SUPPLEMENT TO "MODEL SELECTION FOR TREATMENT CHOICE: PENALIZED WELFARE MAXIMIZATION"

(*Econometrica*, Vol. 89, No. 2, March 2021, 825–848)

ERIC MBAKOP
Department of Economics, University of Calgary

MAX TABORD-MEEHAN
Department of Economics, University of Chicago

APPENDIX A: PROOFS OF MAIN RESULTS

RECALL THAT the planner's objective function is given by

$$W(G) = E_P\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)\cdot \mathbf{1}\{X \in G\}\right]. \tag{A.1}$$

To each treatment allocation $G \in \mathcal{G}$ we associate a function $f_G : \mathbb{R} \times \mathcal{X} \times \{0,1\} \to \mathbb{R}$ defined by

$$f_G(Z) = f_G(Y, X, D) = \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)\cdot \mathbf{1}\{X \in G\},$$

where $Z = (Y, X, D)$. Let $\mathcal{F} := \{f_G : G \in \mathcal{G}\}$ denote the corresponding set of functions associated to decision rules in $\mathcal{G}$. By (A.1), any optimal allocation in $\mathcal{G}$ solves

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

Equivalently, functions associated to optimal allocations solve

$$f^* \in \arg\max_{f\in\mathcal{F}} E_P f(Z).$$

By an abuse of notation, for $G \in \mathcal{G}$, we set

$$W(f_G) = E_P f_G(Z).$$

Given an approximating sequence $\{\mathcal{G}_k\}_k$ of classes of treatment allocations, let $\{\mathcal{F}_k\}_k$ denote the sequence of associated classes of functions.

The following lemma, whose proof is given in Kitagawa and Tetenov (2018) (Lemma A.1), establishes the relevant link between the classes of sets $\{\mathcal{G}_k\}_k$ and the classes of functions $\{\mathcal{F}_k\}_k$. It shows that if a class $\mathcal{G}$ has finite VC dimension, then the associated class $\mathcal{F}$ is a VC-subgraph class with dimension bounded above by that of $\mathcal{G}$.

LEMMA A.1: *Let $\mathcal{G}$ be a VC class of subsets of $\mathcal{X}$ with finite VC dimension V. Let g be a function from $\mathcal{Z} := \mathbb{R} \times \mathcal{X} \times \{0,1\}$ to $\mathbb{R}$. Then the set of functions $\mathcal{F}$ defined by*

$$\mathcal{F} = \left\{g(z)\cdot \mathbf{1}\{x \in G\} : G \in \mathcal{G}\right\}$$

is a VC-subgraph class with dimension at most V.

Eric Mbakop: eric.mbakop@ucalgary.ca
Max Tabord-Meehan: maxtm@uchicago.edu

 https://doi.org/10.3982/ECTA16437

For each $k \geq 1$, let $\hat{f}_{n,k}$ be a maximizer of the empirical welfare over the class $\mathcal{F}_k$; that is,

$$\hat{f}_{n,k} = \arg\max_{f \in \mathcal{F}_k} W_n(f),$$

and for $f \in \mathcal{F}_k$, define the complexity-penalized estimate of welfare by

$$R_{n,k}(f) = W_n(f) - C_n(k) - \sqrt{\frac{k}{n}}.$$

The PWM rule $\hat{f}_{n,\hat{k}}$ is then chosen such that

$$\hat{k} = \arg\max_{k \geq 1} R_{n,k}(\hat{f}_{n,k}).$$

In what follows, we set $\hat{f}_n := \hat{f}_{n,\hat{k}}$ and $R_n(\hat{f}_n) := R_{n,\hat{k}}(\hat{f}_{n,\hat{k}})$.

To bound the regret, we decompose it as follows:

$$W^*_{\mathcal{F}} - W(\hat{f}_n) = \big(W^*_{\mathcal{F}} - R_n(\hat{f}_n)\big) + \big(R_n(\hat{f}_n) - W(\hat{f}_n)\big). \tag{A.2}$$

The following lemma yields (under Assumption 3.4) a sub-Gaussian tail bound for the second term on the right-hand side of the preceding equality.

LEMMA A.2: *Given Assumption* 3.4, *there exists a positive constant* Δ (*that does not depend on* n) *such that*

$$P\big(R_n(\hat{f}_n) - W(\hat{f}_n) > \epsilon\big) \leq \Delta e^{-2c_o n\epsilon^2}$$

for every n.

PROOF: First, note that

$$P\big(R_n(\hat{f}_n) - W(\hat{f}_n) > \epsilon\big) \leq P\Big(\sup_k \big(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k})\big) > \epsilon\Big);$$

then by the union bound,

$$P\Big(\sup_k \big(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k})\big) > \epsilon\Big) \leq \sum_k P\big(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) > \epsilon\big).$$

Now by definition of $R_{n,k}$, we have

$$\sum_k P\big(R_{n,k}(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) > \epsilon\big) = \sum_k P\left(W_n(\hat{f}_{n,k}) - C_n(k) - W(\hat{f}_{n,k}) > \epsilon + \sqrt{\frac{k}{n}}\right).$$

By Assumption 3.4,

$$\sum_k P\left(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon + \sqrt{\frac{k}{n}}\right) \leq \sum_k c_1 e^{-2c_o n(\epsilon + \sqrt{\frac{k}{n}})^2} \leq e^{-2c_o n\epsilon^2} \sum_k c_1 e^{-2kc_o}.$$

By setting

$$\Delta := \sum_k c_1 e^{-2kc_o} < \infty, \tag{A.3}$$

the result follows. *Q.E.D.*

PROOF OF THEOREM 3.1: We follow the general strategy from Bartlett, Boucheron, and Lugosi (2002). For every k, we have

$$W^*_{\mathcal{F}} - W(\hat{f}_n) = \big(W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k}\big) + \big(W^*_{\mathcal{F}_k} - W(\hat{f}_n)\big). \tag{A.4}$$

We first consider the second term in (A.4), and expand it as follows:

$$W^*_{\mathcal{F}_k} - W(\hat{f}_n) = \big(W^*_{\mathcal{F}_k} - R_n(\hat{f}_n)\big) + \big(R_n(\hat{f}_n) - W(\hat{f}_n)\big). \tag{A.5}$$

By the definition of R_n, the first term of expression (A.5) is bounded by

$$W^*_{\mathcal{F}_k} - R_n(\hat{f}_n) \le W^*_{\mathcal{F}_k} - W_n(\hat{f}_{n,k}) + C_n(k) + \sqrt{\frac{k}{n}}.$$

Fix $\delta > 0$, and choose some $f^*_k \in \mathcal{F}_k$ such that $W(f^*_k) + \delta \ge W^*_{\mathcal{F}_k}$.[1] We have

$$W^*_{\mathcal{F}_k} - W_n(\hat{f}_{n,k}) + C_n(k) + \sqrt{\frac{k}{n}} \le W\big(f^*_k\big) + \delta - W_n\big(f^*_k\big) + C_n(k) + \sqrt{\frac{k}{n}}.$$

Taking expectations of both sides and letting δ converge to 0 yields

$$E\big[W^*_{\mathcal{F}_k} - R_n(\hat{f}_n)\big] \le E\big[C_n(k)\big] + \sqrt{\frac{k}{n}}.$$

By Lemma A.2 and a standard integration argument (see, for instance, problem 12.1 in Györfi, Devroye, and Lugosi (1996)), the second term on the right-hand side of (A.5) is bounded by

$$E\big[R_n(\hat{f}_n) - W(\hat{f}_n)\big] \le \sqrt{\frac{\log(\Delta e)}{2c_o n}}.$$

Combining these bounds yields

$$E\big[W^*_{\mathcal{F}} - W(\hat{f}_n)\big] \le E\big[C_n(k)\big] + W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k} + \sqrt{\frac{\log(\Delta e)}{2c_o n}} + \sqrt{\frac{k}{n}},$$

for every k, and our result follows. *Q.E.D.*

[1] If the welfare criterion achieves its maximum on $\mathcal{F}_k$, then f^*_k can be set equal to any maximizer. In general, however, such an optimum may not exist, and thus we must choose f^*_k to be an "almost maximizer" of the welfare criterion on $\mathcal{F}_k$.

PROOF OF LEMMA 3.2: We first establish the inequality

$$P\big(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon\big) \leq \exp\left(-2n\epsilon^2\left(\frac{\kappa}{3M}\right)^2\right). \tag{A.6}$$

By two standard symmetrization arguments, we get

$$E\Big[\sup_{f\in\mathcal{F}_k} W_n(f) - W(f)\Big] \leq 2E\left[\sup_{f\in\mathcal{F}_k} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right] = E\big[C_n(k)\big], \tag{A.7}$$

where we recall that $C_n(k) = E[2\sup_{f\in\mathcal{F}_k} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)|Z_1, Z_2, \ldots, Z_n]$ and $\{\sigma_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher random variables independent from the data $\{Z_i\}_{i=1}^n$. Note that

$$P\big(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon\big) \leq P(\sup_{f\in\mathcal{F}_k}\big(\big(W_n(f) - W(f)\big) - C_n(k) > \epsilon\big),$$

and set $M_{n,k} := \sup_{f\in\mathcal{F}_k}(W_n(f) - W(f)) - C_n(k)$. Combining the preceding inequality with (A.7) yields

$$P\big(W_n(\hat{f}_{n,k}) - W(\hat{f}_{n,k}) - C_n(k) > \epsilon\big) \leq P(M_{n,k} - EM_{n,k} > \epsilon).$$

To control the deviations of $M_{n,k}$ from its mean, we use McDiarmid's inequality (see Györfi, Devroye, and Lugosi (1996, Theorem 9.2); note that $M_{n,k}$ satisfies the bounded difference property with increments bounded by $\frac{3M}{n\kappa}$) which yields the inequality

$$P(M_{n,k} - EM_{n,k} > \epsilon) \leq \exp\left(-2n\epsilon^2\left(\frac{\kappa}{3M}\right)^2\right),$$

from which our result follows.

The second inequality (where C is a universal constant)

$$E\big[C_n(k)\big] \leq C\frac{M}{\kappa}\sqrt{\frac{V_k}{n}}$$

follows from a chaining argument and a control on the universal entropy of VC-subgraph classes (see, for instance, the proof of Lemma A.4 in Kitagawa and Tetenov (2018)), along with Lemma A.1. *Q.E.D.*

PROOF OF LEMMA 3.1: Let us assume for notational simplicity that the quantity $m = n(1-\ell)$ is an integer. We first establish the inequality

$$P\big(W_m(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) - C_m(k) > \epsilon\big) \leq \exp\left(-2n\ell\epsilon^2\left(\frac{\kappa}{M}\right)^2\right). \tag{A.8}$$

By the definition of $C_m(k)$, we have

$$P\big(W(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) - C_m(k) > \epsilon\big) = P\big(W_r(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) > \epsilon\big).$$

Now, working conditionally on $\{Z_i\}_{i=1}^m$, we get by Hoeffding's inequality that

$$P\big(W_r(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) > \epsilon | \{Z_i\}_{i=1}^m\big) \leq \exp\left(-2n\ell\epsilon^2\left(\frac{\kappa}{M}\right)^2\right).$$

Since the right-hand side of the preceding inequality is non-random, the inequality holds unconditionally as well.

We now establish the inequality

$$E\big[C_m(k)\big] \leq C\frac{M}{\kappa\sqrt{(1-\ell)}}\sqrt{\frac{V_k}{n}}.$$

By the definition of $C_m(k)$, we have

$$E\big[C_m(k)\big] = E\big[W_m(\hat{f}_{m,k}) - W_r(\hat{f}_{m,k})\big] = E\big[W_m(\hat{f}_{m,k}) - W(\hat{f}_{m,k}) + W(\hat{f}_{m,k}) - W_r(\hat{f}_{m,k})\big].$$

Note that by the law of iterated expectations, we have

$$E\big[W(\hat{f}_{m,k}) - W_r(\hat{f}_{m,k})\big] = 0,$$

and by Lemma A.4 in Kitagawa and Tetenov (2018) combined with Lemma A.1 there exists some universal constant C such that

$$E\big[W_m(\hat{f}_{m,k}) - W(\hat{f}_{m,k})\big] \leq C\frac{M}{\kappa}\sqrt{\frac{V_k}{m}}.$$

Since $m = (1-\ell)n$, the result follows. *Q.E.D.*

PROOF OF PROPOSITIONS 3.2 AND 3.1: From the inequality

$$\frac{e^{-x}}{\big(1-e^{-x}\big)} \leq \frac{1}{x},$$

and from (A.3) and (A.6), we derive that

$$\Delta \leq 1/2\left(\frac{3M}{\kappa}\right)^2.$$

Similarly, we derive from (A.3) and (A.8) that

$$\Delta \leq 1/(2l)\left(\frac{M}{\kappa}\right)^2.$$

The results then follow by substituting these into the inequalities of Theorem 3.1. *Q.E.D.*

PROOF OF THEOREM 3.2: Our strategy here is to proceed analogously to the proof of Theorem 3.1 with some additional machinery. Let $\hat{f}_n^e$ and $R_n^e(\cdot)$ be defined analogously to the case when the propensity score is known. For every k, we have that

$$W_{\mathcal{F}}^* - W\big(\hat{f}_n^e\big) = \big(W_{\mathcal{F}}^* - W_{\mathcal{F}_k}^*\big) + \big(W_{\mathcal{F}_k}^* - W\big(\hat{f}_n^e\big)\big). \tag{A.9}$$

Adding and subtracting $R_n^e(\hat{f}_n^e)$ to the last term yields

$$W_{\mathcal{F}_k}^* - W(\hat{f}_n^e) = \big(W_{\mathcal{F}_k}^* - R_n^e(\hat{f}_n^e)\big) + \big(R_n^e(\hat{f}_n^e) - W(\hat{f}_n^e)\big). \tag{A.10}$$

Let $f_k^* := \arg\max_{f\in\mathcal{F}_k} W(f)$ (if the supremum is not achieved, apply the argument to a δ-maximizer of the welfare, and let δ tend to zero). Now consider the first term on the right-hand side of (A.10). Expanding yet again gives

$$W_{\mathcal{F}_k}^* - R_n^e(\hat{f}_n^e) = W_{\mathcal{F}_k}^* - W_n(f_k^*) + W_n(f_k^*) - R_n^e(\hat{f}_n^e). \tag{A.11}$$

From the definition of R_n^e, we have

$$W_n(f_k^*) - R_n^e(\hat{f}_n^e) \le W_n(f_k^*) - W_n^e(f_k^*) + C_n^e(k) + \sqrt{\frac{k}{n}} \le \frac{1}{n}\sum_{i=1}^n |\hat{\tau}_i - \tau_i| + C_n^e(k) + \sqrt{\frac{k}{n}}.$$

Hence, considering the above inequality and taking expectations in (A.11) yields

$$E\big[W_{\mathcal{F}_k}^* - R_n^e(\hat{f}_n^e))\big] \le E\left[\frac{1}{n}\sum_{i=1}^n |\hat{\tau}_i - \tau_i|\right] + E\big[C_n^e(k)\big] + \sqrt{\frac{k}{n}},$$

and thus by Assumption 3.7,

$$E\big[W_{\mathcal{F}_k}^* - R_n^e(\hat{f}_n^e))\big] \le O\big(\phi_n^{-1}\big) + E\big[C_n^e(k)\big] + \sqrt{\frac{k}{n}}. \tag{A.12}$$

We now consider the second term on the right-hand side of (A.10). Let $\hat{k}$ be the class k such that

$$\hat{f}_n^e = \hat{f}_{n,\hat{k}}^e.$$

Note that $\hat{k}$ is random. We have

$$R_n^e(\hat{f}_n^e) - W(\hat{f}_n^e) = W_n^e(\hat{f}_{n,\hat{k}}^e) - C_n^e(\hat{k}) - \sqrt{\frac{\hat{k}}{n}} - W(\hat{f}_{n,\hat{k}}^e).$$

By adding and subtracting $W_n(\hat{f}_{n,\hat{k}}^e)$ and the function $\tilde{C}_n(\hat{k})$, we get

$$\begin{aligned} &W_n^e(\hat{f}_{n,\hat{k}}^e) - C_n^e(\hat{k}) - \sqrt{\frac{\hat{k}}{n}} - W(\hat{f}_{n,\hat{k}}^e) \\ &\quad = \big(W_n^e(\hat{f}_{n,\hat{k}}^e) - W_n(\hat{f}_{n,\hat{k}}^e)\big) + \big(\tilde{C}_n(\hat{k}) - C_n^e(\hat{k})\big) \\ &\qquad + \left(W_n(\hat{f}_{n,\hat{k}}^e) - W(\hat{f}_{n,\hat{k}}^e) - \tilde{C}_n(\hat{k}) - \sqrt{\frac{\hat{k}}{n}}\right). \end{aligned} \tag{A.13}$$

Note again that

$$\sup_k \big(W_n^e(\hat{f}_{n,k}^e) - W_n(\hat{f}_{n,k}^e)\big) \le \frac{1}{n}\sum_{i=1}^n |\hat{\tau}_i - \tau_i|,$$

and so by Assumptions 3.7 and 3.8, the first two terms of (A.13) are of order $O(\phi_n^{-1})$ in expectation. By the first part of Assumption 3.8, and an argument similar to the one used in the proof of Lemma A.2, it can be shown that

$$E\left[\sup_k\left(W_n\big(\hat{f}^e_{n,k}\big) - W\big(\hat{f}^e_{n,k}\big) - \tilde{C}_n(k) - \sqrt{\frac{k}{n}}\right)\right] \le \sqrt{\frac{\log(\Delta e)}{2c_0 n}},$$

where Δ and c_o are the same constants that appear in Lemma A.2. We thus get

$$E\big[R^e_n\big(\hat{f}^e_n\big) - W\big(\hat{f}^e_n\big)\big] \le O\big(\phi_n^{-1}\big) + \sqrt{\frac{\log(\Delta e)}{2m}}. \tag{A.14}$$

Now combining (A.12) and (A.14), we conclude that

$$E\big[W^*_{\mathcal{F}_k} - W\big(\hat{f}^e_n\big)\big] \le O\big(\phi_n^{-1}\big) + E\big[C^e_n(k)\big] + \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(\Delta e)}{2m}}.$$

Finally, by Assumption 3.8, we get

$$E\big[W^*_{\mathcal{F}} - W\big(\hat{f}^e_n\big)\big] \le O\big(\phi_n^{-1}\big) + E\big[\tilde{C}_n(k)\big] + W^*_{\mathcal{F}} - W^*_{\mathcal{F}_k} + \sqrt{\frac{k}{n}} + \sqrt{\frac{\log(\Delta e)}{2m}},$$

for all k, and hence the result follows. *Q.E.D.*

PROOF OF LEMMA 3.3: In what follows, we verify that the third condition of Assumption 3.8 is satisfied for the holdout penalty with estimated propensity score, as the first two conditions follow from previous arguments. Set

$$\tilde{C}_m(k) = W_m\big(\hat{f}^e_{m,k}\big) - W_r\big(\hat{f}^e_{m,k}\big).$$

Note that since the propensity score is unknown, the empirical welfare criteria W_m and W_r are infeasible. It can easily be shown that for this choice of $\tilde{C}_m(k)$, we have

$$\big|\tilde{C}_m(k) - C^e_m(k)\big| \le \frac{1}{m}\sum_{i=1}^{m}\big|\hat{\tau}_i^{\,E} - \tau_i\big| + \frac{1}{r}\sum_{i=m+1}^{n}\big|\hat{\tau}_i^{\,T} - \tau_i\big|,$$

which yields

$$E\sup_{k\ge 1}\big|\tilde{C}_m(k) - C^e_m(k)\big| = O\big(\phi_n^{-1}\big). \qquad \textit{Q.E.D.}$$

PROOF OF PROPOSITION 4.1: Let $\mathcal{G}$ be the set of monotone allocations. Let π_k denote the partition of $[0,1]$ formed by the points $x_i = i/2^k$, $i = 0,\dots,2^k$. Let $\{\mathcal{G}_k\}_k$ be the approximating sequence defined in Example 3.2, and define $G^* \in \mathcal{G}$ to be a set such that $W(G^*) = W^*_{\mathcal{G}}$ (if no such G^* exists, the argument proceeds by considering an "almost maximizer"). By definition, for each $G \in \mathcal{G}$, there is an associated function $b_G : [0,1] \to [0,1]$ which determines the boundary of the allocation region, that is, such that $G = \{(x_1, x_2) \in \mathcal{X} : x_2 \le b_G(x_1)\}$.

Fix some $P \in \mathcal{P}_r$, where $\mathcal{P}_r$ is as defined in Assumption 4.1. By definition,

$$W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \le W\big(G^*\big) - W(\tilde{G}_k),$$

where $\tilde{G}_k \in \mathcal{G}_k$ is the allocation such that $b_{\tilde{G}_k}(\cdot)$ is the linear interpolation of b_{G^*} on the partition π_k. We can rewrite this as

$$W(G^*) - W(\tilde{G}_k) = E\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right) \cdot \left(\mathbf{1}\{X \in G^*\} - \mathbf{1}\{X \in \tilde{G}_k\}\right)\right]$$
$$\leq \frac{M}{\kappa} P_X\left(G^* \Delta \tilde{G}_k\right), \tag{A.15}$$

where Δ denotes the symmetric difference operator, $A \Delta B := A \backslash B \cup B \backslash A$. Let

$$M_i = [x_{i-1}, x_i] \times \left[b_{G^*}(x_{i-1}), b_{G^*}(x_i)\right],$$

for $i = 1, \ldots, 2^k$. It follows from the monotonicity of b_{G^*} that the graphs of the restrictions of $b_{G^*}(\cdot)$ and $b_{\tilde{G}_k}(\cdot)$ to $[x_{i-1}, x_i]$ are contained in M_i. Hence we have that

$$P_X\left(G^* \Delta \tilde{G}_k\right) \leq \sum_{i=1}^{2^k} P_X(M_i) = \sum_{i=1}^{2^k} P_X(M_{1i} \times M_{2i}),$$

where $M_{1i} = [x_{i-1}, x_i]$, $M_{2i} = [b_{G^*}(x_{i-1}), b_{G^*}(x_i)]$. By Assumption 4.1,

$$P_X(M_{1i} \times M_{2i}) = \int_{M_{2i}} P_{X_1|x_2}(M_{1i})\, dP_{X_2} \leq \frac{1}{2^k} A P_{X_2}(M_{2_i}).$$

Summing over i:

$$\sum_{i=1}^{2^k} P_X(M_i) \leq \sum_{i=1}^{2^k} \frac{1}{2^k} A P_{X_2}(M_{2i}) \leq \frac{A}{2^k},$$

since the $\{M_{2i}\}_i$ form a partition of $[0, 1]$. We thus obtain that

$$W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k} \leq A \frac{M}{\kappa} 2^{-k},$$

as desired. *Q.E.D.*

APPENDIX B: ADDITIONAL RESULTS

B.1. *Supplement to Remark 3.5*

In this subsection, we provide some simple calculations that justify the comments made in Remark 3.5. Consider first the Rademacher penalty; then Proposition 3.1 shows that

$$E_{P^n}\left[W^*_{\mathcal{G}} - W(\hat{G}_n)\right] \leq \inf_k \left[C \frac{M}{\kappa} \sqrt{\frac{V_k}{n}} + \left(W^*_{\mathcal{G}} - W^*_{\mathcal{G}_k}\right) + \sqrt{\frac{k}{n}}\right] + g(M, \kappa) \frac{M}{\kappa} \sqrt{\frac{1}{n}},$$

where C is the universal constant derived in the bound of EWM in Kitagawa and Tetenov (2018) and g is defined as

$$g(M, \kappa) := 6 \sqrt{\log\left(\frac{3\sqrt{e}}{\sqrt{2}} \frac{M}{\kappa}\right)}.$$

Our first task is to quantify the size of C. By the proof of Lemma A.4 in Kitagawa and Tetenov (2018), we can see that the constant C depends on a universal constant K derived in Theorem 2.6.7 of Van Der Vaart and Wellner (1996), which establishes a bound on the covering numbers of a VC-subgraph class. Inspection of the proof in Van Der Vaart and Wellner (1996) allows us to conclude that a suitable K is given by $K = 3\sqrt{e}/8$. Plugging this into the expression for C derived in Kitagawa and Tetenov (2018) allows us to conclude that a suitable C is given by $C = 36.17$. Turning to $g(M, \kappa)$, we can calculate that in order for it to surpass C by an order of magnitude, we would need M/κ to be about as large as 10^{120}. This gives us a sense of the relative sizes of the terms in our bound.

B.2. *Supplement to Remark 4.1*

In this subsection, we perform a sample splitting exercise to estimate the welfare performance of various decision rules on the JTPA data. To estimate welfare, we split the data into two halves. The first half of the data (the "estimating sample") is used to compute various decision rules. The second half of the data (the "auxiliary sample") is used to estimate the welfare generated by each resulting treatment allocation.

Given a sample of size n and a treatment allocation G, we estimate welfare using

$$\widehat{W}(G) = E_n\left[\frac{Y_i D_i}{e(X_i)}\mathbf{1}\{X_i \in G\} + \frac{Y_i(1-D_i)}{1-e(X_i)}\mathbf{1}\{X_i \notin G\}\right],$$

where $E_n(\cdot)$ is the sample average. We study the welfare performance of three decision rules: EWM on the class $\mathcal{G}_5$ as described in Section 4, PWM with the holdout penalty on the sieve $\{\mathcal{G}_k\}_{k=1}^5$ as described in Section 4, and a "random baseline" which randomly assigns the same fraction of the population as PWM to job training. In Table I, we report the estimated welfare computed on both the estimating and auxiliary samples.

In Table I, we see that EWM has the highest estimated welfare when evaluated on the estimating sample. This is not surprising given that EWM maximizes empirical welfare on the estimating sample by construction. In contrast, when we estimate welfare using the auxiliary sample, we see that PWM has the highest estimated welfare, which shows that PWM can effectively protect against overfitting in this example. However, we stress that this difference was not found to be statistically significant (one-tailed p-value 0.34; see Remark B.1 below for details on how our test was constructed). We also note that the performance of PWM on the auxiliary sample is essentially the same as the performance of the random baseline rule; this is also the case when comparing EWM to a similar random baseline (not formally reported). It is possible that this is a feature specific to the monotone policy class (which we view as an exogenous constraint) in this application,

TABLE I

ESTIMATED WELFARE COMPARISONS FOR JTPA DATA[a]

	PWM	EWM	Random Baseline (Average of 1000 Draws)
Estimating sample	\$16,221	\$16,522	\$15,878
Auxiliary sample	\$16,402	\$16,272	\$16,394
	(384)	(395)	(265)

[a]Standard errors in parentheses (see Remark B.1).

and that a more flexible policy class would be able to outperform the random baseline via more selective targeting.

REMARK B.1: In Table I, we provide standard errors for the estimated welfare computed on the auxiliary sample. These should be interpreted as standard errors for the welfare estimate conditional on the estimated treatment allocation. To compute the standard errors, we proceed as follows: given an auxiliary sample of size m and a fixed treatment allocation G, it follows immediately by the central limit theorem that

$$\sqrt{m}\big(\widehat{W}(G) - E\big[\widehat{W}(G)\big]\big) \xrightarrow{d} N\big(0, V(G)\big),$$

as $m \to \infty$, where $V(G) = \mathrm{Var}(\frac{YD}{e(X)}\mathbf{1}\{X \in G\} + \frac{Y(1-D)}{1-e(X)}\mathbf{1}\{X \notin G\})$. Let $\hat{V}(G)$ be the empirical analog of $V(G)$ computed on the auxiliary sample; then the standard error is given by $\sqrt{\hat{V}(G)/m}$. By a similar argument, we can derive the limiting joint distribution for two distinct policies G_1 and G_2, which allows us to construct a difference-in-means test for the welfare difference between the two policies.

B.3. *A Simulation Study*

In this subsection, we perform a small simulation study to highlight the ability of the PWM rule to reduce $\mathcal{G}$-regret in an empirically relevant setting. We consider a situation where the planner has access to threshold-type allocations over five covariates, as described in Examples 2.2 and 3.1, and wishes to perform best-subset selection. The sieve sequence we consider is the same as in Example 3.1, where $\mathcal{G}_k$ is the set of threshold allocations on $k-1$ out of the five covariates. For example, $\mathcal{G}_1$ contains only the allocations $G = \emptyset$ and $G = \mathcal{X}$, which correspond to threshold allocations that use zero covariates, $\mathcal{G}_2$ contains all threshold allocations on one out of the five covariates, etc. We focus here on the setting with five covariates for computational simplicity, but recent work by Chen and Lee (2018) suggests that solving this problem with ten or more covariates could be feasible in practice.

The problem that the planner faces is choosing how many covariates to use in the allocation: for example, suppose that the distribution P is such that some of the available covariates are irrelevant for assigning treatment. Of course, the planner could perform EWM on all the covariates at once, and by the bound in equation (3) this is guaranteed to produce small regret in large enough samples. However, if the sample is not large, the planner may be able to achieve a reduction in regret by performing PWM. Through the lens of Corollary 3.3, our results say that PWM should behave *as if* we had performed EWM in the smallest class $\mathcal{G}_k$ that contains all of the relevant covariates.

We consider the following data generating process: Let $\mathcal{X} = [0,1]^5$, and

$$X_i = (X_{1i}, X_{2i}, \ldots, X_{5i}) \sim \big(U[0,1]\big)^5.$$

The potential outcomes for unit i are specified as:

$$Y_i(1) = 50\big(2X_{2i} - (1 - X_{1i})^4 - 0.5 + 0.5(X_{3i} - X_{4i})\big) + U_{1i},$$
$$Y_i(0) = 50\big(0.5(X_{3i} - X_{4i})\big) + U_{2i},$$

where U_1 and U_2 are distributed as $U[-80, 80]$ random variables which are independent of each other and of X. The covariates enter the potential outcomes in three different ways:

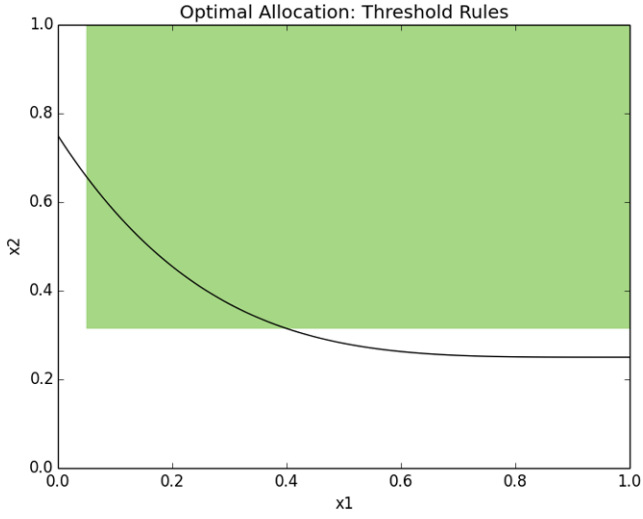

FIGURE 1.—Shaded in green: the best threshold-allocation for our design. Second-best welfare: 21.3. Traced in black: the boundary of the first-best allocation.

- X_{5i} is an irrelevant covariate; it does not play a role in determining potential outcomes at all.
- X_{3i} and X_{4i} affect both treatment and control equally; there will be a nonzero correlation between the observed outcome Y_i and these covariates, but they serve no purpose for treatment assignment.
- X_{1i} and X_{2i} *do* serve a purpose for assigning treatment, and both are used in the optimal threshold allocation. See Figure 1.

Finally, the propensity score $P(D = 1|X)$ is specified to be constant at 0.2.

To implement PWM, we used the holdout penalty, with 3/4 of our sample designated as the estimating sample. In Appendix C, we explain in detail how to implement PWM as a mixed integer linear program.

Our results compare the $\mathcal{G}$-regret of the PWM rule against the regret of performing EWM in $\mathcal{G}_6$ (which corresponds to the class that uses all five covariates) or performing EWM in $\mathcal{G}_3$ computed using 1000 Monte Carlo iterations. Recall that $\mathcal{G}_3$ is the smallest class that contains the optimal threshold allocation. In light of Corollary 3.3, we would hope that PWM behaves similarly to doing EWM in $\mathcal{G}_3$ directly. In Figure 2, we plot the regret of these rules for various sample sizes.

First, we comment on the regret of performing EWM in $\mathcal{G}_6$ (recall that this corresponds to the set of allocations using all five covariates) versus performing EWM in $\mathcal{G}_3$ (which corresponds to the set of allocations that use two of the five covariates). As we would expect, regret decreases as sample size increases. Moreover, performing EWM in $\mathcal{G}_6$ results in larger regret at every sample size: performing EWM in $\mathcal{G}_3$ results in a 34% improvement in regret relative to EWM in $\mathcal{G}_6$ on average, across the sample sizes we consider.

Next, we comment on the performance of PWM. As we had hoped, the regret of PWM is smaller than the regret of performing EWM in $\mathcal{G}_6$ at every sample size: performing PWM results in a 19.8% improvement in regret relative to EWM in $\mathcal{G}_6$ on average, across the sample sizes we consider.

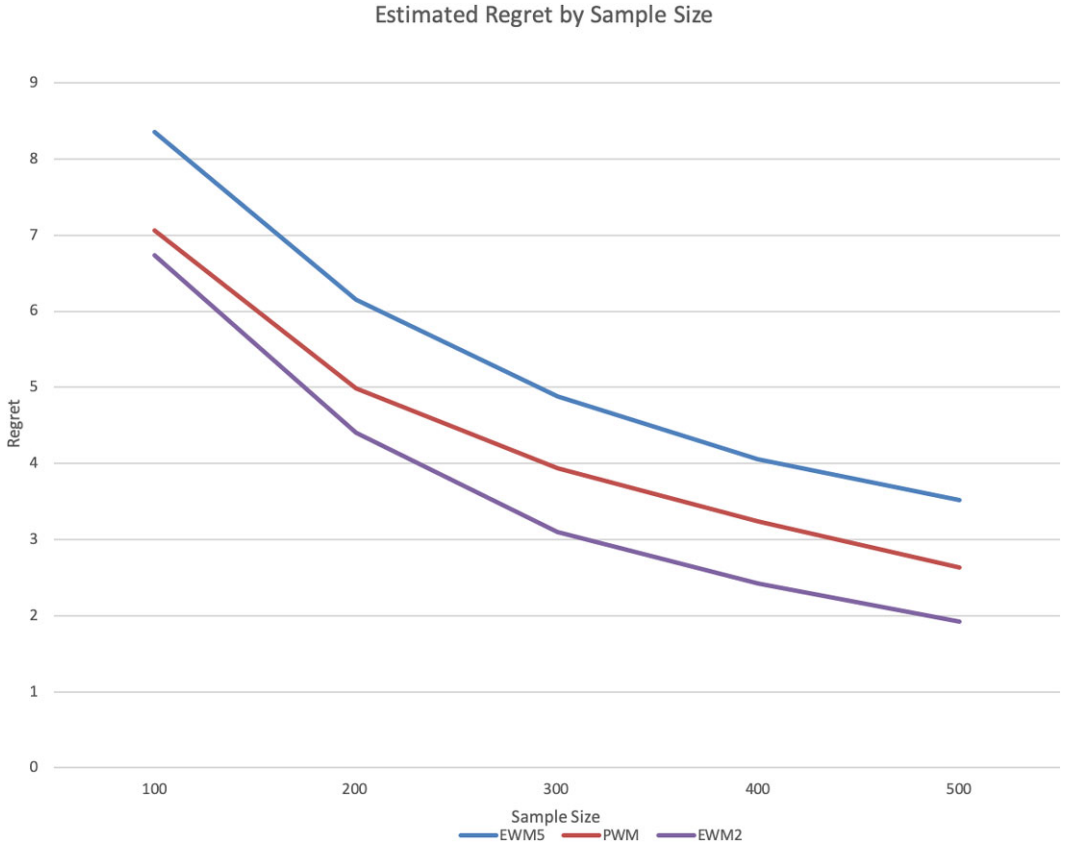

FIGURE 2.—Estimated regret by sample size. Optimal (second-best) welfare: 21.3. EWM5 corresponds to $\mathcal{G}_6$ (five covariates), EWM2 corresponds to $\mathcal{G}_3$ (two covariates).

B.4. *Welfare Maximization With Entropy Restrictions on $\mathcal{G}$*

In this section, we study the treatment choice problem when certain entropy restrictions are imposed on $\mathcal{G}$. First, we derive an upper bound on the maximum regret of EWM under assumptions on the bracketing entropy of $\mathcal{G}$:

Throughout this section, let $\mathcal{X} = [0,1]^{d_x}$. Given a class of sets $\mathcal{G}$ of $\mathcal{X}$, let $\mathcal{H} := \{\mathbf{1}_G : G \in \mathcal{G}\}$. Let $\|\cdot\|_p$ be the $L_p(\mu)$ metric on $\mathcal{H}$, where μ is Lebesgue measure on $\mathcal{X}$. Given $h_1, h_2 \in \mathcal{H}$, with $h_1 \leq h_2$, let $[h_1, h_2] := \{h \in \mathcal{H} : h_1 \leq h \leq h_2\}$. We call the set $[h_1, h_2]$ a *bracket*. Given $\epsilon > 0$, define $N_p^B(\epsilon, \mathcal{H}, \mu)$ to be the smallest k such that, for some pairs (h_j^L, h_j^U), $j = 1, \ldots, k \in \mathcal{H}$, with $h_j^L \leq h_j^U$ and $\|h_j^U - h_j^L\|_p < \epsilon$,

$$\mathcal{H} \subset \bigcup_{j=1}^{k} [h_j^L, h_j^U].$$

We call $H_p^B(\epsilon, \mathcal{H}, \mu) := \log N_p^B(\epsilon, \mathcal{H}, \mu)$ the $L_p(\mu)$ bracketing entropy (in the sense of Alexander (1984)).

Given this definition, we impose the following assumption on the bracketing entropy of $\mathcal{G}$:

ASSUMPTION B.1: *There exist positive constants K, r for which*

$$H_1^B(\epsilon, \mathcal{H}, \mu) \leq K\epsilon^{-r},$$

for all $\epsilon > 0$.

Dudley (1999) provided many examples for which this assumption holds. In particular, by Theorem 8.3.2 in Dudley (1999), if $\mathcal{G}$ is the set of monotone allocations in $[0, 1]^{d_x}$, then Assumption B.1 holds with $r = d_x - 1$ (and the brackets can be constructed in the sense of Alexander (1984)).

As we have emphasized throughout the paper, to obtain bounds on maximum regret for classes of infinite VC dimension, we must impose additional regularity conditions on the DGP. To that end, we consider the following assumption:

ASSUMPTION B.2: *Let $\mathcal{P}_r$ be a set of DGPs such that there exists some constant $A > 0$, where, for every distribution in $\mathcal{P}_r$, the distribution of X is continuous with density bounded above by A.*

With this additional regularity condition, we obtain the following upper bound on maximum regret for EWM:

PROPOSITION B.1: *Under Assumptions* 2.1, 3.1, B.1, *and* B.2, *we have that*

$$\sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E_{P^n}\big[W\big(G^*\big) - W\big(\hat{G}_{\mathrm{EWM}}\big)\big] = O\big(\tau(n)\big),$$

where $\tau(n) = n^{-1/2}$ if $r < 1$, $\tau(n) = \log(n)/\sqrt{n}$ if $r = 1$, and $\tau(n) = n^{-1/(1+r)}$ if $r > 1$.

Note that this result *does not* assume that the first-best allocation is contained in $\mathcal{G}$. From Proposition B.1, we see that for r sufficiently small, EWM converges at a parametric rate (under suitable regularity conditions). Similar results have been obtained in the classification context by Mammen and Tsybakov (1999) and Tsybakov (2004).

Next, we present a lower bound on maximum regret under the following assumption on the $L_1(\mu)$ ϵ-capacity:

Given $\epsilon > 0$, define $D_p(\epsilon, \mathcal{H}, \mu)$ to be the largest k such that there exist functions $h_1, \ldots, h_k \in \mathcal{H}$ with $\|h_i - h_j\|_p > \epsilon$ for $i \neq j$. We call $H_p(\epsilon, \mathcal{H}, \mu) := \log D_p(\epsilon, \mathcal{H}, \mu)$ the $L_p(\mu)$ *epsilon-capacity*.

Given this definition, we impose the following assumption on the ϵ-capacity of $\mathcal{G}$:

ASSUMPTION B.3: *There exist positive constants K_1, K_2, $\epsilon_1 > 0$, $r \geq 1$ such that*

$$K_2\epsilon^{-r} \leq H_1(\epsilon, \mathcal{H}, \mu) \leq K_1\epsilon^{-r},$$

for all $0 < \epsilon \leq \epsilon_1$.

It can be shown that if $\mathcal{G}$ satisfies Assumption B.1, then the upper bound in Assumption B.3 will also hold. However, the reverse may not be true. Dudley (1999) provided many examples for which Assumption B.3 holds, and in particular it holds for the set of monotone allocations in $[0, 1]^{d_x}$ with $r = d_x - 1$ (see Theorem 8.3.2).

With this assumption, we obtain the following lower bound on maximum regret:

PROPOSITION B.2: *Let $\mathcal{P}^*(\mu) \subset \mathcal{P}(M, \kappa)$ be the set of DGPs such that the marginal distribution of X is μ, and $G^* = G^*_{FB}$. Under Assumption B.3, there exists a positive constant B (which depends on M, K_1, K_2, r), such that*

$$\inf_{\hat{G}} \sup_{P \in \mathcal{P}^*(\mu)} E_{P^n}\left[W\left(G^*\right) - W(\hat{G})\right] \geq Bn^{-1/(1+r)},$$

for all $n \geq 1$.

For classes $\mathcal{G}$ such that Assumptions B.1 and B.3 *both* hold, Propositions B.1 and B.2 immediately imply the following rate-optimality result for EWM:

COROLLARY B.1: *Given Assumptions B.1, B.2, and B.3, EWM is rate-optimal over $\mathcal{P}_r \cap \mathcal{P}(M, \kappa)$ for $r > 1$ and rate-optimal up to a* log *factor for $r = 1$.*

PROOF: This follows immediately from the fact that $\mathcal{P}^*(\mu) \subset \mathcal{P}_r \cap \mathcal{P}(M, \kappa)$, and hence

$$\inf_{\hat{G}} \sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E_{P^n}\left[W\left(G^*\right) - W(\hat{G})\right] \geq \inf_{\hat{G}} \sup_{P \in \mathcal{P}^*(\mu)} E_{P^n}\left[W\left(G^*\right) - W(\hat{G})\right]. \quad \textit{Q.E.D.}$$

As we remarked above, for the set of monotone allocations on $[0, 1]^2$, Assumptions B.1 and B.3 hold with $r = 1$. Hence we can conclude that EWM is rate-optimal up to a log-factor for monotone allocations when the distribution of $\mathcal{X}$ is continuous with a bounded density. Note that Corollary B.1 only establishes rate-optimality when r is sufficiently large. For $r < 1$, the lower bound presented in Proposition B.2 is certainly too loose: the set of DGPs used in the proof of Proposition B.2 impose a "hard margin," and hence converge much faster than the parametric rate when $r < 1$.

REMARK B.2: It can be shown that the PWM procedure implemented as in Section 4 can also achieve the rate established in Corollary B.1 (up to a log factor). To see why, note that by using arguments similar to those used in the proof of Propositions B.1, it can be shown that for the holdout penalty,

$$\sup_{P \in \mathcal{P}_r \cap \mathcal{P}(M,\kappa)} E\left[C_m(k)\right] = O\left(\frac{\log(n)}{\sqrt{n}}\right).$$

Combining this result with Proposition 4.1 and Corollary 3.1, we get that the maximum regret of PWM is bounded above by (up to constants)

$$\frac{\log(n)}{\sqrt{n}} + \inf_k \left(2^{-k} + \sqrt{\frac{k}{n}}\right),$$

whose rate of convergence is dominated by the leading term.

APPENDIX C: COMPUTATIONAL DETAILS

In this section, we provide details on how we perform the computations of Section 4 and Appendix B.3. All of our work is implemented in Python paired with Gurobi. We begin with Section 4, then proceed to Appendix B.3.

C.1. *Application Details*

First, we describe how to compute each $\hat{G}_{n,k}$ to solve PWM over monotone allocations. Recall the definition of $\psi_{T,j}(x)$ as defined in Example 3.2. We modify this definition to accommodate the fact that our covariates do not lie in the unit interval. In particular, we restrict ourselves to levels of education that lie in the interval $[5, 20]$, which leads to the following modification:

$$\psi_{T,j}(x) = \begin{cases} 1 - \left| \dfrac{T}{15}(x-5) - j \right|, & x \in \left[\dfrac{j-1}{T/15} + 5, \dfrac{j+1}{T/15} + 5 \right] \cap [5, 20], \\ 0, & \text{otherwise.} \end{cases}$$

Let $\Theta_T = [\theta_0 \;\; \theta_1 \;\; \cdots \;\; \theta_T]'$ and let $\bar{\Theta}_T = [-1 \;\; \theta_0 \;\; \theta_1 \;\; \cdots \;\; \theta_T]'$. Let our two-dimensional covariate be denoted as $x = (x^{(1)}, x^{(2)})$, where $x^{(1)}$ is level of education and $x^{(2)}$ is previous earnings. Let

$$\Psi_T(x) = \left[x^{(2)} \quad \psi_{T,0}\left(x^{(1)}\right) \quad \cdots \quad \psi_{T,T}\left(x^{(1)}\right) \right]'.$$

To compute $\hat{G}_{n,k}$, we solve the following mixed integer linear program (MILP), which modifies the MILP described in Kitagawa and Tetenov (2018) for "Single Linear Index Rules":

$$\begin{aligned} \max_{\substack{\theta_0, \theta_1, \ldots, \theta_T, \\ z_1, \ldots, z_n}} \quad & \sum_{i=1}^{n} \tau_i \cdot z_i \\ \text{subject to} \quad & \frac{\bar{\Theta}_T' \Psi_T(x_i)}{c_{iT}} < z_i \leq \frac{\bar{\Theta}_T' \Psi_T(x_i)}{c_{iT}} + 1, \quad i = 1, \ldots, n, \\ & z_i \in \{0, 1\}, \quad i = 1, \ldots, n, \\ & D_T \Theta_T \leq 0, \end{aligned}$$

where $T = 2^{k-1}$, τ_i is as defined in equation (2), c_T is an appropriate constant (to be discussed in the following sentence), and D_T is the differentiation matrix as defined in Example 3.2. c_{iT} is a constant chosen such that $c_{iT} > \sup_{\Theta_T} |\bar{\Theta}_T' \Psi_T(x_i)|$, which allows us to formulate a set of what are known as "big-M" constraints. To implement such a constraint, it must necessarily be the case that Θ_T is bounded, so in order to implement PWM, we also include an implicit (very large) bound on the possible treatment allocations.[2]

The first two sets of constraints impose that the treatment allocation results in a piecewise linear boundary; the third set of constraints impose that this boundary is monotone. The strength of this formulation is that it imposes monotonicity via a *linear* constraint, which allows us to solve the problem as a MILP.

C.2. *Simulation Details*

We describe a MILP to compute each $\hat{G}_{n,k}$ over threshold allocations on d covariates. Define x to be a $(d+1)$-dimensional vector where $x = (1, x^{(1)}, x^{(2)}, \ldots, x^{(d)})$, with the

[2]Big-M constraints have the potential to cause numerical instabilities when solving MILPs that are poorly formulated. We found that it was important to ensure that the covariates are scaled to within the same order of magnitude and that the `IntFeasTol` and `FeasibilityTol` parameters in Gurobi were set to their smallest possible values.

last d components denoting the d covariates, and suppose $x \in [0,1]^{d+1}$, which is the case in the simulation design. We define the threshold β_k on covariate $x^{(k)}$ to be a $(d+1)$-dimensional vector such that the first component is in $[-1,1]$, all other components other than the $(k+1)$st are zero, and the $(k+1)$st component is one of $\{-1,0,1\}$. Let $A = \{1,2,\ldots,d\}$ index the dimension of the threshold. We modify the MILP described in Kitagawa and Tetenov (2018) for "Multiple Linear Index Rules":

$$
\begin{aligned}
\max_{\substack{\{\beta_a\}_{a\in A},\\ \{z_1^a,\ldots,z_n^a\}_{a\in A},z_1^*,\ldots,z_n^*}} \quad & \sum_{i=1}^{n} \tau_i \cdot z_i^* \\
\text{subject to} \quad & \frac{x_i'\beta_a}{c} < z_i^a \leq \frac{x_i'\beta_a}{c} + 1, \quad i = 1,\ldots,n,\ a \in A, \\
& 1 - |A| + \sum_{a\in A} z_i^a \leq z_i^* \leq \frac{1}{|A|}\sum_{a\in A} z_i^a, \quad i = 1,\ldots,n, \\
& \beta_a^{(1)} \in [-1,1], \quad a \in A, \\
& \beta_a^{(j)} = 0, \quad j > 1,\ j \neq a+1,\ a \in A, \\
& \sum_{a\in A} e_a = k, \\
& -e_a \leq \beta_a^{(1)} \leq e_a, \quad a \in A, \\
& \beta_a^{(a+1)} = y_{a,1} - y_{a,2}, \quad a \in A, \\
& y_{a,1} + y_{a,2} = e_a, \quad a \in A, \\
& \left\{z_i^a\right\}_{a\in A}, z_i^* \in \{0,1\}, \quad i = 1,\ldots,n, \\
& \{e_a\}_{a\in A} \in \{0,1\}, \quad a \in A, \\
& \{y_{a,1}\}_{a\in A}, \{y_{a,2}\}_{a\in A} \in \{0,1\}, \quad a \in A.
\end{aligned}
$$

The constraints serve the following roles: the first two constraints enforce the assignment of observations to treatment, the next two constraints enforce part of the structure of the threshold allocation, the fifth constraint specifies that only k thresholds can be used, and the three subsequent constraints enforce this. Again, we require an appropriately chosen constant c to implement a set of big-M constraints, but in this case the choice is straightforward: $c = d+2$ will suffice since this guarantees that $c > x_i'\beta_a$ for any possible x_i and β_a, by construction.

REMARK C.1: Solving the above program for the simulation design of Appendix B.3 with a sample size of 2000 took approximately one hour and fifteen minutes on a 2018 iMac. In practice, the solution of this MILP could potentially be further optimized using the improvements developed in Bertsimas, King, and Mazumder (2016) and Chen and Lee (2018). Alternatively, careful implementation of a direct parameter search could also considered; see, for example, the work in Zhou, Athey, and Wager (2018) using a tree-based policy class.

APPENDIX D: PROOFS FOR APPENDIX B

PROOF OF PROPOSITION B.1: We follow the general strategy of Theorem 1 in Mammen and Tsybakov (1999). Let $\bar{W}(\cdot) = (\kappa/M)W(\cdot)$ be a normalized version of $W(\cdot)$. Let G^* be a maximizer of $\bar{W}(\cdot)$ in $\mathcal{G}$. Let $\mathcal{P} = \mathcal{P}_r \cap \mathcal{P}(M, \kappa)$ and define

$$T_n = \sqrt{n}\frac{\bar{W}(G^*) - \bar{W}(\hat{G}) - (\bar{W}_n(G^*) - \bar{W}_n(\hat{G}))}{q_n},$$

where $q_n = 1$ if $r < 1$, $q_n = \log(n)$ if $r = 1$, and $q_n = n^{(r-1)/2(r+1)}$ if $r > 1$. By the definition of $\hat{G}$ and G^*, we have that $\bar{W}_n(\hat{G}) \geq \bar{W}_n(G^*)$, $\bar{W}(\hat{G}) \leq \bar{W}(G^*)$, and hence we have that

$$0 \leq \sqrt{n}q_n^{-1}(\bar{W}(G^*) - \bar{W}(\hat{G})) \leq T_n.$$

Now we argue that $E[T_n]$ is uniformly bounded over $\mathcal{P}$ for n sufficiently large, which, given the definition of q_n, implies the statement of the theorem. To that end, note that

$$E[T_n] \leq E[S_n],$$

where

$$\begin{aligned} S_n &= \sup_{G\in\mathcal{G}} \sqrt{n}q_n^{-1}\left|\bar{W}(G^*) - \bar{W}(G) - (\bar{W}_n(G^*) - \bar{W}_n(G))\right| \\ &= \sup_{G\in\mathcal{G}} \sqrt{n}q_n^{-1}\left|\frac{1}{n}\sum_{i=1}^{n}(\bar{g}(Z_i)(\mathbf{1}\{X_i \in G^*\} - \mathbf{1}\{X_i \in G\}) \right. \\ &\quad \left. - E[\bar{g}(Z_i)(\mathbf{1}\{X_i \in G^*\} - \mathbf{1}\{X_i \in G\})])\right|, \end{aligned}$$

with

$$\bar{g}(Z_i) = \frac{\kappa}{M}g(Y_i, D_i, X_i) = \frac{\kappa}{M}\left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right).$$

For the case $r < 1$, we can invoke Lemma D.1 to conclude immediately that

$$\sup_{P\in\mathcal{P}} E[S_n] = O(1),$$

so we are done. For the case $r \geq 1$, note that $S_n \leq 2\sqrt{n}/q_n$, which gives that, for any $D > 0$,

$$E[S_n] \leq D + 2\frac{\sqrt{n}}{q_n}P(S_n > D),$$

hence we can apply Corollary D.1 to the last probability to conclude that $\sup_{P\in\mathcal{P}} E[S_n] = O(1)$. Let $\widetilde{\mathcal{F}} = \{\bar{g}\cdot(\mathbf{1}_{G^*} - \mathbf{1}_G) : G \in \mathcal{G}\}$; then for an appropriate choice of D (which depends on P *only* through K and r, and A), Corollary D.1 gives

$$P(S_n > D) \leq C\exp(-D^2 q_n^2),$$

for some constant C which depends on P *only* through K, r, and A. Hence we can conclude that

$$\sup_{P\in\mathcal{P}} E[T_n] \leq \sup_{P\in\mathcal{P}} E[S_n] = O(1),$$

as desired. *Q.E.D.*

PROOF OF PROPOSITION B.2: Define

$$L_n := \inf_{\hat{G}} \sup_{P\in\mathcal{P}^*(\mu)} E_{P^n}\big[W\big(G^*\big) - W(\hat{G})\big].$$

We follow the general strategy of Theorem 6 in Massart and Nédélec (2006). For every $h \in \mathcal{H} = \{\mathbf{1}_G : G \in \mathcal{G}\}$, set $\tau_h(x) = (M/4)(2h(x) - 1)$, $\gamma_h(x) = (2/M)\tau_h(x)$, and define P_h as the joint distribution on $\mathcal{X} \times \{0, 1\} \times \mathcal{Y}^2$ (i.e., the set of realizations of $(X, D, Y(1), Y(0))$) such that, under P_h, X has distribution μ,

$$Y(1)|\{X = x\} = \begin{cases} \dfrac{M}{2} & \text{with prob. } \dfrac{1+\gamma_f(x)}{2}, \\ -\dfrac{M}{2} & \text{with prob. } \dfrac{1-\gamma_f(x)}{2}, \end{cases}$$

$Y(0)|\{X = x\} = 0$, and D is Bernoulli(0.5) independent of everything else. Note that, by construction, we have that $\tau_h(x) = E_{P_h}[Y(1) - Y(0)|X = x] = \tau(x)$, h describes the first-best decision rule under P_h, and $P_h \in \mathcal{P}^*(\mu)$.

Next, let $\mathcal{C}$ be a finite subset of $\mathcal{H}$; then it follows that

$$\inf_{\hat{G}} \sup_{P\in\mathcal{P}^*(\mu)} E_{P^n}\big[W\big(G^*\big) - W(\hat{G})\big] \geq \inf_{\hat{G}} \sup_{h\in\mathcal{C}} E_h\big[W\big(G^*\big) - W(\hat{G})\big],$$

where $E_h = E_{P_h^n}$. Since, under P_h, G^* is the first-best allocation by construction, we get that

$$W\big(G^*\big) - W(G) = \int_{G^*\Delta G} \big|\tau(X)\big|\, dP_X,$$

for any $G \in \mathcal{G}$. Hence it follows that, given the construction of τ_h,

$$W\big(G^*\big) - W(G) = \frac{M}{4}\mu\big(G^*\Delta G\big).$$

Putting all this together and using the fact that $h = \mathbf{1}_{G^*}$ under P_h:

$$L_n \geq \inf_{\hat{h}\in\mathcal{H}} \sup_{h\in\mathcal{C}} \frac{M}{4} E_h\big[\|h - \hat{h}\|_1\big],$$

where $\hat{h} = \mathbf{1}_{\hat{G}}$, and $\|\cdot\|_1$ is the $L_1(\mu)$ norm. Define the statistic

$$\tilde{h} = \arg\min_{h\in\mathcal{C}} \|h - \hat{h}\|_1;$$

then by the triangle inequality it follows that

$$\inf_{\hat{h}\in\mathcal{H}} \sup_{h\in\mathcal{C}} E_h\big[\|\hat{h} - h\|_1\big] \geq \frac{1}{2} \inf_{\tilde{h}\in\mathcal{C}} \sup_{h\in\mathcal{C}} E_h\big[\|\tilde{h} - h\|_1\big].$$

We now construct the appropriate set $\mathcal{C}$. Let $\mathcal{C}'$ be an ϵ-packing set of $\mathcal{H}$, and let $\mathcal{C}''$ be a $C\epsilon$ cover of $\mathcal{H}$ for some $C > 1$, $\epsilon > 0$ to be specified later. By definition, each $h \in \mathcal{C}'$ lies in some ball of radius $C\epsilon$ centered at a point in $\mathcal{C}''$. So by taking $\mathcal{C}$ to be the intersection of $\mathcal{C}'$ with such a ball in $\mathcal{C}''$ which results in a set of maximal cardinality, we get that for $h_1, h_2 \in \mathcal{C}$, where $h_1 \neq h_2$,

$$\epsilon \leq \|h_1 - h_2\|_1 \leq C\epsilon,$$

and moreover,

$$\log\bigl(|\mathcal{C}|\bigr) \geq H_1(\epsilon, \mathcal{H}, \mu) - H_1(C\epsilon, \mathcal{H}, \mu).$$

To see this, note that since we have constructed $\mathcal{C}$ to have maximal cardinality, it must be the case that

$$|\mathcal{C}| \geq \frac{|\mathcal{C}'|}{|\mathcal{C}''|},$$

and by definition, $|\mathcal{C}'| = H_1(\epsilon, \mathcal{H}, \mu)$, $|\mathcal{C}''| \leq H_1(C\epsilon, \mathcal{H}, \mu)$.

Now, by Markov's inequality,

$$\inf_{\tilde{h} \in \mathcal{C}} \sup_{h \in \mathcal{C}} E_h\bigl[\|\tilde{h} - h\|_1\bigr] \geq \epsilon \inf_{\tilde{h} \in \mathcal{C}}\Bigl(1 - \inf_{h \in \mathcal{C}} P_h^n(\tilde{h} = h)\Bigr),$$

and hence by Lemma 8 in Massart and Nédélec (2006),

$$L_n \geq \frac{M\epsilon}{8}(1 - \alpha),$$

where $\alpha := 0.71$, as long as $\bar{\mathcal{K}} \leq \alpha \log(|\mathcal{C}|)$, where, for some fixed $h_0 \in \mathcal{C}$,

$$\begin{aligned}\bar{\mathcal{K}} &:= \frac{1}{|\mathcal{C}| - 1} \sum_{h \in \mathcal{C}, h \neq h_0} \mathcal{K}\bigl(P_h^n, P_{h_0}^n\bigr) \\ &= \frac{n}{|\mathcal{C}| - 1} \sum_{h \in \mathcal{C}, h \neq h_0} \mathcal{K}(P_h, P_{h_0}),\end{aligned}$$

and $\mathcal{K}(\cdot, \cdot)$ is the Kullback–Leibler divergence. By Lemma D.2, we have that

$$\bar{\mathcal{K}} \leq n \sup_{h \in \mathcal{C}} \|h - h_0\|_1 \leq n\epsilon,$$

where the last inequality follows by the construction of $\mathcal{C}$.

Again by the construction of $\mathcal{C}$, we can choose C such that there exists some positive constant C_1 for which $\log(|\mathcal{C}|) \geq C_1 \epsilon^{-r}$ for $\epsilon \leq \epsilon_1$, and therefore

$$\frac{\bar{\mathcal{K}}}{\log|\mathcal{C}|} \leq \frac{n}{C_1}\epsilon^{1+r}.$$

Hence we can conclude that $L_n \geq (M\epsilon/8)(1 - \alpha)$ whenever

$$\frac{n}{C_1}\epsilon^{1+r} \leq \alpha,$$

that is,

$$\epsilon \leq (\alpha C_1)^{1/(1+r)} n^{-1/(1+r)}.$$

Now, we may also choose C such that $\alpha C_1 \leq \epsilon_1^{1+r}$, so that the constraint $\epsilon \leq \epsilon_1$ is satisfied if we set

$$\epsilon = (\alpha C_1)^{1/(1+r)} n^{-1/(1+r)}.$$

Hence we have that

$$L_n = \inf_{\hat{G}} \sup_{P \in \mathcal{P}^*(\mu)} E_{P^n}\big[W\big(G^*\big) - W(\hat{G})\big] \geq A n^{-1/(1+r)},$$

where A is a constant which depends on K_1, K_2, M, and r as desired. *Q.E.D.*

PROPOSITION D.1: *Let $\{Z_i\}_{i=1}^n$ be a sequence of i.i.d. random vectors with distribution P. Let $Z = (Z_1, Z_2)$, and let $\mathcal{F}$ be a class of real-valued functions of the form $f(z) = f(z_1, z_2) = g(z) \cdot h(z_2)$, where $h \in \mathcal{H}$, $\mathcal{H}$ is a class of functions with values in $\{0, 1\}$, and g is some fixed real-valued function (which may depend on P) such that $|g| \leq 1$. Let P_2 be the marginal distribution of Z_2 and suppose $\mathcal{H}$ satisfies*

$$H_2^B(\epsilon, \mathcal{H}, P_2) \leq K \epsilon^{-\ell}, \tag{D.1}$$

for some constants $K > 0$, $\ell \geq 2$, for all $\epsilon > 0$. Then there exist positive constants C_1, C_2, C_3, C_4 (which depend only on K and ℓ) such that if

$$\xi \leq \frac{\sqrt{n}}{128}, \tag{D.2}$$

and

$$\xi \geq \begin{cases} C_1 n^{(\ell-2)/2(\ell+2)}, & \ell \geq 2, \\ C_2 \log \max(n, e), & \ell = 2, \end{cases} \tag{D.3}$$

then

$$P^n\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(Z_i) - Ef(Z_i)\big] \right| > \xi \right) \leq C_4 \exp\big(-\xi^2\big).$$

PROOF: We follow the general strategy of Theorem 2.3 and Corollary 2.4 in Alexander (1984). Let

$$\nu_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(Z_i) - Ef(Z_i)\big].$$

We begin with a series of definitions. Let $\delta_0 > \delta_1 > \cdots > \delta_N > 0$ be a sequence of real numbers where $\{\delta_j\}_j$ and N are to be specified precisely later in the proof. For every $0 \leq j \leq N$, there exists a set of δ_j-brackets $\mathcal{H}_j^B$ of $\mathcal{H}$ such that $|\mathcal{H}_j^B| = N_2^B(\delta_j, \mathcal{H}, P_2)$. For each $h \in \mathcal{H}$, let $(h_j^L, h_j^U) \in \mathcal{H}_j^B$ be the brackets such that $h_j^L \leq h \leq h_j^U$ and $\|h_j^H - h_j^L\|_2 < \delta_j$.

Define the function $H_\theta(\cdot) : (0, \infty) \to [0, \infty)$ as follows:

$$H_\theta(u) = \begin{cases} Ku^{-\ell}, & u \le 1, \\ -\dfrac{K}{\theta}u + \dfrac{K(1+\theta)}{\theta}, & u \in (1, 1+\theta], \\ 0, & u > 1+\theta. \end{cases}$$

Note that by construction, H_θ is continuous on $[0, \infty)$, and by Assumption (D.1) and the fact that $\mathcal{H}$ has diameter 1 by definition, $N_2^B(\delta_j, \mathcal{H}, P_2) \le \exp(H_\theta(\delta_j))$ for $\theta > 0$. From now on, we fix such a $\theta > 0$, and suppress θ from our notation. For any $f \in \mathcal{F}$, we have by definition that $f = g \cdot h$ for some $h \in \mathcal{H}$, and so given the bracket (h_j^L, h_j^U), define $f_j^L := g \cdot h_j^L \mathbf{1}\{g \ge 0\} + g \cdot h_j^U \mathbf{1}\{g < 0\}$, and $f_j^U := g \cdot h_j^U \mathbf{1}\{g \ge 0\} + g \cdot h_j^L \mathbf{1}\{g < 0\}$, and note that by construction, (f_j^L, f_j^U) is a bracket for f. Let $f_j = f_j^L$, and let $\mathcal{F}_j = \{f_j : f \in \mathcal{F}\}$; then $|\mathcal{F}_j| \le \exp(H(\delta_j))$ and for every $f \in \mathcal{F}$, $\|f - f_j\|_2 < \delta_j$.

By a standard chaining argument,

$$P\Big(\sup_{f \in \mathcal{F}} \big|\nu_n(f)\big| > \xi\Big) \le R_1 + R_2 + R_3,$$

where

$$R_1 = |\mathcal{F}_0| \sup_{f \in \mathcal{F}} P\bigg(\big|\nu_n(f)\big| > \frac{7}{8}\xi\bigg),$$

$$R_2 = \sum_{j=0}^{N-1} |\mathcal{F}_j||\mathcal{F}_{j+1}| \sup_{f \in \mathcal{F}} P\big(\big|\nu_n(f_j - f_{j+1})\big| > \eta_j\big),$$

$$R_3 = P\bigg(\sup_{f \in \mathcal{F}} \big|\nu_n(f_N - f)\big| > \frac{\xi}{16} + \eta_N\bigg),$$

where $\{\eta_j\}_j$ are chosen such that $\sum_{j=0}^N \eta_j \le \xi/16$ and will be specified precisely later in the proof. We now choose $\{\delta_j\}_j$, $\{\eta_j\}_j$, and N to make these three terms sufficiently small.

First, consider R_1. Take δ_0 such that $H(\delta_0) = \xi^2/4$. Then by Hoeffding's inequality,

$$R_1 \le 2|\mathcal{F}_0| \exp\bigg(-2\Big(\frac{7}{8}\xi\Big)^2\bigg) \le 2\exp\big(-\xi^2\big).$$

Next, we develop a bound on R_2. Since by construction $\|f_j - f_{j+1}\|_2 \le 2\delta_j$, it follows by repeated applications of Bennet's inequality (see Lemma D.3) that

$$R_2 \le \sum_{j=0}^{N-1} 2\exp\big(2H(\delta_{j+1})\big) \exp\big(-\psi_1\big(\eta_j, n, 4\delta_j^2\big)\big),$$

where ψ_1 has the properties described in Lemma D.3. Next, consider R_3. Given the construction of $\mathcal{F}_N$ and writing $f = g \cdot h$,

$$\begin{aligned} \big|\nu_n(f_N - f)\big| &\le \big|\nu_n\big(f_N^U - f_N^L\big)\big| + 2\sqrt{n}\big\|f_N^U - f_N^L\big\|_1 \\ &\le \big|\nu_n\big(f_N^U - f_N^L\big)\big| + 2\sqrt{n}\delta_N^2, \end{aligned}$$

since $\|f_N^U - f_N^L\|_1 \leq \|h_N^U - h_N^L\|_1 \leq \delta_N^2$ (where here we use the fact that h_N^U, h_N^L take values in $\{0,1\}$). Take $\delta_N \leq s := (\xi/(32\sqrt{n}))^{1/2}$; then by the above derivation and Bennet's inequality,

$$\begin{aligned} R_3 &\leq P\Big(\sup_{f\in\mathcal{F}} |\nu_n(f_N^U - f_N^L)| > \eta_N\Big) \\ &\leq 2|\mathcal{F}_N| \exp\big(-\psi_1\big(\eta_N, n, \delta_N^2\big)\big). \end{aligned}$$

To complete our bounds on R_2 and R_3, we consider two separate cases. First, suppose $\delta_0 \leq s$ as defined above. Then by taking $N = 0$ and $\eta_0 = \xi/16$, we have that $R_2 = 0$ and

$$R_3 \leq 2|\mathcal{F}_0| \exp\big(-\psi_1\big(\eta_0, n, \delta_0^2\big)\big).$$

Since $\delta_0 \leq s$, we have that

$$2\eta_0 = \frac{\xi}{8} = 4\sqrt{n}\left(\frac{\xi}{32\sqrt{n}}\right) \geq 4\sqrt{n}\delta_0^2,$$

and hence by the properties of ψ_1,

$$\psi_1\big(\eta_0, n, \delta_0^2\big) \geq \frac{1}{4}\psi_1\big(2\eta_0, n, \delta_0^2\big) \geq \frac{1}{4}\eta_0\sqrt{n}.$$

Using Assumption (D.2), we can then conclude that

$$\psi_1\big(\eta_0, n, \delta_0^2\big) \geq \frac{1}{4}\eta_0\sqrt{n} = \frac{\xi}{64}\sqrt{n} \geq 2\xi^2.$$

By the definition of δ_0,

$$|\mathcal{F}_0| \leq \exp\left(\frac{\xi^2}{4}\right),$$

so that putting everything together yields

$$R_2 + R_3 \leq 4\exp\big(-\xi^2\big).$$

Next, consider the case where $\delta_0 > s$. Let N and $\{\delta_j\}_{j=1}^N$ be as in Lemma D.4, where $t = \delta_0$, and s is as defined above. Let $\eta_j = 8\sqrt{2}\delta_j H(\delta_{j+1})^{1/2}$ for $0 \leq j < N$, $\eta_N = 8\sqrt{2}\delta_N H(\delta_N)^{1/2}$. Then by Lemma D.4,

$$\sum_{j=0}^{N} \eta_j = 8\sqrt{2}\sum_{j=0}^{N} \delta_j H(\delta_{j+1})^{1/2} \leq 64\sqrt{2}\int_{s/4}^{\delta_0} H(u)^{1/2}\,du.$$

Now, by the definition of $H(\cdot)$, we have that for $0 < s < t$,

$$\int_s^t H(u)^{1/2}\,du \leq \begin{cases} K^{1/2}\log(1/s), & \ell = 2 \text{ and } t \leq 1, \\ 2K^{1/2}(\ell-2)^{-1}s^{(2-\ell)/2}, & \ell > 2, \end{cases}$$

and so by combining this with Assumptions (D.2) and (D.3) (with C_1 sufficiently large), it can be shown that

$$\sum_{j=0}^{N} \eta_j \leq \frac{\xi}{16},$$

and hence our choice of $\{\eta_j\}_j$ is consistent with our construction (note that when $\ell = 2$, the above inequality only applies when $t \leq 1$; however, we can argue using $\delta_0 \leq 1 + \theta$ that $\sum \eta_j \leq \xi/16 + C'\theta$ for some constant $C' > 0$, for all $\theta > 0$, and hence our result holds for $P(\sup_f |\nu_n(f)| \geq \xi + C'\theta)$ where $\theta > 0$ can be made arbitrarily small). By Assumption (D.3) (with C_1 sufficiently large), it can also be shown that

$$H(s) \leq \frac{\xi\sqrt{n}}{16},$$

and hence it follows that

$$\left(\frac{\eta_j}{4\delta_j^2\sqrt{n}}\right)^2 < \frac{8H(s)}{ns^2} \leq 16,$$

so that by the properties of ψ_1,

$$\psi_1\big(\eta_j, n, 4\delta_j^2\big) \geq \frac{\eta_j^2}{16\delta_j^2}.$$

Using our bound on R_2, we can then conclude that

$$R_2 \leq \sum_{j=0}^{N-1} 2\exp\left(2H(\delta_{j+1}) - \frac{\eta_j^2}{16\delta_j^2}\right) \leq \sum_{j=0}^{N-1} 2\exp\big(-4^{j+1}H(\delta_0)\big).$$

Similarly, we can argue that

$$R_3 \leq 2\exp\big(-4^{N+1}H(\delta_0)\big).$$

Putting these together, and using Assumption (D.3),

$$R_2 + R_3 \leq \sum_{j=0}^{\infty} 2\exp\big(-4^{j+1}H(\delta_0)\big) \leq C\exp\big(-\xi^2\big),$$

where C is a constant that depends only on K and ℓ. *Q.E.D.*

COROLLARY D.1: *Let $\{Z_i\}_{i=1}^n$ be a sequence of i.i.d. random vectors with distribution P. Let $Z = (Z_1, Z_2)$, and let $\mathcal{F}$ be a class of real-valued functions of the form $f(z) = f(z_1, z_2) = g(z) \cdot h(z_2)$, where $h \in \mathcal{H}$, $\mathcal{H}$ is a class of functions with values in $\{0, 1\}$, and g is some fixed real-valued function (which may depend on P) such that $|g| \leq 1$. Suppose $\mathcal{H}$ satisfies Assumption B.1, and suppose that P_2, the marginal distribution of Z_2, has a density with respect to Lebesgue measure bounded above by A. Then there exist positive constants D_1, D_2, D_3 (which depend only on K, r, and A) such that, for $n \geq 3$,*

$$P^n\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[f(Z_i) - Ef(Z_i)\big]\right| > xq_n\right) \leq D_3\exp\big(-x^2q_n^2\big),$$

for $D_1 \leq x \leq D_2\sqrt{n}/q_n$, where

$$q_n = \begin{cases} \log n, & r = 1, \\ n^{(r-1)/2(r+1)}, & r > 1!. \end{cases}$$

PROOF: First, note that since P_2 has density with respect to Lebesgue measure bounded above by A, we get that by Assumption B.1,

$$H_1^B(\epsilon, \mathcal{H}, P_2) \leq C\epsilon^{-r},$$

where C is some constant which depends only on K and A. Next, since $\mathcal{H}$ consists of $\{0,1\}$-valued functions, any ϵ-bracket for $\mathcal{H}$ in L_1 is an $\epsilon^{1/2}$-bracket in L_2 and vice versa. Hence we get that

$$H_2^B(\epsilon, \mathcal{H}, P_2) \leq K'\epsilon^{-2r},$$

for some constant K' which depends only on K, r, and A. The result then follows immediately by Proposition D.1. *Q.E.D.*

LEMMA D.1: *Maintain the assumptions of Proposition B.1 with $r < 1$. Let S_n be as in the proof of Proposition B.1. Then*

$$\sup_{P \in \mathcal{P}} E[S_n] = O(1).$$

PROOF: By definition, $S_n \leq \sqrt{n}S_n^{(1)} + S_n^{(2)}$, where

$$S_n^{(1)} = \sup_{G \in \mathcal{G}: \|f_G\| \leq n^{-1/(2+2r)}} \left| \frac{1}{n} \sum_{i=1}^n \left(\tilde{f}_G(Z_i) - E\left[\tilde{f}_G(Z_i)\right]\right) \right|,$$

$$S_n^{(2)} = \sup_{G \in \mathcal{G}: \|f_G\| \geq n^{-1/(2+2r)}} \sqrt{n} \frac{\left| \frac{1}{n} \sum_{i=1}^n \left(\tilde{f}_G(Z_i) - E\left[\tilde{f}_G(Z_i)\right]\right) \right|}{\|\tilde{f}_G\|^{1-r}},$$

with $\tilde{f}_G = \bar{g} \cdot (\mathbf{1}\{X \in G^*\} - \mathbf{1}\{X \in G\})$ and $\|\cdot\|$ the $L_2(P)$ norm, and we have used the fact that $\|\tilde{f}_G\| \leq 1$. We will use Lemma 5.13 in van de Geer (2000) to bound each of these quantities. To apply the lemma, let g in her notation be $\tilde{f}_G$ in ours, and g_0 in her notation be zero. Set $\alpha = 2r$, $\beta = 0$ in the statement of her lemma. It remains to verify condition (5.40) in her lemma for the class $\widetilde{\mathcal{F}} = \{\tilde{f}_G : G \in \mathcal{G}\}$, but this follows by Assumption B.1 by combining the arguments from the proof of Corollary D.1 and the proof of Proposition D.1. By inequality (5.42) in her lemma,

$$\sup_{P \in \mathcal{P}} n^{1/(1+r)} E\left[S_n^{(1)}\right] = O(1),$$

and hence since $r < 1$,

$$\sup_{P \in \mathcal{P}} \sqrt{n} E\left[S_n^{(1)}\right] = O(1).$$

By inequality (5.43),

$$\sup_{P\in\mathcal{P}} E\big[S_n^{(2)}\big] = O(1).$$

Combining both of these together gives our desired result. *Q.E.D.*

LEMMA D.2: *Let P_f be specified as in the proof of Proposition* B.2. *Then for $f, g \in \mathcal{H}$ such that $f \neq g$,*

$$\mathcal{K}(P_f, P_g) \leq \|f - g\|_1,$$

where $\mathcal{K}(\cdot,\cdot)$ is the Kullback–Leibler divergence.

PROOF: Let $Q_{f,x}(\cdot)$ denote the probability mass function of $(Y(1), D)|X = x$ under P_f (recall that $Y(0)|X = x$ is degenerate, so we omit it from the calculation). If $f \neq g$, a direct calculation shows that

$$\mathcal{K}(Q_{f,x}, Q_{g,x}) = \frac{1}{2}\log(3).$$

Hence

$$\begin{aligned}\mathcal{K}(P_f, P_g) &= \int_{\mathcal{X}} \mathcal{K}(Q_{f,x}, Q_{g,x})\mathbf{1}\big\{f(x) \neq g(x)\big\}\, d\mu \\ &= \frac{1}{2}\log 3\|f - g\|_1 \leq \|f - g\|_1.\end{aligned}$$ *Q.E.D.*

LEMMA D.3—Bennet's Inequality: see Theorem 2.9 in Boucheron, Lugosi, and Massart (2013): *Let $\{Z_i\}_{i=1}^n$ be a sequence of independent random vectors with distribution P. Let f be some function taking values in* $[0, 1]$ *and define*

$$\nu_n(f) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[f(Z_i) - Ef(Z_i)\big].$$

Then for any $\xi \geq 0$,

$$P^n\big(|\nu_n(f)| > \xi\big) \leq 2\exp\big(-\psi_1(\xi, n, \alpha)\big),$$

where $\alpha = \operatorname{var}(\nu_n(f))$ and

$$\psi_1(\xi, n, \alpha) = \xi\sqrt{n}h\left(\frac{\xi}{\sqrt{n}\alpha}\right),$$

with

$$h(x) = \big(1 + x^{-1}\big)\log(1 + x) - 1.$$

Importantly, ψ has the following two relevant properties:

$$\psi_1(\xi, n, \alpha) \geq \psi_1(C\xi, n, \rho\alpha) \geq C^2\rho^{-1}\psi_1(\xi, n, \alpha),$$

for $C \leq 1$, $\rho \geq 1$, *and*

$$\psi_1(\xi, n, \alpha) \geq \begin{cases} \dfrac{\xi^2}{4\alpha}, & \text{if } \xi < 4\sqrt{n}\alpha, \\ \dfrac{\xi}{2}\sqrt{n}, & \text{if } \xi \geq 4\sqrt{n}\alpha. \end{cases}$$

LEMMA D.4—Lemma 3.1 in Alexander (1984): *Let* $H : (0, t] \to \mathbb{R}^+$ *be a decreasing function, and let* $0 < s < t$. *Let* $\delta_0 := t$, $\delta_{j+1} := s \vee \sup\{x \leq \delta_j/2 : H(x) \geq 4H(\delta_j)\}$ *for* $j \geq 0$, *and* $N := \min\{j : \delta_j = s\}$. *Then*

$$\sum_{j=0}^{N} \delta_j H(\delta_j)^{1/2} \leq 8 \int_{s/4}^{t} H(u)^{1/2}\, du.$$

REFERENCES

ALEXANDER, K. S. (1984): "Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm," *The Annals of Probability*, 12, 1041–1067. [12,13,20,26]

BARTLETT, P. L., S. BOUCHERON, AND G. LUGOSI (2002): "Model Selection and Error Estimation," *Machine Learning*, 48, 85–113. [3]

BERTSIMAS, D., A. KING, AND R. MAZUMDER (2016): "Best Subset Selection via a Modern Optimization Lens," *The Annals of Statistics*, 44, 813–852. [16]

BOUCHERON, S., G. LUGOSI, AND P. MASSART (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London: Oxford University Press. [25]

CHEN, L.-Y., AND S. LEE (2018): "Best Subset Binary Prediction," *Journal of Econometrics*, 206, 39–56. [10, 16]

DUDLEY, R. M. (1999): *Uniform Central Limit Theorems*, Vol. 23. Cambridge: Cambridge Universsity Press. [13]

GYÖRFI, L., L. DEVROYE, AND G. LUGOSI (1996): *A Probabilistic Theory of Pattern Recognition*. New York: Springer. [3,4]

KITAGAWA, T., AND A. TETENOV (2018): "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616. [1,4,5,8,9,15,16]

MAMMEN, E., AND A. B. TSYBAKOV (1999): "Smooth Discrimination Analysis," *The Annals of Statistics*, 27, 1808–1829. [13,17]

MASSART, P., AND É. NÉDÉLEC (2006): "Risk Bounds for Statistical Learning," *The Annals of Statistics*, 34, 2326–2366. [18,19]

TSYBAKOV, A. B. (2004): "Optimal Aggregation of Classifiers in Statistical Learning," *Annals of Statistics*, 32, 135–166. [13]

VAN DE GEER, S. A. (2000): *Empirical Processes in M-Estimation*, Vol. 6. Cambridge: Cambridge University Press. [24]

VAN DER VAART, A. W., AND J. A. WELLNER (1996): "Weak Convergence," in *Weak Convergence and Empirical Processes*. New York: Springer, 16–28. [9]

ZHOU, Z., S. ATHEY, AND S. WAGER (2018): "Offline Multi-Action Policy Learning: Generalization and Optimization," arXiv preprint, arXiv:1810.04778. [16]

Co-editor Ulrich K. Müller handled this manuscript.

Manuscript received 15 June, 2018; final version accepted 25 September, 2020; available online 29 September, 2020.