SUPPLEMENT TO "WHO SHOULD BE TREATED? EMPIRICAL WELFARE MAXIMIZATION METHODS FOR TREATMENT CHOICE"

(*Econometrica*, Vol. 86, No. 2, March 2018, 591–616)

TORU KITAGAWA
Cemmap and Department of Economics, University College London

ALEKSEY TETENOV
Department of Economics, University of Bristol

APPENDIX 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

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

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 $\mathbb{R}$*

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

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 $\mathcal{Z} \times \mathbb{R}$, $A = \{(z_1, t_1), \ldots, (z_{v+1}, t_{v+1})\}$. Denote the collection of subgraphs of $\mathcal{F}$ by $\mathrm{SG}(\mathcal{F}) \equiv \{\mathrm{SG}(f_G), G \in \mathcal{G}\}$. We want to show that $\mathrm{SG}(\mathcal{F})$ does not shatter A.

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

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 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 $S \in \mathrm{SG}(\mathcal{F})$ that picks out the set (possibly empty)

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

Toru Kitagawa: t.kitagawa@ucl.ac.uk
Aleksey Tetenov: a.tetenov@bristol.ac.uk

 https://doi.org/10.3982/ECTA13288

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. *Q.E.D.*

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 i.i.d. 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 $\|f\|_{L_2(P)} = [\int_{\mathcal{Z}} f^2 \, dP]^{1/2}$, 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,*

$$\text{(i)} \quad E\left(\max_{i \leq n} Y_i\right) \leq \lambda\sqrt{2 \ln n},$$

$$\text{(ii)} \quad E\left(\max_{i \leq n} |Y_i|\right) \leq \lambda\sqrt{2 \ln(2n)}.$$

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, that is, 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 i.i.d. 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,

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

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

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

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 $B_0 = \{\mathbf{f}^{(0)}\}$ and construct $\{B_k : k = 1, \ldots, \bar{K}\}$ 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 $\{\mathbf{f}^{(k)} \in B_k : k = 1, \ldots, \bar{K}\}$ be a random sequence such that $\mathbf{f}^{(k)} \in \arg\min_{\mathbf{f}\in\mathbf{B}_k} \rho(\mathbf{f}, \mathbf{f}^*)$. Since B_k is a cover with radius $2^{-k}\bar{F}$, $\rho(\mathbf{f}^{(k)}, \mathbf{f}^*) \le 2^{-k}\bar{F}$ holds. In addition, we have

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

By a telescope sum,

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

We hence obtain

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

We apply Lemma A.2 to obtain

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

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

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

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

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

where the last line follows from the fact that $N(\varepsilon\bar{F},\mathbf{F},\rho)$ 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,

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

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

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

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

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

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

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 $B_0 = \{\mathbf{f}^{(0)}\}$ and construct $\{B_k : k = 1, \ldots, \bar{K}\}$ 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 $\{\mathbf{f}^{(k)} \in B_k : k = 1, \ldots, \bar{K}\}$ to be a random sequence such that $\mathbf{f}^{(k)} \in \arg\min_{\mathbf{f}\in\mathbf{B}_k} \rho(\mathbf{f}, \mathbf{f}^*)$. 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

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

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

$$\begin{aligned} E_{P^n}\Big[\sup_{f\in\mathcal{F}}\big(E_n(f)-E_P(f)\big)\Big] &\le C_1\sqrt{\frac{v}{n}}E_{P^n}(\delta_n) = C_1\sqrt{\frac{v}{n}}E_{P^n}\Big(\Big[\sup_{f\in\mathcal{F}}E_n\big(f^2\big)\Big]^{1/2}\Big) \\ &\le C_1\sqrt{\frac{v}{n}}\Big[E_{P^n}\Big(\sup_{f\in\mathcal{F}}E_n\big(f^2\big)\Big)\Big]^{1/2}. \end{aligned} \tag{A.8}$$

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

$$\begin{aligned} E_n\big(f^2\big) &= E_n\big(f^2 - E_P\big(f^2\big)\big) + E_P\big(f^2\big) \\ &= E_n\big[\big(f-\|f\|_{L_2(P)}\big)\big(f+\|f\|_{L_2(P)}\big)\big] + \|f\|_{L_2(P)}^2 \\ &\le 2\bar{F}E_n\big[f-\|f\|_{L_2(P)}\big] + \|f\|_{L_2(P)}^2 \\ &\le 2\bar{F}E_n\big[f-E_P(f)\big] + \|f\|_{L_2(P)}^2 \\ &\big(\because \|f\|_{L_2(P)} \ge E_P(f) \text{ by the Cauchy–Schwarz inequality}\big). \end{aligned}$$

Combining this inequality with (A.8) yields

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

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

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

For $\frac{v}{n} \le \frac{\delta^2}{\bar{F}^2C_1^2}$, that is, $n \ge \frac{C_1^2\bar{F}^2v}{\delta^2}$, the upper bound can be further bounded by $(1+\sqrt{2})\times C_1\delta\sqrt{\frac{v}{n}}$, so the conclusion of the lemma follows with $C_2 = (1+\sqrt{2})C_1$. *Q.E.D.*

A.2. *Proofs of Theorems 2.1 and 2.2*

PROOF OF THEOREM 2.1: Define

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

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

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

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

$$W_{\mathcal{G}}^* - W(\hat{G}_{\mathrm{EWM}}) \le 2\sup_{f\in\mathcal{F}}\big|E_n(f)-E_P(f)\big|. \tag{A.9}$$

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

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

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

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

We consider a suitable subclass $\mathcal{P}^* \subset \mathcal{P}(1, \kappa)$, 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 $x_1, \ldots, x_v \in \mathcal{X}$ be v points that are shattered by $\mathcal{G}$. We constrain P_X (the marginal distribution of X) to being supported only on $(x_1, \ldots, x_v)$. 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 $\mathbf{b} = (b_1, \ldots, b_v) \in \{0, 1\}^v$ be a bit vector used to index a member of $\mathcal{P}^*$, that is, $\mathcal{P}^*$ consists of a finite number of DGPs. For each $j = 1, \ldots, v$, and depending on $\mathbf{b}$, construct the following conditional distribution of Y_1 given $X = x_j$: if $b_j = 1$,

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

and, if $b_j = 0$,

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

where $\gamma \in [0, \frac{1}{2}]$ 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 subclass 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

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

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 jth element is $\hat{b}_j = 1\{x_j \in \hat{G}\}$. Consider $\pi(\mathbf{b})$ a prior distribution for $\mathbf{b}$ such that $b_1, \ldots, b_v$ are i.i.d. and $b_1 \sim \text{Ber}(1/2)$. The welfare loss satisfies the following inequalities:

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

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

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

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 $\pi(\mathbf{b})$. For each $j = 1, \ldots, (v)$, let

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

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

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

Hence,

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

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

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

where $E_{\tilde{P}^n}(\cdot)$ 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

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

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

$$E\left|B\left(k,\frac{1}{2}\right)-\frac{k}{2}\right| \le \sqrt{E\left(B\left(k,\frac{1}{2}\right)-\frac{k}{2}\right)^2} \quad (\because \text{Cauchy–Schwarz inequality})$$
$$= \sqrt{\frac{k}{4}},$$

we obtain

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

Hence, the Bayes risk is bounded from below by

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

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

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

The condition $1-2\gamma \ge \frac{1}{2}$ is equivalent to $n \ge 16v$. Multiplying M to this lower bound completes the proof. *Q.E.D.*

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 $\sup_{f\in\mathcal{F}} E_P(f^2) \le \delta^2$ and $\sup_{f\in\mathcal{F}}\|f\|_\infty \le \bar{F}$ 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,*

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

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

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

and its sample analogue

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

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

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

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

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

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

On the other hand, with Assumption 2.2 (MA) imposed, $P_X(G^*_{\mathrm{FB}} \triangle G)$ can be bounded from above by a function of $d(G^*_{\mathrm{FB}}, G)$, 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*

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

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

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

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

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

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

This, in turn, implies

$$P_X\big(G^*_{\mathrm{FB}} \triangle G\big) \le \left(\frac{M}{\kappa\eta\alpha}\right)^{\frac{\alpha}{1+\alpha}} (1 + \alpha) d\big(G^*_{\mathrm{FB}}, G\big)^{\frac{\alpha}{1+\alpha}}. \qquad \textit{Q.E.D.}$$

PROOF OF THEOREM 2.3: Let $a = \sqrt{kt}\varepsilon_n$ with $k \ge 1$, $t \ge 1$, and $\varepsilon_n > 0$, where $t \ge 1$ 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

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

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

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

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 $d(G^*_{\mathrm{FB}}, G) = -E_P(h(Z_i; G))$. Since $d(G^*_{\mathrm{FB}}, G) \geq 0$ for all $G \in \mathcal{G}$, it holds that $-E_P(h) \geq 0$ for all $h \in \mathcal{H}$.

Since we have

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

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

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

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

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

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

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

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

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

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

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

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

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 $\mathcal{H}_0 = \{h\in\mathcal{H} : -E_P(h) \le a^2\}$ and $\mathcal{H}_j = \{h\in\mathcal{H} : r^{2(j-1)}a^2 < -E_P(h) \le r^{2j}a^2\}$, $j=1,2,\ldots$. Then,

$$\begin{aligned}
V_a &\le \sup_{h\in\mathcal{H}_0}\left\{\frac{E_n(h)-E_P(h)}{-E_P(h)+a^2}\right\} + \sum_{j\ge1}\sup_{h\in\mathcal{H}_j}\left\{\frac{E_n(h)-E_P(h)}{-E_P(h)+a^2}\right\} \\
&\le \frac{1}{a^2}\left[\sup_{h\in\mathcal{H}_0}\big(E_n(h)-E_P(h)\big) + \sum_{j\ge1}\big(1+r^{2(j-1)}\big)^{-1}\sup_{h\in\mathcal{H}_j}\big(E_n(h)-E_P(h)\big)\right] \qquad \text{(A.20)} \\
&\le \frac{1}{a^2}\left[\sup_{-E_P(h)\le a^2}\big(E_n(h)-E_P(h)\big) + \sum_{j\ge1}\big(1+r^{2(j-1)}\big)^{-1}\sup_{-E_P(h)\le r^{2j}a^2}\big(E_n(h)-E_P(h)\big)\right].
\end{aligned}$$

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

$$\begin{aligned}
V_a \le \frac{1}{a^2}\Bigg[&\sup_{\|h\|_{L_2(P)}\le c_1^{1/2}a^{\frac{\alpha}{1+\alpha}}}\big(E_n(h)-E_P(h)\big) \\
&+ \sum_{j\ge1}\big(1+r^{2(j-1)}\big)^{-1}\sup_{\|h\|_{L_2(P)}\le c_1^{1/2}r^{\frac{\alpha}{1+\alpha}j}a^{\frac{\alpha}{1+\alpha}}}\big(E_n(h)-E_P(h)\big)\Bigg].
\end{aligned}$$

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

$$E_{P^n}(V_a) \le C_2\frac{c_1^{\frac12}}{a^2}\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}}\sum_{j\ge0}\frac{r^{\frac{\alpha}{1+\alpha}j}}{1+r^{2(j-1)}} \le C_2 c_1^{\frac12}\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}\left(\frac{r^2}{1-r^{-\frac{2+\alpha}{1+\alpha}}}\right) \le c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2}$$

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 $c_2 = C_2 c_1^{\frac12}\left(\frac{r^2}{1-r^{-\frac{2+\alpha}{1+\alpha}}}\right) \vee 1$ is a constant greater than or equal to 1 and depends only on

$(M, \kappa, \eta, \alpha)$, as $r > 1$ is fixed. We plug this upper bound into (A.19) to obtain

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

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

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

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

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

which can be reduced to an innocuous restriction $n \geq 1$ 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

$$c_2\sqrt{\frac{v}{n}}a^{\frac{\alpha}{1+\alpha}-2} \leq \frac{\varepsilon_n}{a} = \frac{1}{\sqrt{kt}} \leq \frac{1}{\sqrt{k}} \quad \text{and}$$
$$a^{\frac{2\alpha}{1+\alpha}-2} = a^{2(\frac{\alpha}{1+\alpha}-2)}a^2 \leq \left[\varepsilon_n^{\frac{\alpha}{1+\alpha}-2}\right]^2\varepsilon_n^2 = c_2^{-2}v^{-1}n\varepsilon_n^2,$$

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

$$\begin{aligned} V_a &\leq \frac{1}{\sqrt{k}} + \sqrt{2\frac{c_1c_2^{-2}v^{-1}n\varepsilon_n^2 + 8}{nk\varepsilon_n^2}} + \frac{2}{3nk\varepsilon_n^2} \\ &= \frac{1}{\sqrt{k}} + \sqrt{\frac{2c_1c_2^{-2}v^{-1}}{k} + \frac{8}{nk\varepsilon_n^2}} + \frac{2}{3nk\varepsilon_n^2} \\ &\leq \frac{1}{\sqrt{k}} + \sqrt{\frac{2c_1c_2^{-2}v^{-1}}{k} + \frac{8}{k}} + \frac{2}{3k} \quad \text{for } n\varepsilon_n^2 \geq 1. \end{aligned} \tag{A.24}$$

Note that condition $n\varepsilon_n^2 \geq 1$ 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

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

Hence, (A.18) yields

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

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

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

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

PROOF OF THEOREM 2.4: As in the proof of Theorem 2.2, we work with the normalized outcome support, $Y_{1,i}, Y_{0,i} \in [-\frac{1}{2}, \frac{1}{2}]$. 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 $x_1, \ldots, x_v \in \mathcal{X}$ be v points that are shattered by $\mathcal{G}$, and let γ be a positive number satisfying $\gamma \leq \min\{\eta, \frac{1}{2}\}$, whose proper choice will be given later. We fix the marginal distribution of X at the one supported only on $(x_1, \ldots, x_v)$ and having the probability mass function

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

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 $\mathcal{P}^* = \{P_\mathbf{b} : \mathbf{b} \in \{0, 1\}^{v-1}\}$. 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 that

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

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, $\mathcal{P}^* \subset \mathcal{P}_{\mathrm{FB}}(1, \kappa, \eta, \alpha)$ holds.

Let $\pi(\mathbf{b})$ be a prior distribution for $\mathbf{b}$ such that $b_1, \ldots, b_{v-1}$ are i.i.d. and $b_1 \sim \text{Ber}(1/2)$. 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 jth element is $\hat{b}_j = 1\{x_j \in \hat{G}\}$.

The welfare loss can be bounded from below as follows:

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

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 $\hat{G}$. The infimum in the last line is taken over decision sets $G_n = \{x_j : b_n(x_j) = 1\}$ that are constrained to contain $\{x_v\}$, that is, $b_n(x_v) = 1$.

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

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

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 $\exp\{-2\sqrt{2}\}$, so we obtain the

following lower bound:

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

Recall that γ is constrained to $\gamma \leq \min\{\eta, \frac{1}{4}\}$. 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, that is, multiply M to (A.26) and substitute η/M for η appearing in (A.26) and (A.27). *Q.E.D.*

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 that $\hat{W}_n^{\tau}(\hat{G}_{m\text{-hybrid}}) - \hat{W}_n^{\tau}(\tilde{G}) \geq 0$ for any $\tilde{G} \in \mathcal{G}$. The following inequalities therefore follow:

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

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

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

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 $\hat{W}_n(G) \equiv E_n[\frac{Y_i(1-D_i)}{1-e(X_i)} + \hat{\tau}_i^e \cdot 1\{X_i \in G\}]$, respectively, yields a

similar upper bound

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

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 $E_{P^n}[\sup_{G\in\mathcal{G}}|W_n^\tau(G) - W(G)|]$ is $n^{-1/2}$, same as that of $E_{P^n}[\sup_{G\in\mathcal{G}}|W_n(G) - W(G)|]$, since the proof of Theorem 2.1 can be applied to the following class of functions:

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

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 $E_{P^n}[\sup_{G\in\mathcal{G}}|W_n(G) - W(G)|]$, (A.31) leads to the uniform convergence rate of $\phi_n^{-1} \vee n^{-1/2}$ for the e-hybrid rule. *Q.E.D.*

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 by Shin Kanaya and Dennis Kristensen in unpublished work on bandwidth choice.

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 $B = 0$,

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

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

$$\frac{\partial g}{\partial B}(A, \tilde{B}) = -\frac{\frac{\partial f}{\partial B}(\tilde{X}, A, \tilde{B})}{\frac{\partial f}{\partial X}(\tilde{X}, A, \tilde{B})} = \frac{1}{1 - \frac{\alpha}{1+\alpha}A\tilde{X}^{-\frac{1}{1+\alpha}}} = \frac{\tilde{X}}{\frac{\tilde{X}}{1+\alpha} + \frac{\alpha}{1+\alpha}\big(\tilde{X} - A\tilde{X}^{\frac{\alpha}{1+\alpha}}\big)}$$

$$= \frac{\tilde{X}}{\frac{\tilde{X}}{1+\alpha} + \frac{\alpha}{1+\alpha}\tilde{B}} \leq 1 + \alpha.$$

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

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 i.i.d. from $P_X(\cdot)$. Then,*

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

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 $\tilde{G} = G^*_{\mathrm{FB}}$ in (A.29) and rewrite (A.29) in terms of the normalized welfare loss for $\hat{G}_{m\text{-hybrid}}$,

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

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

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

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 $G \in \mathcal{G}$,

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

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

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

inequality (A.32) implies

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

Denote event $\{V_a^\tau < \frac{1}{2}\}$ by Ω_t, which is equivalent to event $\{d(G^*_{\mathrm{FB}}, \hat{G}_{m\text{-hybrid}}) \le 2\rho_n + k\varepsilon_n^2 t\}$. The same line of argument that leads to (A.25) in the proof of Theorem 2.3 leads to, for $t \ge 1$,

$$P^n(\Omega_t) = P^n\big(d\big(G^*_{\mathrm{FB}}, \hat{G}_{m\text{-hybrid}}\big) \le 2\rho_n + k\varepsilon_n^2 t\big) \ge 1 - \exp(-t), \tag{A.34}$$

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

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

where

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

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

$$\Lambda_1 = \big\{\mathcal{V}_{0,n} \le n^{-\lambda}\big\}, \qquad \Lambda_2 = \big\{P_X\big(G^*_{\mathrm{FB}} \triangle \hat{G}_{m\text{-hybrid}}\big) \ge n^{-\lambda}\big\}.$$

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

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

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 $\Lambda_1 \cap \Lambda_2 \cap \Omega_t$,

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

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 $t \ge 1$,

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

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 $d(G^*_{\mathrm{FB}}, \hat{G}_{m\text{-hybrid}}) \leq 1$. In the fourth line, $A_{3,n}$ follows from Lemma A.9, $A_{4,n}$ follows from $d(G^*_{\mathrm{FB}}, \hat{G}_{m\text{-hybrid}}) \leq P_X(G^*_{\mathrm{FB}} \triangle \hat{G}_{m\text{-hybrid}})$ and $P_X(G^*_{\mathrm{FB}} \triangle \hat{G}_{m\text{-hybrid}}) < n^{-\lambda}$ 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

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

where the third line follows from Jensen's inequality since $(1+\alpha)/2 \leq 1$. Hence, $A_{1,n}$ satisfies $\sup_{P \in \mathcal{P}_m} A_{1,n} = O(\tilde{\psi}_n^{-(1+\alpha)})$. 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 $E_{P^n}(d(G^*_{\mathrm{FB}}, \hat{G}_{m\text{-hybrid}}))$ over $P \in \mathcal{P}_m \cap \mathcal{P}_{\mathrm{FB}}(M, \kappa, \eta, \alpha)$ is bounded by the convergence rates of the $A_{1,n}$ and $A_{2,n}$,

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

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

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

and that inequality (A.33) is replaced by

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

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. *Q.E.D.*

APPENDIX B: INFERENCE FOR WELFARE GAIN

In the proposed EWM procedure, the maximized empirical welfare $W_n(\hat{G}_{\mathrm{EWM}})$ can be seen as an estimate of $W(\hat{G}_{\mathrm{EWM}})$, the welfare level attained by implementing the estimated treatment rule.[1] In situations where propensity scores are known, this section pro-

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

vides 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 $\hat{G}_{\mathrm{EWM}}$ 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 $\nu_n(G) = \sqrt{n}(V_n(G) - V(G))$, where $V_n(G) \equiv W_n(G) - W_n(G_0)$. If there is a random variable $\bar{\nu}_n$ such that $\nu_n(\hat{G}) \leq \bar{\nu}_n$ 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 that

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

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

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

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

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

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

[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 1. Such $\hat{\mathcal{G}}$ can be obtained by applying the technique of contact set estimation in the context of stochastic dominance testing. See Linton, Song, and Whang (2010) and Donald and Hsu (2016), as well as the literature on moment inequalities with moment selection (Andrews and Shi (2013), among others).

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 $V_n^{\mathrm{dm}}(\hat{G}) \equiv W_n^{\mathrm{dm}}(\hat{G}) - W_n^{\mathrm{dm}}(G_0)$. Resampling in this case is from observations $Z_i^{\mathrm{dm}} = (Y_i^{\mathrm{dm}}, D_i, X_i)$, with outcomes $Y_i^{\mathrm{dm}} = Y_i - E_n[Y_i]$ demeaned by the outcome mean in the original sample.

APPENDIX C: COMPUTING EWM TREATMENT RULES

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

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

where each g_i is a function of the data, that is, for the EWM rule $\hat{G}_{\mathrm{EWM}}$, $g_i = \frac{1}{n}(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)})$, for the e-hybrid rule $\hat{G}_{e\text{-hybrid}}$, $g_i = \hat{\tau}_i^e/n$, and for the m-hybrid rule $\hat{G}_{m\text{-hybrid}}$, $g_i = \hat{\tau}^m(X_i)/n$. 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

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

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

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

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 observed that the maximum score objective function could be rewritten as a Mixed Integer Linear Programming problem with additional binary parameters $(z_1, \ldots, z_n)$ that replace the indicator functions $1\{X_i^T\beta \geq 0\}$. The equality $z_i = 1\{X_i^T\beta \geq 0\}$ 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:

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

$$\text{s.t.} \quad \frac{X_i^T\beta}{C_i} < z_i \leq 1 + \frac{X_i^T\beta}{C_i} \quad \text{for } i = 1, \ldots, n, \tag{C.4}$$
$$z_i \in \{0, 1\},$$

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) imposed 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 contain large sets of observations that differ from each other in only one covariate. Suppose that m observations $i_1, \ldots, i_m$ differ only in the value of the last covariate: $X_{i_1} = (1, \tilde{x}_1, \ldots, \tilde{x}_{d_x-1}, x_{d_x,i_1}), \ldots, X_{i_m} = (1, \tilde{x}_1, \ldots, \tilde{x}_{d_x-1}, x_{d_x,i_m})$, 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 $V(\hat{G}_{\mathrm{EWM}})$ of the empirical welfare-maximizing policy requires computing $\bar{\nu}_n^* = \sup_{G \in \mathcal{G}} \sqrt{n}(V_n^*(G) - V_n(G))$ 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

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

The optimization problem for $\bar{\nu}_n^*$ is analogous to the optimization problem for $\hat{G}_{\mathrm{EWM}}$. Furthermore, solving it does not require the knowledge of $\hat{G}_{\mathrm{EWM}}$; 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 ≤ age ≤ 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 ≥ 40) OR (length of unemployment ≥ 2 years).

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

$$\mathcal{G} = \bigl\{G_{\beta^1,\ldots,\beta^J}, \beta^1, \ldots, \beta^J \in \mathbf{B}\bigr\},$$
$$G_{\beta^1,\ldots,\beta^J} = \bigl\{x : x^T\beta^1 \geq 0, \ldots, x^T\beta^J \geq 0\bigr\}.$$

Then the optimization problem (C.1) becomes

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

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

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

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

$$1 - J + \sum_{j=1}^{J} z_i^j \leq z_i^* \leq J^{-1} \sum_{j=1}^{J} z_i^j \quad \text{for } 1 \leq i \leq n, \tag{C.9}$$
$$z_i^1, \ldots, z_i^J, z_i^* \in \{0, 1\} \quad \text{for } 1 \leq i \leq n.$$

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

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

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

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

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

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

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

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

This problem cannot be rewritten as a linear optimization problem in the same way as (C.3) because the factor $\min\{1, \frac{Kn}{\sum_{i=1}^{n} 1\{X_i^T \beta \geq 0\}}\}$ 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:

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

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. [22]

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. [10]

DONALD, S. G., AND Y.-C. HSU (2016): "Improving the Power of Tests of Stochastic Dominance," *Econometric Reviews*, 35 (4), 553–585. [22]

DUDLEY, R. M. (1999): *Uniform Central Limit Theorems*. Cambridge: Cambridge University Press. [2]

FLORIOS, K., AND S. SKOURAS (2008): "Exact Computation of Max Weighted Score Estimators," *Journal of Econometrics*, 146 (1), 86–91. [24]

LINTON, O., K. SONG, AND Y. WHANG (2010): "An Improved Bootstrap Test of Stochastic Dominance," *Journal of Econometrics*, 154, 186–202. [22]

LUGOSI, G. (2002): "Pattern Classification and Learning Theory," in *Principles of Nonparametric Learning*, ed. by L. Györfi. Vienna: Springer, 1–56. [2]

MASSART, P., AND É. NÉDÉLEC (2006): "Risk Bounds for Statistical Learning," *The Annals of Statistics*, 34 (5), 2326–2366. [2]

TSYBAKOV, A. B. (2004): "Optimal Aggregation of Classifiers in Statistical Learning," *The Annals of Statistics*, 32 (1), 135–166. [11]

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. New York: Springer. [2,5,19,23]

Co-editor Liran Einav handled this manuscript.

Manuscript received 9 March, 2015; final version accepted 29 November, 2017; available online 29 November, 2017.