\in[-h,h]^{d_x}\right\}\right)^2,
\end{aligned}$$

where $c_{16} = c_5^2 d_x^\beta$. Under Assumption E.1 (PX) and being conditional on $\Omega_{\lambda,n}$,

$$\begin{aligned}
&\max_{1\leq i\leq n}\left|\sum\nolimits_{j\in J_i}\left(\mu\left(X_j\right) - \mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2 \leq c_{16}\frac{h^{2\beta}}{t_n^2}\left[\frac{1}{h^{d_x}}\sup_{B\in\mathcal{B}_h} P_{X,n}(B)\right]^2 \\
&\leq c_{16}\frac{h^{2\beta}}{t_n^2}\left[\frac{1}{h^{d_x}}\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B) - P_X\left(B\right)\right) + \sup_{B\in\mathcal{B}_h} P_X\left(B\right)\right)\right]^2 \\
&\leq c_{16}\frac{h^{2\beta}}{t_n^2}\left[\frac{1}{h^{d_x}}\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B) - P_X\left(B\right)\right) + 2^{d_x}\cdot\bar{p}_X\right]^2 \\
&\leq c_{16}\frac{h^{2\beta}}{t_n^2}\left\{\frac{2}{h^{2d_x}}\left[\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B) - P_X\left(B\right)\right)\right]^2 + 2^{2d_x+1}\cdot\bar{p}_X^2\right\},
\end{aligned}$$

where $\mathcal{B}_h$ is the class of hypercubes in $\mathbb{R}^{d_x}$, $\mathcal{B}_h \equiv \left\{\prod\nolimits_{k=1}^{d_x}\left[x_k - h, x_k + h\right] : \left(x_1, \ldots, x_{d_x}\right) \in \mathcal{X}\right\}$, and the last inequality follows since $(a+b)^2 \leq 2a^2 + 2b^2$. Accordingly,

$$\begin{aligned} & E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum\nolimits_{j\neq i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2 \cdot 1\left\{\Omega_{\lambda,n}\right\}\right] \\ \leq\ & c_{17}\frac{h^{2\beta}}{t_n^2} + 2c_{16}\frac{h^{2\beta}}{t_n^2}\frac{1}{h^{2d_x}}E_{P^n}\left\{\left[\sup\nolimits_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right]^2\right\} \\ \leq\ & c_{17}\frac{h^{2\beta}}{t_n^2} + 4c_{16}\frac{h^{2\beta}}{t_n^2}\frac{1}{h^{2d_x}}\left\{\begin{array}{c} Var\left(\sup\nolimits_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right) \\ +\left[E_{P^n}\left(\sup\nolimits_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right)\right]^2 \end{array}\right\}, \end{aligned}$$

where $c_{17} = 2^{2d_x+1}c_{16}\bar{p}_X^2$. In order to bound the variance and the squared mean terms in the curly brackets, we apply Lemma E.3 and Lemma A.5 with $\bar{F} = 1$ and $\delta = \bar{p}_X\left(2h\right)^{d_x/2}$. Let $v_{\mathcal{B}_h} < \infty$ be the VC-dimension of $\mathcal{B}_h$ that depends only on d_x. For all n satisfying $nh^{d_x} \geq \frac{C_1 v_{\mathcal{B}_h}}{2^{d_x}\bar{p}_X^2}$, we have

$$\begin{aligned} Var\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right) &\leq \frac{2}{n}E_{P^n}\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right)+\frac{1}{4n} \\ &\leq 2^{\frac{d_x}{2}+1}C_2\bar{p}_X\frac{\sqrt{v_{\mathcal{B}_h}h^{d_x}}}{n^{3/2}}+\frac{1}{4n} \quad \text{and} \\ \left[E_{P^n}\left(\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)\right)\right]^2 &\leq 2^{d_x}C_2^2\bar{p}_X^2\frac{v_{\mathcal{B}_h}h^{d_x}}{n}. \end{aligned}$$

As a result, there exist positive constants c_{18}, and c_{19} that depend only on β, d_x, and $\bar{p}_X$, such that

$$E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum_{j\neq i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2 \cdot 1\left\{\Omega_{\lambda,n}\right\}\right] \leq c_{17}\frac{h^{2\beta}}{t_n^2}+c_{18}\frac{h^{2\beta}}{t_n^2\left(nh^{d_x}\right)}+c_{19}\frac{h^{2\beta}}{t_n^2\left(nh^{d_x}\right)^{3/2}}$$

holds for all n satisfying $nh^{d_x} \geq \frac{C_1 v_{\mathcal{B}_h}}{2^{d_x}\bar{p}_X^2}$. Since $nh^{d_x} \to \infty$ by the assumption, focusing on the leading term yields

$$\limsup_{n\to\infty}\sup_{P\in P_\mu} E_{P^n}\left[2\max_{1\leq i\leq n}\left|\sum_{j\in J_i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2 \cdot 1\left\{\Omega_{\lambda,n}\right\}\right] \leq O\left(\frac{h^{2\beta}}{t_n^2}\right). \qquad \text{(E.10)}$$

In order to bound the second term in the right hand side of (E.9), note first that

$$\begin{aligned} \left|\sum_{j\in J_i}\xi_j\omega_j\left(X_i\right)\right|^2 &\leq \frac{1}{nh^{d_x}\lambda^2(X_i)}\left\|\frac{1}{\sqrt{nh^{d_x}}}\sum_{j\in J_i}\xi_j U\left(\frac{X_j-X_i}{h}\right)K\left(\frac{X_j-X_i}{h}\right)\right\|^2 \\ &\leq \frac{K_{\max}^2}{nh^{d_x}t_n^2}\max_{1\leq k\leq \dim(U)}\eta_{ik}^2 \end{aligned}$$

holds conditional on $\Omega_{\lambda,n}$, where η_{ik}, $1 \leq k \leq \dim\left(U\right)$, is the k-th entry of vector

$$\frac{1}{\sqrt{nh^{d_x}}}\sum_{j\in J_i}\xi_j U\left(\frac{X_j-X_i}{h}\right)1\left\{\left(X_j-X_i\right)\in\left[-h,h\right]^{d_x}\right\}.$$

Therefore,

$$E_{P^n}\left[\max_{1\le i\le n}\left|\sum_{j\in J_i}\xi_j\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]\le\frac{K_{\max}^2}{nh^{d_x}t_n^2}E_{P^n}\left[\max_{1\le i\le n,1\le k\le\dim(U)}\eta_{ik}^2\right]. \tag{E.11}$$

Conditional on $(X_1,\dots,X_n)$, η_{ik} has mean zero and every summand in η_{ik} lies in the interval, $\left[-\frac{M}{\sqrt{nh^{d_x}}}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\},\frac{M}{\sqrt{nh^{d_x}}}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}\right]$. The Hoeffding's inequality then implies that, for every $1\le i\le n$ and $1\le k\le\dim(U)$, it holds

$$\begin{aligned}
& P^n\left(|\eta_{ik}|\ge t|X_1,\dots,X_n\right)\\
\le\ & 2\exp\left(-\frac{t^2}{\frac{2M^2}{nh^{d_x}}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right)\\
\le\ & 2\exp\left(-\frac{t^2}{\frac{2M^2}{nh^{d_x}}\max_{1\le i\le n}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right),\ \ \forall t>0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& E_{P^n}\left[\exp\left(\frac{\eta_{ik}^2}{\frac{2M^2}{nh^{d_x}}\max_{1\le i\le n}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right)|X_1,\dots,X_n\right]\\
=\ & 1+\int_1^\infty P^n\left(\exp\left(\frac{\eta_{ik}^2}{\frac{2M^2}{nh^{d_x}}\max\limits_{1\le i\le n}\sum\limits_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}}\right)\ge t'|X_1,\dots,X_n\right)dt'\\
=\ & 1+\int_1^\infty P^n\left(|\eta_{ik}|\ge\sqrt{\frac{2M^2}{nh^{d_x}}\max_{1\le i\le n}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}\log t'}|X_1,\dots,X_n\right)dt'\\
\le\ & 1+2\int_1^\infty\exp\left(-2\log t'\right)dt'\\
=\ & 1+2\int_1^\infty\left(t'\right)^{-2}dt'\\
=\ & 3
\end{aligned}$$

for all $1\le i\le n$ and $1\le k\le\dim(U)$. We can therefore apply Lemma 1.6 of Tsybakov (2009) to bound $E_{P^n}\left[\max_{i,k}\eta_{ik}^2|X_1,\ldots,X_n\right]$,

$$\begin{aligned}
& E_{P^n}\left[\max_{1\leq i\leq n,1\leq k\leq \dim(U)}\eta_{ik}^2|X_1,\ldots,X_n\right] \\
\leq\; & 2M^2\max_{1\leq i\leq n}\left[\frac{1}{nh^{d_x}}\sum_{j\in J_i}1\left\{(X_j-X_i)\in[-h,h]^{d_x}\right\}\right]\log\left(3\dim\left(U\right)n\right) \\
\leq\; & 2M^2\left[\frac{1}{h^{d_x}}\sup_{B\in\mathcal{B}_h}\left(P_{X,n}(B)-P_X\left(B\right)\right)+2^{d_x}\bar{p}_X\right]\log\left(3\dim\left(U\right)n\right).
\end{aligned}$$

By applying Lemma A.5 with $\bar{F}=1$ and $\delta=\bar{p}_X\left(2h\right)^{d_x/2}$, the unconditional expectation of $\max_{i,k}\eta_{ik}^2$ can be bounded as

$$E_{P^n}\left[\max_{1\leq i\leq n,1\leq k\leq \dim(U)}\eta_{ik}^2\right]\leq 2M^2\left[C_2 2^{d_x/2}\bar{p}_X\sqrt{\frac{v_{\mathcal{B}_h}}{nh^{d_x}}}+2^{d_x}\bar{p}_X\right]\log\left(3\dim\left(U\right)n\right) \tag{E.12}$$

for all n such that $nh^{d_x}\geq\frac{C_1 v_{\mathcal{B}_h}}{2^{d_x}\bar{p}_X^2}$. Plugging (E.12) back into (E.11) and focusing on the leading term give

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_\mu}E_{P^n}\left[\max_{0\leq i\leq n}\left|\sum\nolimits_{j\neq i}\xi_j\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]\leq O\left(\frac{\log n}{nh^{d_x}t_n^2}\right). \tag{E.13}$$

Combining (E.9), (E.10), and (E.13), we obtain

$$\begin{aligned}
& E_{P^n}\left[\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|^2\right] \\
\leq\; & E_{P^n}\left[\max_{1\leq i\leq n}\left|\hat{\mu}_{-i}(X_i)-\mu\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]+M^2P^n\left(\Omega_{\lambda,n}^c\right) \\
\leq\; & 2E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum\nolimits_{j\neq i}\left(\mu\left(X_j\right)-\mu\left(X_i\right)\right)\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right] \\
& +2E_{P^n}\left[\max_{1\leq i\leq n}\left|\sum\nolimits_{j\neq i}\xi_j\omega_j\left(X_i\right)\right|^2\cdot 1\left\{\Omega_{\lambda,n}\right\}\right]+M^2P^n\left(\Omega_{\lambda,n}^c\right), \\
=\; & O\left(\frac{h^{2\beta}}{t_n^2}\right)+O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)+M^2P^n\left(\Omega_{\lambda,n}^c\right),
\end{aligned}$$

so the desired conclusion is proven if $P^n\left(\Omega_{\lambda,n}^c\right)$ is shown to converge faster than the $O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)$ term.

To find the convergence rate of $P^n\left(\Omega_{\lambda,n}^c\right)$, consider first the case of non-stochastic J_i. By applying Lemma E.1 (iii) with the sample size set at $(n-1)$, we have

$$\begin{aligned} P^n\left(\{\lambda\left(X_i\right) \leq c_6, \text{ for some } 1 \leq i \leq n\}\right) &= nP^n\left(\{\lambda\left(X_n\right) \leq c_6\}\right) \\ &= n\int P^n\left(\lambda(X_n) \leq c_6 | X_n\right) dP_X \\ &= n\int P^{n-1}\left(\lambda(x) \leq c_6\right) dP_X(x) \qquad\qquad \text{(E.14)} \\ &\leq 2n\left[\dim U\right]^2 \exp\left(-\frac{c_7}{2} nh^{d_x}\right). \end{aligned}$$

For the case of stochastic J_i, by viewing n_{J_i} as a binomial random variable with parameters $(n-1)$ and π with $\kappa < \pi < 1-\kappa$, and recalling that, when P_X satisfies Assumption E.1 (PX), the conditional distributions $P_{X|D=d}$, $d \in \{1,0\}$ also satisfy the support and density conditions stated in Assumption E.1 (PX), we can apply the exponential inequality shown in Lemma E.1 (iii) to bound $P^{n-1}\left(\lambda(x) \leq c_6 | n_{J_n}\right)$. Hence, with $\Omega_{\pi,n} \equiv \left\{\left|\frac{n_{J_n}}{n-1} - \pi\right| \leq \frac{1}{2}\pi\right\} = \left\{\frac{(n-1)\pi}{2} \leq n_{J_n} \leq \frac{3(n-1)\pi}{2}\right\}$ used above, we have

$$\begin{aligned} P^{n-1}\left(\lambda(x) \leq c_6\right) &\leq P^{n-1}\left(\{\lambda(x) \leq c_6\} \cap \Omega_{\pi,n}\right) + P^{n-1}\left(\Omega_{\pi,n}^c\right) \\ &\leq \max_{n_{J_n} \in \Omega_{\pi,n}} P^{n-1}\left(\lambda(x) \leq c_6 | n_{J_n}\right) + P^{n-1}\left(\Omega_{\pi,n}^c\right). \\ &\leq 2\left[\dim U\right]^2 \exp\left(-\frac{c_7\pi}{4} nh^{d_x}\right) + 2\exp\left(-\frac{\pi^2}{4}n\right), \end{aligned}$$

Plugging this upper bound into (E.14) and focusing on the leading term leads to

$$P^n\left(\{\lambda\left(X_i\right) \leq c_6, \text{ for some } 1 \leq i \leq n\}\right) \leq O\left(n\exp\left(-c_7\frac{\pi}{4}nh^{d_x}\right)\right).$$

Hence, in either of the non-stochastic or the stochastic J_i case, since $t_n \leq c_6$ holds for all large n and the obtained upper bounds are uniform over $P \in \mathcal{P}_\mu$, we conclude

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{P}_\mu} E_{P^n}\left[\max_{1 \leq i \leq n}\left|\hat{\mu}_{-i}(X_i) - \mu\left(X_i\right)\right|^2\right] \leq O\left(\frac{h^{2\beta}}{t_n^2}\right) + O\left(\frac{\log n}{nh^{d_x}t_n^2}\right) + O\left(n\exp(-nh^{d_x})\right).$$

Since $t_n = (\log n)^{-1}$ by assumption, $O(n\exp\left(-nh^{d_x}\right))$ converges faster than $O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)$, the leading terms are given by the first two terms, $O\left(\frac{h^{2\beta}}{t_n^2}\right) + O\left(\frac{\log n}{nh^{d_x}t_n^2}\right)$. □

Proof of Corollary E.1. By noting the following inequalities,

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{\tau}^m(X_i)-\tau\left(X_i\right)\right|\right] &\leq E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{m}_1(X_i)-m_1\left(X_i\right)\right|\right] \\ &\quad +E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{m}_0(X_i)-m_0\left(X_i\right)\right|\right] \\ E_{P^n}\left[\max_{1\leq i\leq n}\left(\hat{\tau}^m(X_i)-\tau\left(X_i\right)\right)^2\right] &\leq 2E_{P^n}\left[\max_{1\leq i\leq n}\left(\hat{m}_1(X_i)-m_1\left(X_i\right)\right)^2\right] \\ &\quad +2E_{P^n}\left[\max_{1\leq i\leq n}\left(\hat{m}_0(X_i)-m_0\left(X_i\right)\right)^2\right], \end{aligned}$$

we obtain the current corollary by applying Lemma E.4. The resulting uniform convergence rate is given by $\psi_n = n^{\frac{1}{2+d_x/\beta_m}}$. When the assumption (2.10) in Theorem 2.6 is concerned, the corresponding rate is given by $\tilde{\psi}_n = \left[\left(\frac{\log n}{n}\right)^{\frac{1}{2+d_x/\beta_m}}(\log n)^2\right]^{-1}$. □

Proof of Corollary E.2. (i) Assume that n is large enough so that $\varepsilon_n \leq \kappa/2$ holds. Given $\hat{e}\left(X_i\right) \in [\varepsilon_n, 1-\varepsilon_n]$, $\hat{\tau}_i^e - \tau_i$ can be expressed as

$$\hat{\tau}_i^e - \tau_i = \frac{Y_i D_i}{e(X_i)}\left[\frac{e(X_i)-\hat{e}(X_i)}{\hat{e}(X_i)}\right] + \frac{Y_i\left(1-D_i\right)}{1-e(X_i)}\left[\frac{e(X_i)-\hat{e}(X_i)}{1-\hat{e}\left(X_i\right)}\right],$$

so

$$\left|\hat{\tau}_i^e - \tau_i\right| \leq \frac{M}{\kappa}\cdot\frac{1}{\hat{e}\left(X_i\right)\left(1-\hat{e}\left(X_i\right)\right)}\cdot\left|\hat{e}(X_i)-e(X_i)\right|$$

holds. On the other hand, when $\hat{e}\left(X_i\right) \notin [\varepsilon_n, 1-\varepsilon_n]$, $\hat{\tau}_i^e = 0$ and $|\tau_i| \leq \frac{M}{\kappa}$ imply $|\hat{\tau}_i^e - \tau_i| \leq \frac{M}{\kappa}$. Hence, the following bounds are valid,

$$\left|\hat{\tau}_i^e - \tau_i\right| \leq \begin{cases} \frac{M}{\kappa}\cdot\frac{4}{\kappa(2-\kappa)}\cdot\left|\hat{e}(X_i)-e(X_i)\right| & \text{if } \hat{e}\left(X_i\right) \in \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right], \\ \frac{M}{\kappa}\cdot\frac{1}{\varepsilon_n(1-\varepsilon_n)} & \text{if } \hat{e}\left(X_i\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]. \end{cases} \tag{E.15}$$

Hence,

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\hat{\tau}_i^e - \tau_i\right|\right] &= E_{P^n}\left[\left|\hat{\tau}_n^e - \tau_n\right|\right] \\ &\leq \frac{M}{\kappa}\cdot\frac{4}{\kappa\left(2-\kappa\right)}\cdot E_{P^n}\left[\left|\hat{e}(X_n)-e(X_n)\right|\right] \\ &\quad +\frac{M}{\kappa}\cdot\frac{1}{\varepsilon_n\left(1-\varepsilon_n\right)}\cdot P^n\left(\hat{e}\left(X_n\right) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right). \end{aligned}$$

By Lemma E.4 (i), $\sup_{P\in\mathcal{P}_e} E_{P^n}\left[|\hat{e}(X_n)-e(X_n)|\right] \le O(n^{-\frac{1}{2+d_x/\beta_e}})$, so the conclusion follows if $P^n\left(\ \hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right)$ is shown to converge faster than $O(n^{-\frac{1}{2+d_x/\beta_e}})$. To see this claim is true, note that

$$\begin{aligned} P^n\left(\ \hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right) &= \int_{\mathcal{X}} P^{n-1}\left(\ \hat{e}\left(x\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right)dP_X\left(x\right) \\ &\le \int_{\mathcal{X}} P^{n-1}\left(\ \left|\hat{e}\left(x\right)-e(x)\right|\ge\frac{\kappa}{2}\right)dP_X\left(x\right) \\ &\le c_9\exp\left(-\frac{c_{10}\kappa^2}{4}nh^{d_x}\right) \end{aligned}$$

holds for all n satisfying $c_8h^{\beta}<\kappa/2$, where the c_8, c_9, and c_{10} are the constants defined in Lemma B.2 (i). Since ε_n is assumed to converge at a polynomial rate, $\frac{1}{\varepsilon_n(1-\varepsilon_n)}P^n\left(\hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right)$ converges faster than $O(n^{-\frac{1}{2+d_x/\beta_e}})$.

(ii) By (E.15), we have

$$\begin{aligned} E_{P^n}\left[\max_{1\le i\le n}\left|\hat{\tau}_i^e-\tau_i\right|^2\right] &\le \left(\frac{4M}{\kappa^2\left(2-\kappa\right)}\right)^2 E_{P^n}\left[\max_{1\le i\le n}\left|\hat{e}(X_i)-e(X_i)\right|^2\right] \\ &\quad+\left(\frac{M}{\kappa\varepsilon_n\left(1-\varepsilon_n\right)}\right)^2 P^n\left(\ \hat{e}\left(X_i\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\text{ for some }1\le i\le n\right). \end{aligned} \tag{E.16}$$

By Lemma E.4 (ii), the first term in (E.16) converges at rate $O\left(n^{-\frac{2}{2+d_x/\beta}}\left(\log n\right)^{\frac{2}{2+d_x/\beta}+2}\right)$. To find the convergence rate of the second term in (E.16), consider

$$\begin{aligned} & P^n\left(\ \hat{e}\left(X_i\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\text{ for some }1\le i\le n\right) \\ \le\ & nP^n\left(\hat{e}\left(X_n\right)\notin\left[\frac{\kappa}{2},1-\frac{\kappa}{2}\right]\right) \\ \le\ & c_9n\exp\left(-\frac{c_{10}\kappa^2}{4}nh^{d_x}\right), \end{aligned}$$

where the last line follows from Lemma B.2 (i). Since ε_n converges at polynomial rate, we conclude the second term in (E.16) converges faster than the first term. □

References

Audibert, J.-Y. and A. B. Tsybakov (2007): "Fast Learning Rates for Plug-in Classifiers," *The Annals of Statistics*, 35, 608–633.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities, A Nonasymptotic Theory of Independence*, Oxford University Press.

Tsybakov, A. B. (2009): *Introduction to Nonparametric Estimation*, Springer.