Online Supplement to "Equality-Minded Treatment Choice"

Toru Kitagawa* and Aleksey Tetenov†

November 19, 2019

Abstract

This online supplement contains the materials and proofs omitted from Kitagawa and Tetenov (2019), "Equality-minded Treatment Choice."

B Illustrative Example

In this section, we illustrate the properties of rank-dependent SWFs in comparison with the utilitarian one in a simple setting with the Gini SWF, $W_{Gini}(F) = \int_0^\infty (1-F(y))^2 dy$. We first compare the welfare ordering on the parametric family of log-normal outcome distributions. Second, we consider a simple treatment choice problem with binary X in order to illustrate how the optimal rules fundamentally differ between the two SWFs.

First, consider the welfare ordering over the family of log-normal distributions of outcomes, $Y \sim \log N(\mu, \sigma^2)$, ignoring the treatment choice problem. The mean of Y is given by $E(Y) = \exp(\mu + \sigma^2/2)$. The Gini inequality coefficient for $\log N(\mu, \sigma^2)$ is given by $2\Phi\left(\sigma/\sqrt{2}\right) - 1$ (see, e.g., Cowell (1995)), where $\Phi(\cdot)$ is the cdf of the standard normal distribution. By (5), we have

$$W_{Gini}(\mu, \sigma) \equiv 2\exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi\left(\frac{\sigma}{\sqrt{2}}\right)\right]. \tag{B.1}$$

This welfare function is increasing in μ, whereas it is not monotonic in σ. For instance, when $\mu = 0$, $W_{Gini}(\mu, \sigma)$ is decreasing in σ for $\sigma < 0.87$ and increasing for $\sigma > 0.87$. See Figure B.1 for a plot of $W_{Gini}(\mu, \sigma)$ over $\sigma \in [0, 2]$ holding $\mu = 0$ fixed. The U-shape of the Gini social

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

†University of Geneva, Email: aleksey.tetenov@unige.ch

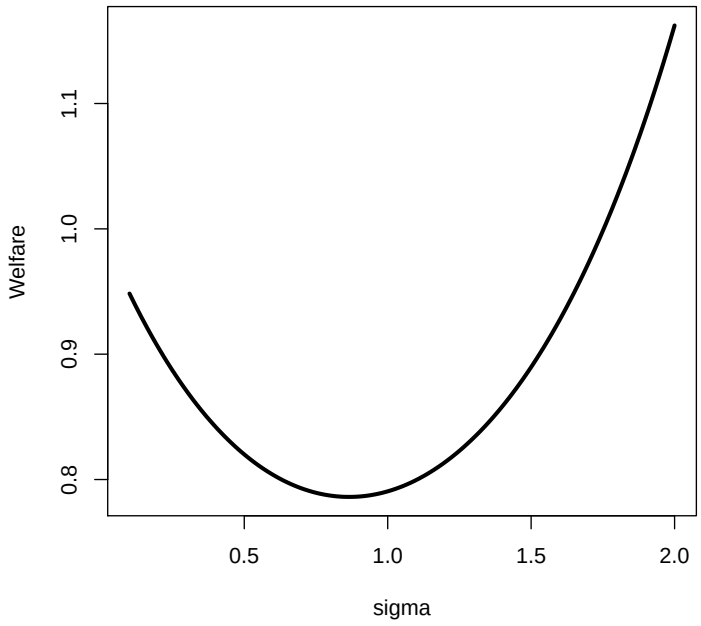

Figure B.1: Equality-minded welfare for $\log N(0, \sigma^2)$.

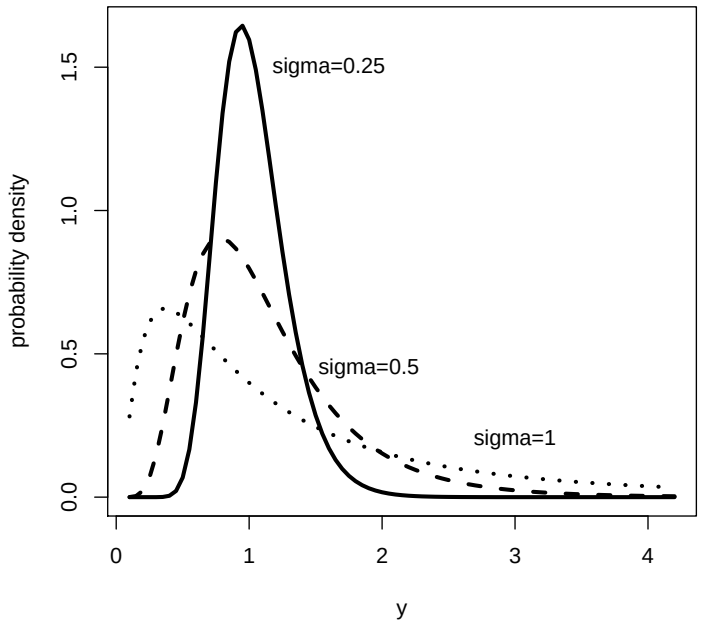

Figure B.2: Density of $\log N(0, \sigma^2)$.

welfare indicates that for $\sigma < 0.87$, the negative contribution to the social welfare from an increase in the Gini coefficient dominates the positive contribution from an increase in the mean, while for $\sigma > 0.87$, this relationship reverses. In Figure B.2, we plot the densities of the log-normal distributions for $\sigma = 0.25$, 0.5, and 1. Since $E(Y)$ is monotonically increasing both in μ and σ, higher σ is always preferable in terms of the utilitarian social welfare. In contrast, as shown in the welfare values plotted in Figure B.1, the Gini social welfare yields the complete opposite welfare ordering over the three log-normal distributions in Figure B.2.

Consider now the treatment choice problem. Suppose there is only one binary covariate $X \in \{a, b\}$ with $\Pr(X = a) = \Pr(X = b) = 1/2$. Consider the following parameterization of the potential outcome distributions:

$$\begin{aligned} &Y_1|(X = a) \sim \log N(\mu_a, \sigma_a^2), \quad Y_0|(X = a) \sim \log N(0, 0.8^2), \\ &Y_1|(X = b) \sim \log N(\mu_b, \sigma_b^2), \quad Y_0|(X = b) \sim \log N(0, 0.8^2). \end{aligned} \tag{B.2}$$

According to Theorem 2.1, it suffices to consider non-randomized rules to search for an optimal one. We therefore consider ranking the following four policies: $\mathcal{G} = \{\emptyset, \{a\}, \{b\}, \{a, b\}\} \equiv \{G_\emptyset, G_a, G_b, G_{ab}\}$.

Suppose $\sigma_a = \sigma_b = 0.8$ and $\mu_a, \mu_b > 0$. Then, in each subpopulation of $X = a$ and $X = b$, the distribution of Y_1 stochastically dominates the distribution of Y_0. Since the rank-dependent social welfare is clearly monotonic in the first-order stochastic dominance relationship, treating both $\{X = a\}$ and $\{X = b\}$ maximizes the Gini social welfare. This optimal rule indeed coincides with that of the utilitarian welfare case. In general, when stochastic dominance relationships between $Y_1|X$- and $Y_0|X$-distributions are present for all X, the optimal rule for the rank-dependent social welfare agrees with the utilitarian one and can be obtained by solving the treatment choice problem separately in each subpopulation.

These results change drastically once we let $\sigma_a \neq \sigma_b$. Suppose we fix $\mu_a = \mu_b = 0$, while we vary both σ_a and σ_b over $[0.6, 1.2]$. As the mean of a log normal random variable is

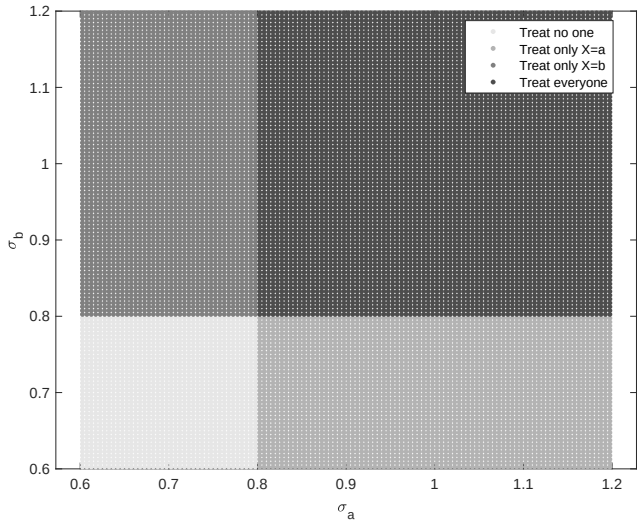

Figure B.3: Optimal policies under the additive welfare. Log-normal potential outcome distributions with $\mu_a = \mu_b = 0$.

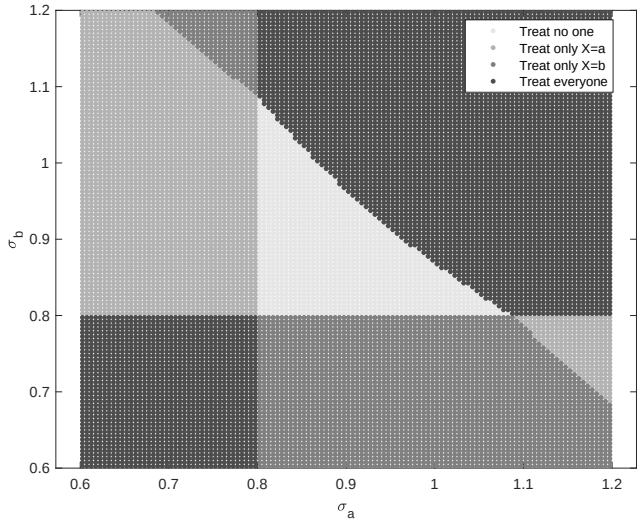

Figure B.4: Optimal policies under the Gini welfare. Log-normal potential outcome distributions with $\mu_a = \mu_b = 0$.

increasing in σ, the optimal treatment rule for the additive welfare is obtained by

$$G^*_{Add} = \begin{cases} G_\emptyset & \text{if } \sigma_a < 0.8 \text{ and } \sigma_b < 0.8, \\ G_a & \text{if } \sigma_a \geq 0.8 \text{ and } \sigma_b < 0.8, \\ G_b & \text{if } \sigma_a < 0.8 \text{ and } \sigma_b \geq 0.8, \\ G_{ab} & \text{if } \sigma_a \geq 0.8 \text{ and } \sigma_b \geq 0.8. \end{cases}$$

In Figure B.3, we plot the optimal treatment rule under the additive welfare at each grid point of $(\sigma_a, \sigma_b) \in [0.6, 1.2]^2$. Since the additive social welfare is separable over the subpopulations, a treatment preferable for one subpopulation does not depend on the treatment assigned to the other subpopulation. The regions in which different rules from $\mathcal{G}$ are optimal form a quadrant partition, as shown in Figure B.3.

In Figure B.4, we plot the optimal policies in terms of the Gini social welfare. The regions in which different rules from $\mathcal{G}$ are optimal are strikingly different compared with the additive welfare case (G^*_{Add}) shown in Figure B.3. In the neighborhood of $(\sigma_a, \sigma_b) = (0.8, 0.8)$, the subpopulations to be treated under the Gini social welfare are the converse of those to be treated under the utilitarian welfare. This is because the Gini social welfare is decreasing in σ in the neighborhood of $\sigma = 0.8$ (Figure B.1), while the additive welfare is monotonically increasing in σ. Another notable difference is that in contrast to the quadrant partition observed in the additive welfare case, the partition in the equality-minded welfare case is more complex. Some treatment rules are optimal in disconnected regions, e.g., G_{ab} is optimal in the south-west and the north-east regions of the plot. Furthermore, the region in which G_a is optimal can border the region in which G_b is optimal. On the border between these regions, the policy maker chooses whether to treat $X = a$ only or $X = b$ only, rather than whether to additionally treat the other subpopulation.

The non-additive Gini SWF can be locally approximated by an additive SWF in a neighborhood of the baseline outcome distribution (Kasy, 2016) as follows. Let F_0 be the baseline outcome distribution. Then the Gini SWF evaluated at an outcome distribution F local to F_0 is approximately

$$W_{Gini}(F) \approx W_{Gini}(F_0) + \int_0^\infty IF(y; W_{Gini}, F_0) dF(y), \tag{B.3}$$

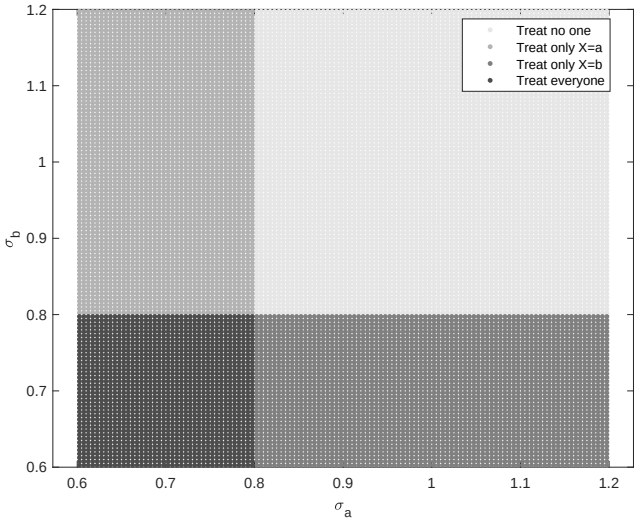

Figure B.5: Optimal policies under the additive approximation of the Gini welfare. Log-normal potential outcome distributions with $\mu_a = \mu_b = 0$.

where $IF(y; W_{Gini}, F_0)$ is the influence function of $W_{Gini}(\cdot)$ at F_0 (see, e.g., Wasserman (2006) for the definition of the influence function)

$$IF(y; W_{Gini}, F_0) = -2W_{Gini}(F_0) + 2\int_0^y (1 - F_0(\tilde{y}))d\tilde{y}.$$

We examine in the current example how policies derived from this additive approximation differ from those maximizing the original non-additive Gini SWF. We set the baseline outcome distribution F_0 to $\log N(0, 0.8^2)$ (the outcome distribution of Y_0 in the population) and then evaluate the additive approximation (B.3) for distributions yielded by each treatment rule at different parameter values (σ_a, σ_b).

Figure B.5 plots policies that maximize the additive approximation (B.3) to the Gini SWF. Since the approximation is additive with respect to F, the treatment chosen for subpopulation with $X = a$ does not depend on the treatment chosen for the subpopulation with $X = b$. We hence obtain a quadrant partition similar to Figure B.2. The additive approximation, however, applies a concave function to the outcomes and recommends treatment only for subpopulations with $\sigma_x < 0.8$ (i.e., where $Y_1|X$ has a distribution with a lower variance, albeit also a lower mean).

A comparison of Figures B.4 and B.5 shows that the optimal policies under the additive

approximation agree with those under Gini SWF in the neighborhood of the baseline distribution, when both σ_a and σ_b are in $[0.6, 0.9]$. On the other hand, when either of the potential outcome distributions $P(Y_1|X=a)$ or $P(Y_1|X=b)$ sufficiently deviates from the baseline distribution, the additive approximation no longer yields the same treatment preference as the Gini SWF it is meant to approximate.

C EWM with Estimated Propensity Score

Unknown propensity score is common in observational studies. This section considers the equality-minded EWM approach with estimated propensity scores and investigates the influence of the lack of knowledge on propensity scores on the uniform convergence rate of the welfare loss criterion.

Let $\hat{e}(x)$ be an estimator for the propensity score $\Pr(D=1|X=x)$. The empirical welfare criterion of assignment policy $\{X \in G\}$ with the estimated propensity scores plugged in is given by

$$\widehat{W}^e_\Lambda(G) = \int_0^\infty \Lambda(\widehat{F}^e_G(y) \vee 0)dy,$$
$$\widehat{F}^e_G(y) \equiv 1 - \frac{1}{n}\sum_{i=1}^n \left[\frac{D_i}{\hat{e}(X_i)} \cdot 1\{X_i \in G\} + \frac{(1-D_i)}{1-\hat{e}(X_i)} \cdot 1\{X_i \notin G\}\right] \cdot 1\{Y_i > y\}.$$

The equality-EWM rule with estimated propensity score is defined accordingly as

$$\widehat{G}^e \in \arg\max_{G \in \mathcal{G}} \widehat{W}^e_\Lambda(G).$$

To characterize the uniform convergence rate of the welfare loss of $\widehat{G}^e$, we first assume that $\hat{e}(\cdot)$ is uniformly consistent to the true propensity score $e(\cdot)$ in the following sense.

Assumption C.1. For a class of data generating processes $\mathcal{P}_e$, there exist sequences $\phi_n, \tilde{\phi}_n \to \infty$ as $n \to \infty$ such that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_e} \phi_n E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right] < \infty, \tag{C.1}$$
$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_e} \phi_n E_{P^n}\left[\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right|\right] < \infty,$$

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_e} E_{P^n}\left[\left(\tilde{\phi}_n \max_{1\le i\le n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\right)^2\right]<\infty, \quad \text{and} \tag{C.2}$$

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_e} E_{P^n}\left[\left(\tilde{\phi}_n \max_{1\le i\le n}\left|\frac{1}{1-e(X_i)}-\frac{1}{1-\hat{e}(X_i)}\right|\right)^2\right]<\infty$$

hold.

When the class of data generating processes $\mathcal{P}_e$ constrains the propensity score to a parametric family with compact support of X, a parametric estimator $\hat{e}(X_i)$ satisfies this assumption with $\phi_n = \tilde{\phi}_n = n^{1/2}$. When the propensity scores are estimated nonparametrically instead, ϕ_n and $\tilde{\phi}_n$ are generally slower than $n^{1/2}$. The rates of ϕ_n and $\tilde{\phi}_n$ for nonparametrically estimated propensity scores depend on the smoothness of $e(\cdot)$ and the dimension of X, as we discuss further below.

Theorem C.1. *Suppose Assumptions 2.1, 2.2 and 3.1 hold. For a class of data generating processes $\mathcal{P}_e$, if an estimator for the propensity score satisfies Assumption C.1, then*

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^e)\right] \le O\left(\phi_n^{-1}\vee\sqrt{\frac{v}{n}}\right). \tag{C.3}$$

Proof. See Appendix D.

This theorem extends Theorem 2.5 (e) of Kitagawa and Tetenov (2018a) to the cases of rank-dependent social welfare or unbounded outcome or both. The shown uniform convergence rate implies that the parametrically estimated propensity score achieving $\phi_n = n^{1/2}$ does not affect the convergence rate property of the welfare loss. With nonparametrically estimated propensity score, on the other hand, the uniform welfare loss convergence rate can be slower than the one with the known propensity score obtained in Theorem 3.1. For instance, if $\hat{e}(X_i)$ is estimated by local polynomial regression (with proper trimming), then for a suitably defined $\mathcal{P}_e$, we have $\phi_n = n^{\frac{1}{2+d_x/\beta_e}}$ and $\tilde{\phi}_n = \log n \cdot (\log n/n)^{\frac{1}{2+d_x/\beta_e}}$, where $\beta_e \ge 1$ is the parameter constraining smoothness of $e(\cdot)$ in terms of the degree of the Hölder class of functions and $d_x \ge 1$ is the dimension of X. Since $\frac{1}{2+d_x/\beta_e} < \frac{1}{2}$, the upper bound of the uniform convergence rate shown in Theorem C.1 implies

$$\sup_{P\in\mathcal{P}_e\cap\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^e)\right] \le O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right), \tag{C.4}$$

as long as the VC-dimension of $\mathcal{G}$ is either constant or does not grow too fast as the sample size increases. For a formal derivation of (C.4) and the precise construction of the local polynomial estimator for $e(\cdot)$, see Appendix D.

D Additional proofs

Proof of (18) in Theorem 3.1. Similarly to inequalities (A.14) and (A.15) shown in Appendix A, the average welfare regret for the normalized cdf case can be bounded by

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^R)\right] \leq 2|\Lambda'(0)| \int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^R(y) - F_G(y)\right|\right] dy. \tag{D.1}$$

We hence focus on bounding $\int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^R(y) - F_G(y)\right|\right] dy$.

Let $w_G(Z_i)$ be as defined in (A.16), and let

$$A_{n,G} \equiv \{\widehat{F}_G(-\infty) < 1\} = \{n^{-1}\sum_{i=1}^n w_G(Z_i) > 0\}$$

denote the event that the normalizing term in $\widehat{F}_G^R(y)$ at policy G is nonzero, and $A_{n,G}^c \equiv \{\widehat{F}_G(-\infty) = 1\} = \{n^{-1}\sum_{i=1}^n w_G(Z_i) = 0\}$ be the complement of $A_{n,G}$. Using the indicator functions for $A_{n,G}$ and $A_{n,G}^c$, $\widehat{F}_G^R(y)$ can be written as

$$\widehat{F}_G^R(y) = \left[1 - \frac{1}{n}\sum_{i=1}^n w_{G,i}^R 1\{Y_i > y\}\right] \cdot 1\{A_{n,G}\} + [1 - 1\{y < \min_{1\leq i\leq n} Y_i\}] \cdot 1\{A_{n,G}^c\}, \tag{D.2}$$

where

$$w_{G,i}^R = \frac{w_G(Z_i)}{n^{-1}\sum_{i=1}^n f_G(Z_i) + 1}, \quad f_G(Z_i) = w_G(Z_i) - 1. \tag{D.3}$$

By the triangle inequality,

$$\begin{aligned}
&\left|\widehat{F}_G^R(y) - F_G(y)\right| \leq \left|\widehat{F}_G^R(y) - \widehat{F}_G(y)\right| + \left|\widehat{F}_G(y) - F_G(y)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \left[w_{G,i}^R - w_G(Z_i)\right] 1\{Y_i > y\}\right| \cdot 1\{A_{n,G}\} + 1\{y < \min_{1\leq i\leq n} Y_i\} \cdot 1\{A_{n,G}^c\} + \left|\widehat{F}_G(y) - F_G(y)\right|.
\end{aligned} \tag{D.4}$$

Let

$$S_n^- \equiv \inf_{G\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n f_G(Z_i),$$

$$S_n \equiv \sup_{G\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} f_G(Z_i) \right|.$$

For $\delta \in (0,1)$ fixed, define

$$\Omega_{n,\delta} \equiv \{S_n - E_{P^n}(S_n) \leq \delta/2\} = \{-S_n \geq -E_{P^n}(S_n) - \delta/2\}.$$

By Lemma A.3, $\{f_G : G \in \mathcal{G}\}$ is a VC-subgraph class of functions with VC-dimension at most v with $E_P(f_G) = 0$ and an envelope $\|f_G\|_\infty \leq \frac{1-\kappa}{\kappa}$. Hence, by Lemma A.4,

$$E_{P^n}(S_n) \leq C_1 \frac{1-\kappa}{\kappa}\sqrt{\frac{v}{n}}$$

holds, where C_1 is the universal constant defined in Lemma A.4. Accordingly, for all $n > n(\delta, v) \equiv \left(\frac{C_1(1-\kappa)}{\kappa(1-\delta)}\right)^2 v$, $-E_{P_n}(S_n) - \delta/2 > -1 + \delta/2$ holds. Since $S_n^- \geq -S_n$ holds, $\Omega_{n,\delta}$ being true and $n > n(\delta, v)$ imply $S_n^- > -1 + \delta/2$. Hence, on $\Omega_{n,\delta}$ and for $n > n(\delta, v)$, we have $0 \leq w_{G,i}^R \leq (2/\delta) w_G(Z_i)$ and

$$\left| w_{G,i}^R - w_G(Z_i) \right| = w_{G,i}^R \left| \frac{1}{n}\sum_{i=1}^{n} f_G(Z_i) \right| \leq \frac{2}{\delta} \cdot w_G(Z_i) S_n. \tag{D.5}$$

On $\Omega_{n,\delta}^c$ and for G such that $A_{n,G}$ is true, we have $0 \leq w_{G,i}^R \leq n$ and

$$\left| w_{G,i}^R - w_G(Z_i) \right| \leq n \frac{1-\kappa}{\kappa}. \tag{D.6}$$

Combining (D.5) and (D.6), (D.4) can be rewritten as

$$\begin{aligned}
&\left| \widehat{F}_G^R(y) - F_G(y) \right| \\
\leq &\frac{2}{\delta} \cdot S_n \cdot \frac{1}{n}\sum_{i=1}^{n} w_G(Z_i) 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta} \cap A_{n,G}\} + n\frac{1-\kappa}{\kappa} \cdot \frac{1}{n}\sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}\} \\
&+ 1\{y < \min_{1\leq i\leq n} Y_i\} \cdot 1\{\Omega_{n,\delta} \cap A_{n,G}^c\} + 1\{y < \min_{1\leq i\leq n} Y_i\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}^c\} + \left| \widehat{F}_G(y) - F_G(y) \right|
\end{aligned} \tag{D.7}$$

Note that $\{S_n^- > -1\}$ is equivalent to $\left\{ \inf_{G\in\mathcal{G}} n^{-1}\sum_{i=1}^n w_G(Z_i) > 0 \right\}$, implying that $A_{n,G}$ is true for all $G \in \mathcal{G}$. Hence, for $n > n(\delta, v)$, $\Omega_{n,\delta} \cap A_{n,G} = \Omega_{n,\delta}$, and $\Omega_{n,\delta} \cap A_{n,G}^c = \emptyset$ hold for all $G \in \mathcal{G}$. By also noting $w_G(Z_i) \leq \frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)}$, (D.7) can be further bounded by

$$
\begin{aligned}
\leq & \frac{2}{\delta} \cdot S_n \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)} \right] 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}\} \\
& + n \frac{1-\kappa}{\kappa} \cdot \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}\} \\
& + 1\{y < \min_{1 \leq i \leq n} Y_i\} \cdot 1\{\Omega_{n,\delta}^c \cap A_{n,G}^c\} + \left| \widehat{F}_G(y) - F_G(y) \right| \\
\leq & \frac{2}{\delta} \cdot S_n \left[P(Y_1 > y) + P(Y_0 > y) \right] \cdot 1\{\Omega_{n,\delta}\} \\
& + \frac{2}{\delta} \cdot \left(1 - \frac{\delta}{2} \right) \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)} \right) 1\{Y_i > y\} - P(Y_1 > y) - P(Y_0 > y) \right] \cdot 1\{\Omega_{n,\delta}\} \\
& + n \frac{1-\kappa}{\kappa} \cdot \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c\} + \left| \widehat{F}_G(y) - F_G(y) \right|, \\
\leq & \frac{2}{\delta} \cdot S_n \left[P(Y_1 > y) + P(Y_0 > y) \right] \\
& + \frac{2}{\delta} \cdot \left(1 - \frac{\delta}{2} \right) \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\left(\frac{D_i}{e(X_i)} + \frac{1-D_i}{1-e(X_i)} \right) 1\{Y_i > y\} - P(Y_1 > y) - P(Y_0 > y) \right] \\
& + n \frac{1-\kappa}{\kappa} \cdot \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c\} + \left| \widehat{F}_G(y) - F_G(y) \right|,
\end{aligned}
\tag{D.8}
$$

where the second inequality follows from the fact that $S_n \leq \left(1 - \frac{\delta}{2}\right)$ holds on $\Omega_{n,\delta}$ and for $n > n(\delta, v)$, and $n\frac{1-\kappa}{\kappa} \geq 1$ and $n^{-1} \sum_{i=1}^{n} 1\{Y_i > y\} \geq 1\{y < \min_{1 \leq i \leq n} Y_i\}$ hold for all y. Since the second term in (D.8) has mean zero, $E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G^R(y) - F_G(y) \right|\right]$ can be bounded by

$$
\begin{aligned}
& E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G^R(y) - F_G(y) \right| \right] \\
\leq & \underbrace{\frac{2}{\delta} \cdot E_{P^n}\left[S_n\right] \cdot \left(P(Y_1 > y) + P(Y_0 > y) \right)}_{\text{(i)}} + \underbrace{n \frac{1-\kappa}{\kappa} \cdot E_{P^n}\left[\frac{1}{n} \sum_{i=1}^{n} 1\{Y_i > y\} \cdot 1\{\Omega_{n,\delta}^c\} \right]}_{\text{(ii)}} \\
& + \underbrace{E_{P^n}\left[\sup_{G \in \mathcal{G}} \left| \widehat{F}_G(y) - F_G(y) \right| \right]}_{\text{(iii)}}.
\end{aligned}
\tag{D.9}
$$

By Assumption 3.1 (TC) and Lemma A.4, the integral of term (i) in (D.9) can be bounded as

$$
\int_0^\infty \text{(i)} dy \leq \frac{4C_1}{\delta} \cdot \frac{1-\kappa}{\kappa} \sqrt{\frac{v}{n}} \Upsilon, \tag{D.10}
$$

where we use $E_{P^n}[S_n] \leq C_1 \frac{1-\kappa}{\kappa}\sqrt{\frac{v}{n}}$ and $\int_0^\infty P(Y_d > y)dy \leq \int_0^\infty \sqrt{P(Y_d > y)}dy \leq \Upsilon$.

Consider term (ii); by the Cauchy-Schwarz inequality,

$$\begin{aligned}(\text{ii}) &\leq n\frac{1-\kappa}{\kappa}\sqrt{E_{P^n}\left[\left(\frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\}\right)^2\right]}\sqrt{P^n(\Omega_{n,\delta}^c)} \\ &\leq n\frac{1-\kappa}{\kappa}\sqrt{P(Y>y)}\sqrt{P^n(\Omega_{n,\delta}^c)}\end{aligned}$$

Bernstein's inequality (see, e.g., Theorem 12.2 in Boucheron et al. (2013)) implies that

$$P^n(\Omega_{n,\delta}^c) \leq 2P^n\left(-S_n^- - E_{P^n}(-S_n^-) \geq \frac{\delta}{2}\right) \leq 2\exp\left\{-\frac{(\delta/2)^2 n}{2[2(\Sigma_f^2+\sigma_f^2)+\bar{f}\delta/2]}\right\},$$

where $\Sigma_f^2 \equiv E_{P^n}\left[\sup_{G\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n f_G^2(Z_i)\right] \leq \left(\frac{1-\kappa}{\kappa}\right)^2$, $\sigma_f^2 \equiv \sup_{G\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n E_P(f_G^2(Z_i)) \leq \left(\frac{1-\kappa}{\kappa}\right)^2$, and $\bar{f} \equiv \sup_{G\in\mathcal{G}}\|f_G\|_\infty \leq \frac{1-\kappa}{\kappa}$. Hence,

$$\begin{aligned}\sqrt{P^n(\Omega_{n,\delta})} &\leq \sqrt{2}\exp\left\{-\frac{\delta^2 n}{16[2(\Sigma_f^2+\sigma_f^2)+\bar{f}\delta/2]}\right\} \\ &\leq \sqrt{2}\exp\left\{-\frac{\delta^2\kappa^2 n}{16[4(1-\kappa)^2+(1-\kappa)\kappa\delta/2]}\right\} \\ &\leq \sqrt{2}\exp\left\{-c_1(\delta)\kappa^2 n\right\}\end{aligned}$$

holds, where $c_1(\delta) = \delta^2/(64+8\delta) > 0$. The integral of term (ii) can be therefore bounded by

$$\int_0^\infty (\text{ii})dy \leq \frac{2\sqrt{2}(1-\kappa)\Upsilon}{\kappa}\cdot n\exp\left\{-c_1(\delta)\kappa^2 n\right\}. \tag{D.11}$$

As shown in the proof of equation (17) in Theorem 3.1, Lemma A.5 applies to term (iii) to yield

$$\int_0^\infty (\text{iii})dy \leq \frac{2C_T\cdot\Upsilon}{\kappa}\sqrt{\frac{v}{n}} \tag{D.12}$$

Combining (D.1), (D.9), (D.10), (D.11), and (D.12), and setting $\delta = 1/2$, we conclude

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G}^R)\right] \leq \frac{\Lambda'(0)\Upsilon}{\kappa}\left[C_1^R\sqrt{\frac{v}{n}} + 4\sqrt{2}n\exp\{-C_2^R\kappa^2 n\}\right]$$

for all $n > n(1/2, v) = C_3^R\left(\frac{1-\kappa}{\kappa}\right)^2 v$, where $C_1^R = 16C_1 + 4C_T$, $C_2^R = c_1(1/2) = 1/272$, and $C_3^R = 4C_1^2$. □

Proof of Theorem 3.2. We consider a suitable subclass $\mathcal{P}^* \subset \mathcal{P}$, for which the worst case welfare loss can be bounded from below by a distribution-free term that converges at rate $n^{-1/2}$. Specifically, we restrict distributions of potential outcomes to those whose supports are restricted to $[0, \Upsilon]$. Any such distribution satisfies Assumption 3.1 (TC), since $\int_0^\infty \sqrt{P(Y_d > y)}dy = \int_0^\Upsilon \sqrt{P(Y_d > y)}dy \leq \Upsilon$.

To simplify the proof, we normalize the range of outcomes to $Y \in [0, 1]$. We rescale the ourcome to $Y \in [0, \Upsilon]$ in the final step of the proof by multiplying Υ to the regret lower bound, as the rank-dependent SWF is equivariant to a multiplicative positive constant to Y.

The construction of $\mathcal{P}^*$ proceeds as follows. We restrict the range of outcomes to binary $Y \in \{0, 1\}$. By the definition of VC-dimension, there exists a set of v points in $\mathcal{X}$, denoted $x_1, \ldots, x_v \in \mathcal{X}$ that are shattered by $\mathcal{G}$. We constrain the marginal distribution of X to be supported only on $(x_1, \ldots, x_v)$. Let $\tau^* \in (0, 1]$ stated in the current theorem be given. We put mass $p \equiv \frac{\tau^*}{v-1}$ at x_i for all $i < v$, and mass $1 - \tau^*$ at x_v. The constructed marginal distribution of X is common in $\mathcal{P}^*$. Let the distribution of the treatment indicator D be independent of (Y_0, Y_1, X), and let D 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 bit vector used to index a member of $\mathcal{P}^*$, i.e., $\mathcal{P}^* = \{P_\mathbf{b} : \mathbf{b} \in \{0, 1\}^{v-1}\}$ consists of a finite number of DGPs. For each $j = 1, \ldots, (v - 1)$, and depending on $\mathbf{b}$, construct the following conditional distributions of potential outcomes given $X = x_j$; if $b_j = 1$,

$$Y_0|(X = x_j) \sim Ber\left(\frac{1-\gamma}{2}\right), \quad Y_1|(X = x_j) \sim Ber\left(\frac{1+\gamma}{2}\right), \tag{D.13}$$

and, if $b_j = 0$,

$$Y_0|(X = x_j) \sim Ber\left(\frac{1+\gamma}{2}\right), \quad Y_1|(X = x_j) \sim Ber\left(\frac{1-\gamma}{2}\right), \tag{D.14}$$

where $Ber(m)$ denotes the Bernoulli distribution with mean m and $\gamma \in (0, 1)$ is chosen properly in a later step of the proof. For $j = v$, we set the distribution of potential outcomes to be degenerate at the maximum value of Y, $P(Y_0 = Y_1 = 1|X = x_v) = 1$. Clearly, $P_\mathbf{b} \in \mathcal{P}$ for every $\mathbf{b} \in \{0, 1\}^{v-1}$. We accordingly define $\mathcal{P}^* = \left\{P_\mathbf{b} : \mathbf{b} \in \{0, 1\}^{v-1}\right\} \subset \mathcal{P}$.

Note that when the outcome distribution is Bernoulli with mean μ, the equality-minded welfare function equals $W_\Lambda = \Lambda(1 - \mu)$, which is a non-decreasing function of μ. Hence,

given knowledge of $P_{\mathbf{b}}$, an optimal treatment assignment rule for the equality-minded welfare coincides with that for the utilitarian welfare case,

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

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

$$W_\Lambda(G^*_{\mathbf{b}}) = \Lambda(1-\mu^*),$$
$$\mu^* \equiv p(v-1)\left(\frac{1+\gamma}{2}\right) + (1-\tau^*) = \tau^*\left(\frac{1+\gamma}{2}\right) + (1-\tau^*),$$

which does not depend on $\mathbf{b}$.

Let $\widehat{G}$ be an arbitrary treatment choice rule as a function of observations $Z_i \equiv (Y_i, D_i, X_i)$, $i = 1, \ldots, n$, and $\hat{\mathbf{b}} \in \{0,1\}^{(v-1)}$ be a binary vector whose j-th element is $\hat{b}_j = 1\{x_j \in \widehat{G}\}$. Let $\mu_{\widehat{G}}$ be the mean of outcome Y when the treatment assignment rule $\widehat{G}$ is implemented for a given realization of the sample. Outcomes are binary for all $P \in \mathcal{P}^*$, hence

$$\mu_{\widehat{G}} \equiv \int_{\widehat{G}} \Pr(Y_1 = 1|X = x) dP_X(x) + \int_{\widehat{G}^c} \Pr(Y_0 = 1|X = x) dP_X(x).$$

Consider $\pi(\mathbf{b})$, a prior distribution for $\mathbf{b}$, such that $b_1, \ldots, b_{v-1}$ are iid and $b_1 \sim Ber(1/2)$. The welfare loss satisfies the following inequalities:

$$\begin{aligned}
\sup_{P\in\mathcal{P}} E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda(G) - W_\Lambda(\widehat{G})\right] &\geq \sup_{P_{\mathbf{b}}\in\mathcal{P}^*} E_{P^n_{\mathbf{b}}}\left[W_\Lambda(G^*_{\mathbf{b}}) - W_\Lambda(\widehat{G})\right] \\
&\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[W_\Lambda(G^*_{\mathbf{b}}) - W_\Lambda(\widehat{G})\right] d\pi(\mathbf{b}) \\
&= \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\Lambda(1-\mu^*) - \Lambda(1-\mu_{\widehat{G}})\right] d\pi(\mathbf{b}) \\
&\geq \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[|\Lambda'(1-\mu_{\widehat{G}})|(\mu^* - \mu_{\widehat{G}})\right] d\pi(\mathbf{b}) \\
&\geq |\Lambda'(\tau^*)| \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b}), \qquad \text{(D.15)}
\end{aligned}$$

where the fourth line follows since $\Lambda(\cdot)$ is convex and non-increasing. The fifth line follows from the observation that for all $P \in \mathcal{P}^*$, $\mu_G \geq 1 - \tau^*$ for any treatment rule G, therefore $|\Lambda'(1-\mu_{\widehat{G}})| \geq |\Lambda'(\tau^*)|$.

Consider now bounding $\int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b})$ from below. Building on the lower bound calculation for the classification risk of the empirical risk minimizing classifier in Lugosi

(2002), the proof of Theorem 2.2 in Kitagawa and Tetenov (2018a) considers bounding a similar quantity, though the current construction of $\mathcal{P}^*$ is different from the construction in that paper. Therefore, in what follows, we reproduce the proof of Theorem 2.2 in Kitagawa and Tetenov (2018a) with some necessary modifications.

Consider

$$\begin{aligned}\int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[\mu^* - \mu_{\widehat{G}}\right] d\pi(\mathbf{b}) &\geq \gamma \int_{\mathbf{b}} E_{P^n_{\mathbf{b}}}\left[P_X(G^*_{\mathbf{b}} \triangle \widehat{G})\right] d\pi(\mathbf{b}) \\ &= \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\{b(X) \neq \hat{b}(X)\}\right) dP^n\left(Z_1,\ldots,Z_n|\mathbf{b}\right) d\pi(\mathbf{b}) \\ &\geq \inf_{\widehat{G}} \gamma \int_{\mathbf{b}} \int_{Z_1,\ldots,Z_n} P_X\left(\{b(X) \neq \hat{b}(X)\}\right) dP^n\left(Z_1,\ldots,Z_n|\mathbf{b}\right) d\pi(\mathbf{b})\end{aligned}$$

where each $b(X)$ and $\hat{b}(X)$ is an element of $\mathbf{b}$ and $\hat{\mathbf{b}}$ such that $b(x_j) = b_j$, $\hat{b}(x_j) = \hat{b}_j$, and $b(x_v) = \hat{b}(x_v) = 0$. Note that the last expression can be seen as the minimized Bayes risk with the loss function corresponding to the classification error for predicting binary unknown random variable $b(X)$. Hence, the minimizer of the Bayes risk is attained by the Bayes classifier,

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

where $\pi(b_j|Z_1,\ldots,Z_n)$ is the posterior of b_j. The minimized Bayes risk is given by

$$\begin{aligned}&\gamma \int_{Z_1,\ldots,Z_n} E_X\left[\min\left\{\pi\left(b(X) = 1|Z_1,\ldots,Z_n\right), 1 - \pi\left(b(X) = 1|Z_1,\ldots,Z_n\right)\right\}\right] d\tilde{P}^n \\ = \quad &\gamma \int_{Z_1,\ldots,Z_n} \sum_{j=1}^{v-1} p\left[\min\left\{\pi\left(b_j = 1|Z_1,\ldots,Z_n\right), 1 - \pi(b_j = 1|Z_1,\ldots,Z_n)\right\}\right] d\tilde{P}^n,\end{aligned} \tag{D.16}$$

where $\tilde{P}^n$ is the marginal likelihood of $\{(Y_i, D_i, X_i) : i = 1,\ldots,n\}$ corresponding to prior $\pi(\mathbf{b})$. For each $j = 1,\ldots,(v-1)$ let

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

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

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

Hence,

$$\begin{aligned}
& \min\left\{\pi\left(b_j=1|Z_1,\ldots,Z_n\right), 1-\pi(b_j=1|Z_1,\ldots,Z_n)\right\} \\
= & \frac{\min\left\{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-}, \left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}\right\}}{\left(\frac{1+\gamma}{2}\right)^{k_j^+}\left(\frac{1-\gamma}{2}\right)^{k_j^-}+\left(\frac{1+\gamma}{2}\right)^{k_j^-}\left(\frac{1-\gamma}{2}\right)^{k_j^+}} \\
= & \frac{\min\left\{1, (\frac{1+\gamma}{1-\gamma})^{k_j^+-k_j^-}\right\}}{1+(\frac{1+\gamma}{1-\gamma})^{k_j^+-k_j^-}} \\
= & \frac{1}{1+a^{|k_j^+-k_j^-|}}, \text{ where } a=\frac{1+\gamma}{1-\gamma}>1. \qquad \text{(D.17)}
\end{aligned}$$

Coarsen an observation of (Y_i, D_i) into $\tilde{Y}_i$ defined as

$$\tilde{Y}_i = \begin{cases} 1 & \text{if } Y_iD_i+(1-Y_i)(1-D_i)=1, \\ -1 & \text{otherwise.} \end{cases} \qquad \text{(D.18)}$$

Since $k_j^+ - k_j^- = \sum_{i:X_i=x_j}\tilde{Y}_i$, plugging (D.17) into (D.16) yields

$$\gamma\sum_{j=1}^{v-1} pE_{\tilde{P}^n}\left[\frac{1}{1+a^{\left|\sum_{i:X_i=x_j}\tilde{Y}_i\right|}}\right] \geq \frac{\gamma}{2}\sum_{j=1}^{v-1} pE_{\tilde{P}^n}\left[\frac{1}{a^{\left|\sum_{i:X_i=x_j}\tilde{Y}_i\right|}}\right] \geq \frac{\gamma}{2}p\sum_{i=1}^{v-1} a^{-E_{\tilde{P}^n}\left|\sum_{i:X_i=x_j}\tilde{Y}_i\right|},$$

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

$$E_{\tilde{P}^n}\left|\sum\nolimits_{i:X_i=x_j}\tilde{Y}_i\right| = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}E\left|2B(k,\frac{1}{2})-k\right|$$

holds, where $B(k, \frac{1}{2})$ is a random variable following the binomial distribution 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

$$E_{\tilde{P}^n}\left|\sum\nolimits_{i:X_i=x_j}\tilde{Y}_i\right| \leq \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\sqrt{k}$$

$$
\begin{aligned}
&= E\sqrt{B\left(n,p\right)} \\
&\leq \sqrt{np}. \quad (\because \text{ Jensen's inequality}).
\end{aligned}
$$

Hence, the Bayes risk (D.16) is bounded from below by

$$
\begin{aligned}
&\frac{\gamma}{2}p(v-1)a^{-\sqrt{np}} \\
\geq\ &\frac{\gamma}{2}p(v-1)e^{-(a-1)\sqrt{np}} \quad (\because 1+x\leq e^{x}\ \forall x) \\
=\ &\frac{p\gamma}{2}(v-1)e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}, \qquad \text{(D.19)}
\end{aligned}
$$

therefore

$$
\int_{\mathbf{b}} E_{P_{\mathbf{b}}^{n}}\left[\mu^{*}-\mu_{\widehat{G}}\right]d\pi(\mathbf{b}) \geq \frac{p\gamma}{2}(v-1)e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}. \qquad \text{(D.20)}
$$

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 $\gamma=\sqrt{\frac{v-1}{n\tau^{*}}}$. Then for all $n\geq 4(v-1)/\tau^{*}$, $\gamma\leq 1/2$ and since $p=\frac{\tau^{*}}{v-1}$,

$$
-\frac{2\gamma}{1-\gamma}\sqrt{np}=-\frac{2}{1-\gamma}\sqrt{\frac{v-1}{n\tau^{*}}}\sqrt{\frac{n\tau^{*}}{v-1}}=-\frac{2}{1-\gamma}\geq -4.
$$

Then

$$
\frac{p\gamma}{2}(v-1)e^{-\frac{2\gamma}{1-\gamma}\sqrt{np}}\geq\frac{p\gamma}{2}(v-1)e^{-4}=\frac{\tau^{*}}{2}\sqrt{\frac{v-1}{n\tau^{*}}}e^{-4}=\frac{e^{-4}}{2}\sqrt{\tau^{*}}\sqrt{\frac{v-1}{n}}.
$$

Inserting this bound into (D.20) and multiplying by Υ provides a lower bound for (D.15). This completes the proof. □

Proof of Proposition 4.1. Similarly to inequality (A.15) shown in Appendix A, the average welfare regret of the capacity-constrained estimated policy satisfies

$$
E_{P^{n}}\left[\sup_{G\in\mathcal{G}}W_{\Lambda}^{K}(G)-W_{\Lambda}^{K}(\widehat{G}^{K})\right]\leq 2|\Lambda'(0)|\int_{0}^{\infty}E_{P^{n}}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}_{G}^{K}(y)-F_{G}^{K}(y)\right|\right]dy. \qquad \text{(D.21)}
$$

We hence focus on bounding $E_{P^{n}}\left[\sup_{G\in\mathcal{G}}\left|\widehat{F}_{G}^{K}(y)-F_{G}^{K}(y)\right|\right]$.

Expressing $\widehat{F}_{G}^{K}(y)$ and $F_{G}^{K}(y)$ as

$$
\widehat{F}_{G}^{K}(y)=1-\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_{G,i}^{K}\cdot 1\{Y_{i}>y\},
$$

$$F_G^K(y) = 1 - E_P[w_G^K(Z) \cdot 1\{Y > y\}],$$

where

$$w_G^K(Z) = \frac{1-D}{1-e(X)} + \min\left\{1, \frac{K}{P_X(G)}\right\} \tilde{w}_G(Z),$$
$$\widehat{w}_{G,i}^K = \frac{1-D_i}{1-e(X_i)} + \min\left\{1, \frac{K}{P_{X,n}(G)}\right\} \tilde{w}_G(Z_i),$$

where $\tilde{w}_G(Z_i) = \left[\frac{D_i}{e(X_i)} - \frac{1-D_i}{1-e(X_i)}\right] \cdot 1\{X_i \in G\}$. Note that $\|\tilde{w}_G\|_\infty \leq \kappa^{-1}$. Define

$$\tilde{F}_G^K(y) = 1 - \frac{1}{n}\sum_{i=1}^n w_G^K(Z_i) \cdot 1\{Y_i > y\}.$$

We consider

$$\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^K(y) - F_G^K(y)\right| \leq \underbrace{\sup_{G\in\mathcal{G}} \left|\widehat{F}_G^K(y) - \tilde{F}_G^K(y)\right|}_{\text{(iv)}} + \underbrace{\sup_{G\in\mathcal{G}} \left|\tilde{F}_G^K(y) - F_G^K(y)\right|}_{\text{(v)}}, \tag{D.22}$$

and derive bounds for $\int_0^\infty E_{P^n}[(\text{iv})]dy$ and $\int_0^\infty E_{P^n}[(\text{v})]dy$.

For term (iv), we have

$$\begin{aligned}
\left|\widehat{F}_G^K(y) - \tilde{F}_G^K(y)\right| &\leq \frac{1}{n}\sum_{i=1}^n \left|\widehat{w}_{G,i}^K - w_G^K(Z_i)\right| \cdot 1\{Y_i > y\} \\
&\leq \kappa^{-1} \left|\frac{K}{\max\{K, P_{X,n}(G)\}} - \frac{K}{\max\{K, P_X(G)\}}\right| \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\} \\
&\leq \frac{1}{\kappa K} \left|P_{X,n}(G) - P_X(G)\right| \cdot \frac{1}{n}\sum_{i=1}^n 1\{Y_i > y\}. \\
&= \frac{1}{\kappa K} \left|P_{X,n}(G) - P_X(G)\right| \cdot P(Y > y) \\
&\quad + \frac{1}{\kappa K} \left|P_{X,n}(G) - P_X(G)\right| \cdot \frac{1}{n}\sum_{i=1}^n \left[1\{Y_i > y\} - P(Y > y)\right]. \quad \text{(D.23)}
\end{aligned}$$

Note that by Lemma A.4, $E_{P^n}\left[\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)|\right] \leq C_1\sqrt{v/n}$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&E_{P^n}\left[\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)| \cdot \frac{1}{n}\sum_{i=1}^n \left[1\{Y_i > y\} - P(Y > y)\right]\right] \\
\leq &\sqrt{E_{P^n}\left[\left(\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)|\right)^2\right]} \cdot \sqrt{\frac{P(Y > y)(1 - P(Y > y))}{n}}
\end{aligned}$$

$$\leq \sqrt{\frac{P(Y>y)}{n}}, \tag{D.24}$$

where the second inequality follows from

$$\sup_{G\in\mathcal{G}} |P_{X,n}(G) - P_X(G)|^2 \leq \frac{1}{n}\sum_{i=1}^{n}(1\{X_i \in G\} - P_X(G))^2 \leq 1.$$

Hence, by noting $1 \leq \sqrt{v}$,

$$\int_0^\infty E_{P^n}[(\text{iv})] \leq \frac{2\,(C_1+1)\,\Upsilon}{\kappa K}\cdot\sqrt{\frac{v}{n}} \tag{D.25}$$

Next, consider term (v). Let $\tilde{F}^K_\emptyset(y) = 1 - n^{-1}\sum_{i=1}^n \frac{1-D_i}{1-e(X_i)}\cdot 1\{Y_i > y\}$. We decompose term (v) as follows:

$$\begin{aligned}
\left|\tilde{F}^K_G(y) - F^K_G(y)\right| &\leq \left|(\tilde{F}^K_G(y) - \tilde{F}^K_\emptyset(y)) - (F^K_G(y) - F^K_\emptyset(y))\right| + |\tilde{F}^K_\emptyset(y) - F^K_\emptyset(y)| \\
&= \min\left\{1, \frac{K}{P_X(G)}\right\}\left|\frac{1}{n}\sum_{i=1}^n \tilde{w}_G(Z_i)\cdot 1\{Y_i > y\} - E_P[\tilde{w}_G(Z)\cdot 1\{Y>y\}]\right| \\
&\quad + |\tilde{F}^K_\emptyset(y) - F^K_\emptyset(y)| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \tilde{w}_G(Z_i)\cdot 1\{Y_i > y\} - E_P[\tilde{w}_G(Z)\cdot 1\{Y>y\}]\right| + |\tilde{F}^K_\emptyset(y) - F^K_\emptyset(y)|.
\end{aligned} \tag{D.26}$$

Hence,

$$\begin{aligned}
&\int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\tilde{F}^K_G(y) - F^K_G(y)\right|\right] \\
&\leq \int_0^\infty E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \tilde{w}_G(Z_i)\cdot 1\{Y_i>y\} - E_P[\tilde{w}_G(Z)\cdot 1\{Y>y\}]\right|\right] dy \\
&\quad + \int_0^\infty E_{P^n}\left[|\tilde{F}^K_\emptyset(y) - F^K_\emptyset(y)|\right]\, dy \\
&\leq 2C_T\frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}} + \int_0^\infty E_{P^n}\left[|\tilde{F}^K_\emptyset(y) - F^K_\emptyset(y)|\right]\, dy,
\end{aligned} \tag{D.27}$$

where the second inequality follows from Lemma A.5 with $M = 2\Upsilon$ and $\bar{F} = \kappa^{-1}$, where C_T is the universal constant defined there.

To bound the second term in (D.27),

$$\int_0^\infty E_{P^n}\left[|\tilde{F}^K_\emptyset(y) - F^K_\emptyset(y)|\right] dy \leq \int_0^\infty \sqrt{Var\left(\frac{1}{n}\sum_{i=1}^n \frac{1-D_i}{1-e(X_i)}\cdot 1\{Y_{0i} > y\}\right)} dy$$

$$= \int_0^\infty \sqrt{\frac{1}{n}\left\{E_P\left[\left(\frac{1}{1-e(X)}\right)P(Y_0>y|X)\right] - P(Y_0>y)^2\right\}}$$
$$\leq \int_0^\infty \sqrt{\frac{1}{n}E_P\left[\left(\frac{1}{1-e(X)}\right)P(Y_0>y|X)\right]}$$
$$\leq \frac{1}{\kappa\sqrt{n}}\int_0^\infty \sqrt{P(Y_0>y)}dy \leq \frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}}. \tag{D.28}$$

Combining (D.21), (D.22), (D.25), (D.27), and (D.28), and noting $1 \leq \sqrt{v}$, we conclude

$$E_{P^n}\left[\sup_{G\in\mathcal{G}} W_\Lambda^K(G) - W_\Lambda^K(\widehat{G}^K)\right] \leq \left(\frac{C_{K1}}{K} + C_{K2}\right)|\Lambda'(0)|\frac{\Upsilon}{\kappa}\sqrt{\frac{v}{n}},$$

where $C_{K1} = 4(C_1+1)$ and $C_{K2} = 2(2C_T+1)$. □

Proof of Theorem C.1. For any $G \in \mathcal{G}$, it holds

$$\begin{aligned} W_\Lambda(G) - W_\Lambda(\widehat{G}^e) &\leq \widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G) - \widehat{W}_\Lambda(\widehat{G}^e) + \widehat{W}_\Lambda^e(\widehat{G}^e) \\ &\quad + W_\Lambda(G) - W_\Lambda(\widehat{G}^e) - \widehat{W}_\Lambda(G) + \widehat{W}_\Lambda(\widehat{G}^e) \\ &\leq 2\sup_{G\in\mathcal{G}}|\widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G)| + 2\sup_{G\in\mathcal{G}}\left|\widehat{W}_\Lambda(G) - W_\Lambda(G)\right|, \end{aligned} \tag{D.29}$$

where the first inequality uses $\widehat{W}_\Lambda^e(\widehat{G}^e) - \widehat{W}_\Lambda^e(G) \geq 0$. The mean of the second term in the right-hand side of (D.29) is $O(n^{-1/2})$ as shown in equation (17) of Theorem 3.1.

For the first term in the right-hand side of (D.29), following the inequalities shown in (A.14), we have

$$|\widehat{W}_\Lambda(G) - \widehat{W}_\Lambda^e(G)| \leq |\Lambda'(0)|\int_0^\infty |\widehat{F}_G(y) - \widehat{F}_G^e(y)|dy. \tag{D.30}$$

For every y, the upper bound of $|\widehat{F}_G(y) - \widehat{F}_G^e(y)|$ uniform in G can be obtained as

$$\begin{aligned} &|\widehat{F}_G(y) - \widehat{F}_G^e(y)| \\ \leq &\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| D_i 1\{Y_i > y\}1\{X_i \in G\} \\ &+ \frac{1}{n}\sum_{i=1}^n \left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right| (1-D_i)1\{Y_i > y\}1\{X_i \notin G\} \\ \leq &\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| \cdot 1\{Y_{1i} > y\} + \frac{1}{n}\sum_{i=1}^n \left|\frac{1}{1-e(X_i)} - \frac{1}{1-\hat{e}(X_i)}\right| \cdot 1\{Y_{0i} > y\} \end{aligned}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\cdot P(Y_{1i}>y)}_{\text{(vi)}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\cdot [1\{Y_{1i}>y\}-P(Y_{1i}>y)]}_{\text{(vii)}}$$
$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{1-e(X_i)}-\frac{1}{1-\hat{e}(X_i)}\right|\cdot P(Y_{0i}>y)}_{\text{(viii)}}$$
$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{1-e(X_i)}-\frac{1}{1-\hat{e}(X_i)}\right|\cdot [1\{Y_{0i}>y\}-P(Y_{0i}>y)]}_{\text{(ix)}}. \tag{D.31}$$

We derive the convergence rates of the integrated means of terms (vi) - (ix) in (D.31), separately; by Assumption C.1,

$$\int_0^\infty E_{P^n}[\text{(vi)}]dy \le E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\right]\cdot\Upsilon = O(\phi_n^{-1}).$$
$$\int_0^\infty E_{P^n}[\text{(vii)}]dy \le \int_0^\infty E_{P^n}\left[\max_{1\le i\le n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\cdot\frac{1}{n}\sum_{i=1}^{n}[1\{Y_{1i}>y\}-\ P(Y_{1i}>y)]\right]dy$$
$$\le \sqrt{E_{P^n}\left[\left(\max_{1\le i\le n}\left|\frac{1}{e(X_i)}-\frac{1}{\hat{e}(X_i)}\right|\right)^2\right]}\int_0^\infty\sqrt{\frac{P(Y_{1i}>y)(1-P(Y_{1i}>y))}{n}}dy$$
$$\le O(\tilde{\phi}_n^{-1})\cdot\frac{\Upsilon}{\sqrt{n}} = O\left(\tilde{\phi}_n^{-1}/\sqrt{n}\right).$$

Similarly, we obtain $\int_0^\infty E_{P^n}[\text{(viii)}]dy \le O(\phi_n^{-1})$ and $\int_0^\infty E_{P^n}[\text{(ix)}]dy \le O\left(\tilde{\phi}_n^{-1}/\sqrt{n}\right)$.

These convergence rates for terms (vi) - (ix) and (D.30) imply that

$$E_{P^n}\left[\sup_{G\in\mathcal{G}}\left|\widehat{W}_\Lambda(G)-\widehat{W}^e_\Lambda(G)\right|\right] = O\left(\phi_n^{-1}+\frac{\tilde{\phi}_n^{-1}}{\sqrt{n}}\right)$$

Hence, by (D.29) and noting that $\tilde{\phi}_n^{-1}n^{-1/2}$ converges faster than $n^{-1/2}$, we conclude

$$E_{P^n}\left[\sup_{G\in\mathcal{G}}W_\Lambda(G)-W_\Lambda(\widehat{G}^e)\right] \le O\left((\phi_n^{-1}+\tilde{\phi}_n^{-1}n^{-1/2})\vee n^{-1/2}\right) = O\left(\phi_n^{-1}\vee n^{-1/2}\right).$$

□

E Equality-minded EWM with Nonparametrically Estimated Propensity Score

In this appendix, we consider the equality-minded EWM approach with unknown propensity score estimated nonparametrically by local polynomial regressions. We provide regularity conditions under which the nonparametric estimator of the propensity score satisfies Assumption C.1 with an explicit characterization of ϕ_n and $\tilde{\phi}_n$.

We consider the leave-one-out local polynomial estimator for $e(\cdot)$, i.e., $\hat{e}(X_i)$ is constructed by fitting the local polynomials excluding the i-th observation. 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 $\|x\| \equiv \left(x_1^2 + \cdots + x_{d_x}^2\right)$. Let $K(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}$ be a kernel function and $h > 0$ be a bandwidth, whose dependence on the sample size is implicit in the notation. At each X_i, $i = 1, \ldots, n$, we define the leave-one-out local polynomial coefficient estimators with degree $l \geq 0$ as

$$\hat{\theta}(X_i) = \arg\min_{\theta} \sum_{j \neq i} \left[D_j - \theta^T U\left(\frac{X_j - X_i}{h}\right) \right]^2 K\left(\frac{X_j - X_i}{h}\right),$$

where $U\left(\frac{X_j - X_i}{h}\right)$ is the vector with elements indexed by the multi-index s, i.e., $U\left(\frac{X_j - X_i}{h}\right) \equiv \left(\left(\frac{X_j - X_i}{h}\right)^s\right)_{|s| \leq l}$. With a slight abuse of notation, we define $U(0) = (1, 0, \ldots, 0)^T$. Let $\lambda_n(X_i)$ be the smallest eigenvalue of $B(X_i) \equiv \left(nh^{d_x}\right)^{-1} \sum_{j \neq i} U\left(\frac{X_j - X_i}{h}\right) U^T\left(\frac{X_j - X_i}{h}\right) K\left(\frac{X_i - X_j}{h}\right)$. Accordingly, we construct the leave-one-out local polynomial fit for $e(X_i)$ by

$$\tilde{e}(X_i) = U^T(0)\hat{\theta}(X_i) \cdot 1\left\{\lambda_n(X_i) \geq t_n\right\} \tag{E.1}$$

where t_n is a positive sequence that slowly converges to zero, such as $t_n \propto (\log n)^{-1}$. This trimming constant regularizes the regressor matrix of the local polynomial regression and simplifies the proof of the uniform consistency of the local polynomial estimator.

To characterize $\mathcal{P}_e$ in Assumption C.1, we impose the following restrictions, which are identical to Assumption E.2 in Kitagawa and Tetenov (2018b).

Assumption E.1. *(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 $L_e < \infty$. [1]

(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 $\mathbb{R}^{d_x}$ and $B(x,r)$ be the open ball centered at $x \in \mathbb{R}^{d_x}$ with radius r. There exist constants $\underline{c}$ and r_0 such that

$$Leb\left(\mathcal{X} \cap B(x,r)\right) \geq \underline{c} Leb(B(x,r)) \quad \forall 0 < r \leq r_0, \forall x \in \mathcal{X}, \tag{E.2}$$

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 $K(\cdot)$ has support $[-1,1]^{d_x}$, $\int_{\mathbb{R}^{d_x}} K(u)du = 1$, and $\sup_u K(u) \leq K_{\max} < \infty$.

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.2) 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 $Leb\left(\mathcal{X} \cap B(x,r)\right)$ converges to zero as $r \to 0$ faster than the rate of r^2 for some x on the boundary of $\mathcal{X}$.

The next lemma collects several properties of the local polynomial estimators that are useful to prove the bound shown in (C.4). These claims are borrowed from Theorem 3.2 in Audibert and Tsybakov (2007) and Lemma E.4 in Kitagawa and Tetenov (2018b).

Lemma E.1. *Let $\mathcal{P}_e$ consist of the data generating processes satisfying Assumption E.1 (Smooth-e) and (PX). Let $\tilde{e}(X_i)$ be the leave-one-out estimator for the propensity score defined in (E.1) whose kernel function satisfies E.1 (Ker).*

(i) There exist positive constants c_2, c_3, and c_4 that depend only on β_e, d_x, L_e, $\underline{c}$, r_0, $\underline{p}_X$,

[1] 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 $f_x(x') \equiv \sum_{|s| \leq \beta-1} \frac{(x'-x)^s}{s!} D^s f(x)$. 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 $|f_x(x') - f(x)| \leq L \|x - x'\|^{\beta}$.

and $\bar{p}_X$*, such that, for any* $0 < h < r_0/\underline{c}$*, any* $c_4 h^{\beta_e} < \delta$*, and any* $n \geq 2$*,*

$$P^{n-1}\left(|\tilde{e}(x) - e\left(x\right)| > \delta\right) \leq c_2 \exp\left(-c_3 n h^{d_x} \delta^2\right),$$

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

$$\sup_{P \in \mathcal{P}_e} \int_{\mathcal{X}} E_{P^{n-1}}\left[|\tilde{e}(x) - e\left(x\right)|\right] dP_X(x) \leq O(h^{\beta_e}) + O\left(\frac{1}{\sqrt{n h^{d_x}}}\right)$$

holds. Hence, a choice of bandwidth that optimizes the upper bound of the convergence rate is $h \propto n^{-\frac{1}{2\beta_e + d_x}}$ *and the resulting uniform convergence rate is*

$$\sup_{P \in \mathcal{P}_e} \int_{\mathcal{X}} E_{P^{n-1}}\left[|\tilde{e}(x) - e\left(x\right)|\right] dP_X(x) \leq O\left(n^{-\frac{1}{2 + d_x/\beta_e}}\right). \tag{E.3}$$

(iii)

$$\sup_{P \in \mathcal{P}_e} E_{P^n}\left[(\max_{1 \leq i \leq n} |\tilde{e}(X_i) - e\left(X_i\right)|)^2\right] \leq O\left(\frac{h^{2\beta_e}}{t_n^2}\right) + O\left(\frac{\log n}{n h^{d_x} t_n^2}\right)$$

holds. In particular, when the bandwidth is chosen as in claim (ii) of the current proposition, the resulting uniform convergence rate is

$$\sup_{P \in \mathcal{P}_e} E_{P^n}\left[(\max_{1 \leq i \leq n} |\tilde{e}(X_i) - e\left(X_i\right)|)^2\right] \leq O\left(t_n^{-2} \log n \cdot n^{-\frac{2}{2 + d_x/\beta_e}}\right). \tag{E.4}$$

Making use of Lemma E.1, the next proposition shows a propensity score estimator constructed by suitably trimming $\tilde{e}(X_i)$ satisfies Assumption C.1 with an explicit characterization of the growing sequences ϕ_n and $\tilde{\phi}_n$.

Proposition E.1. *Let* $\mathcal{P}_e$ *consist of data generating processes that satisfy Assumption E.1 (Smooth-e) and (PX). Let* $\tilde{e}(X_i)$ *be the leave-one-out local polynomial estimator with degree* $l = (\beta_e - 1)$*, trimming sequence for the least eigenvalue* $t_n = (\log n)^{-1}$*, bandwidth sequence* $h \propto n^{-\frac{1}{2\beta_e + d_x}}$*, and whose kernel satisfies Assumption E.1 (Ker). Let*

$$\hat{e}(X_i) \equiv \min\left\{1 - \epsilon_n, \max\{\epsilon_n, \tilde{e}(X_i)\}\right\} \in [\epsilon_n, 1 - \epsilon_n] \tag{E.5}$$

with a sequence of trimming constants ϵ_n *that satisfies* $\epsilon_n = O(n^{-a})$ *for some* $a > 0$*. Then,* $\hat{e}(X_i)$ *satisfies Assumption C.1 with* $\phi_n = n^{\frac{1}{2 + d_x/\beta_e}}$ *and* $\tilde{\phi}_n = (\log n)^{-3/2} \cdot n^{\frac{1}{2 + d_x/\beta_e}}$*.*

Proof of Proposition E.1. Assume that n is large enough so that $\varepsilon_n \leq \kappa/2$ holds. Since $\hat{e}(X_i) = \tilde{e}(X_i)$ whenever $\tilde{e}(X_i) \in \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right] \subset [\epsilon_n, 1-\epsilon_n]$, the following bounds are valid

$$\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right| \leq \begin{cases} \frac{2}{\kappa^2}\left|\tilde{e}(X_i) - e(X_i)\right| & \text{if } \tilde{e}(X_i) \in \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right] \\ (\kappa\epsilon_n)^{-1} & \text{if } \tilde{e}(X_i) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]. \end{cases}$$

Hence,

$$\begin{aligned} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right] &= E_{P^n}\left[\left|\frac{1}{e(X_n)} - \frac{1}{\hat{e}(X_n)}\right|\right] \\ &\leq \frac{2}{\kappa^2}E_{P^n}|\tilde{e}(X_n) - e(X_n)| + (\kappa\epsilon_n)^{-1}P^n\left(\tilde{e}(X_n) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right) \end{aligned} \tag{E.6}$$

By Lemma E.1 (ii),

$$\sup_{P\in\mathcal{P}_e} E_{P^n}|\tilde{e}(X_n) - e(X_n)| = \sup_{P\in\mathcal{P}_e}\int_{\mathcal{X}} E_{P^{n-1}}\left[|\tilde{e}(x) - e(x)|\right] dP_X(x) \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right).$$

Furthermore, by Lemma E.1 (i),

$$\begin{aligned} P^n\left(\tilde{e}(X_n) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right) &= \int_{\mathcal{X}} P^{n-1}\left(\tilde{e}(x) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right) dP_X(x) \\ &\leq \int_{\mathcal{X}} P^{n-1}\left(|\tilde{e}(x) - e(x)| > \frac{\kappa}{2}\right) dP_X(x) \\ &\leq c_2 \exp\left(-\frac{c_3\kappa^2}{4}nh^{d_x}\right) \end{aligned} \tag{E.7}$$

holds for all n satisfying $c_4 h^{\beta_e} < \kappa/2$, where c_2, c_3, and c_4 are the constants defined in Lemma E.1 (i). Since ε_n is assumed to converge at a polynomial rate, $\epsilon_n^{-1}P^n\left(\hat{e}(X_n) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right)$ converges faster than $O(n^{-\frac{1}{2+d_x/\beta_e}})$. Thus, from (E.6), we conclude $\sup_{P\in\mathcal{P}_e} E_{P^n}\left[\frac{1}{n}\sum_{i=1}^{n}|\hat{e}(X_i) - e(X_i)|\right] \leq O\left(n^{-\frac{1}{2+d_x/\beta_e}}\right)$, i.e., $\phi_n = n^{\frac{1}{2+d_x/\beta_e}}$.

For the bounds for the mean of the squared maximum, we have

$$\begin{aligned} &E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right)^2\right] \\ \leq &\frac{4}{\kappa^4}E_{P^n}\left[\left(\max_{1\leq i\leq n}|\tilde{e}(X_n) - e(X_n)|\right)^2\right] + (\kappa\epsilon_n)^{-2}P^n\left(\tilde{e}(X_n) \notin \left[\frac{\kappa}{2}, 1-\frac{\kappa}{2}\right]\right) \end{aligned}$$

By Lemma E.1 (iii) and (E.7), $E_{P^n}\left[\left(\max_{1\leq i\leq n}\left|\frac{1}{e(X_i)} - \frac{1}{\hat{e}(X_i)}\right|\right)^2\right] \leq O\left((\log n)^3 \cdot n^{-\frac{2}{2+d_x/\beta_e}}\right)$, i.e., $\tilde{\phi}_n = (\log n)^{-3/2} \cdot n^{\frac{1}{2+d_x/\beta_e}}$.

The other convergence rate bounds in Assumption C.1 can be shown similarly. □

Combining Proposition E.1 with Theorem C.1 proves the claim made in equation (C.4).

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.

Cowell, F. (1995): *Measuring Inequality*, Hemel Hempstead: Harvester Wheatsheaf, 2 ed.

Kasy, M. (2016): "Partial Identification, Distributional Preferences, and the Welfare Ranking of Policies," *Review of Economics and Statistics*, 98, 111–131.

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

——— (2018b): "Online Appendix to "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice"," URL: https://www.econometricsociety.org/sites/default/files/13288_Data_and_Programs.zip.

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

Wasserman, L. (2006): *All of Nonparametric Statistics*, Springer.