SUPPLEMENTARY MATERIAL FOR "MINIMAL DISPERSION APPROXIMATELY BALANCING WEIGHTS: ASYMPTOTIC PROPERTIES AND PRACTICAL CONSIDERATIONS"

BY YIXIN WANG AND JOSÉ R. ZUBIZARRETA

A. PROOF FOR THE UNCONSTRAINED DUAL FORMULATION

Proof of Theorem 1

Proof. We first present a vanilla form of the dual.

LEMMA 1. *The dual of the optimization problem (*1*) is*

$$\begin{aligned} &\underset{\lambda}{\textit{minimize}} && l(\lambda) \\ &\textit{subject to} && \lambda \geq 0 \end{aligned}$$

where

$$l(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\{-Z_j n\rho(Q_j^\top \lambda) + Q_j^\top \lambda\} + \lambda^\top d,$$

$$A_{K\times n} = \begin{pmatrix} B_1(X_1) & B_1(X_2) & \dots & B_1(X_n) \\ \vdots & \vdots & \vdots & \vdots \\ B_K(X_1) & B_K(X_2) & \dots & B_K(X_n) \end{pmatrix}_{K\times n},$$

$$Q_{2K\times n} = \begin{pmatrix} A_{K\times n} \\ -A_{K\times n} \end{pmatrix}_{2K\times n},$$

and

$$d_{2K\times 1} = \begin{pmatrix} \delta_{K\times 1} \\ \delta_{K\times 1} \end{pmatrix}_{2K\times 1}.$$

We prove this lemma towards the end of this section.

We then write $\lambda_{2K\times 1} = \begin{pmatrix} \lambda_{+,K\times 1} \\ \lambda_{-,K\times 1} \end{pmatrix}_{2K\times 1}$. We have

$$\begin{aligned} l(\lambda) &= \frac{1}{n}\sum_{j=1}^{n}\{-Z_j n\rho(A_j^\top \lambda_+ - A_j^\top \lambda_-) + (A_j^\top \lambda_+ - A_j^\top \lambda_-)\} + \lambda_+^\top \delta + \lambda_-^\top \delta \\ &= \frac{1}{n}\sum_{j=1}^{n}[-Z_j n\rho\{A_j^\top(\lambda_+ - \lambda_-)\} + A_j^\top(\lambda_+ - \lambda_-)] + (\lambda_+^\top + \lambda_-^\top)\delta. \end{aligned}$$

Suppose the optimizer is $\lambda^\dagger_{2K\times 1} = \begin{pmatrix} \lambda^\dagger_{+,K\times 1} \\ \lambda^\dagger_{-,K\times 1} \end{pmatrix}_{2K\times 1}$. We claim that $\lambda^\dagger_{+,k}\cdot\lambda^\dagger_{-,k} = 0, k = 1,\dots,K$, where the index k points to the kth entry of a vector.

We prove this claim by contradiction. Suppose the opposite. If $\lambda^\dagger_{+,k} > 0$ and $\lambda^\dagger_{-,k} > 0$ for some k, then

$$\lambda^{\dagger\dagger\top} = [\lambda^\dagger_+ - \{0,\dots,0,\min(\lambda^\dagger_{+,k},\lambda^\dagger_{-,k}),0,\dots,0\}, \lambda^\dagger_- - \{0,\dots,0,\min(\lambda^\dagger_{+,k},\lambda^\dagger_{-,k}),0,\dots,0)\}]$$

has

$$l(\lambda^{\dagger\dagger}) = l(\lambda^\dagger) - 2\min(\lambda^\dagger_{+,k},\lambda^\dagger_{-,k})\cdot\delta < l(\lambda^\dagger)$$

by $\delta > 0$ and $\min(\lambda^{\dagger}_{+,k}, \lambda^{\dagger}_{-,k}) > 0$. This contradicts the fact that $\lambda^{\dagger}$ is the optimizer. Theorem 1 then follows by rewriting $\lambda_+ - \lambda_-$ as λ and deducing $\lambda_+ + \lambda_- = |\lambda|$ from $\lambda^{\dagger}_{+,k} \cdot \lambda^{\dagger}_{-,k} = 0, k = 1, \ldots, K$.□

Proof of Lemma 1

Proof. Rewriting problem (1) in matrix notation,

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \sum_{i=1}^{n} Z_i h(s_i) \\ \text{subject to} \quad & Q_{2K\times n} s_{n\times 1} \leq d_{2K\times 1} \end{aligned}$$

where

$$\begin{aligned} s_{n\times 1} &= (s_i)_{n\times 1} = (\frac{1}{n} - Z_i w_i)_{n\times 1}, \\ A_{K\times n} &= \begin{pmatrix} B_1(X_1) & B_1(X_2) & \ldots & B_1(X_n) \\ \vdots & \vdots & \vdots & \vdots \\ B_K(X_1) & B_K(X_2) & \ldots & B_K(X_n) \end{pmatrix}_{K\times n}, Q_{2K\times n} = \begin{pmatrix} A_{K\times n} \\ -A_{K\times n} \end{pmatrix}_{2K\times n}, \\ d_{2K\times 1} &= \begin{pmatrix} \delta_{K\times 1} \\ \delta_{K\times 1} \end{pmatrix}_{2K\times 1}. \end{aligned}$$

Again as special cases, stable balancing weights have $h(x) = (\frac{1}{n} - \frac{1}{r} - x)^2$ and entropy balancing has $h(x) = (\frac{1}{n} - x)\log(\frac{1}{n} - x)$.

The problem is now in the form of Tseng & Bertsekas (1987) and Tseng & Bertsekas (1991).

The dual of this problem is

$$\begin{aligned} \underset{\lambda}{\text{maximize}} \quad & g(\lambda) \\ \text{subject to} \quad & \lambda \geq 0, \end{aligned}$$

where $g(\lambda) = -\sum_{j=1}^{n} h_j^*(Q_j^{\top}\lambda) - < \lambda, d >$, and $h_j^*(\cdot)$ is the convex conjugate of $Z_j h(\cdot)$.

$$\begin{aligned} h_j^*(t) &= \sup_{s_j}\{ts_j - Z_j h(s_j)\} \\ &= \sup_{w_j}\{-tZ_j w_j + \frac{t}{n} - Z_j h(\frac{1}{n} - Z_j w_j)\} \\ &= \sup_{w_j}\{-tZ_j w_j + \frac{t}{n} - Z_j h(\frac{1}{n} - w_j)\} \\ &= -tZ_j w_j^* + \frac{t}{n} - Z_j h(\frac{1}{n} - w_j^*), \end{aligned}$$

where w_j^* satisfies the first order condition

$$\begin{aligned} & -tZ_j + Z_j h'(\frac{1}{n} - w_j^*) = 0, \\ \Rightarrow & h'(\frac{1}{n} - w_j^*) = t, \\ \Rightarrow & w_j^* = \frac{1}{n} - (h')^{-1}(t). \end{aligned}$$

Therefore,

$$\begin{aligned} h_j^*(t) &= -tZ_j\frac{1}{n} + tZ_j(h')^{-1}(t) + \frac{t}{n} - Z_j h\{(h')^{-1}(t)\}, \\ &= -Z_j[\frac{t}{n} - t(h')^{-1}(t) + h\{(h')^{-1}(t)\}] + \frac{t}{n}. \end{aligned}$$

Denote $\rho(\cdot)$ as

$$\rho(t) = \frac{t}{n} - t(h')^{-1}(t) + h\{(h')^{-1}(t)\}.$$

This gives

$$h_j^*(t) = -Z_j\rho(t) + \frac{t}{n}.$$

Also we notice that

$$\begin{aligned} \rho'(t) &= \frac{1}{n} - (h')^{-1}(t) - t\{(h')^{-1}(t)\}' + h'\{(h')^{-1}(t)\} \cdot \{(h')^{-1}(t)\}' \\ &= \frac{1}{n} - (h')^{-1}(t) - t\{(h')^{-1}(t)\}' + t\{(h')^{-1}(t)\}' \\ &= \frac{1}{n} - (h')^{-1}(t). \end{aligned}$$

This implies

$$w^* = \rho'(t).$$

The dual formulation thus becomes

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & l(\lambda) \\ \text{subject to} \quad & \lambda \geq 0 \end{aligned}$$

where

$$l(\lambda) = \frac{1}{n}\sum_{j=1}^{n}\{-Z_j n\rho(Q_j^\top\lambda) + Q_j^\top\lambda\} + \lambda^\top d.$$

B. Proof of the Asymptotic properties

Proof of Theorem 2

Proof. The proof utilizes the Bernstein's inequality as in Fan et al. (2016).
We first prove the following lemma.

LEMMA 2. *There exists a global minimizer $\lambda^\dagger$ such that*

$$||\lambda^\dagger - \lambda_1^*||_2 = O_p(K^{1/2}(\log K)/n + K^{1/2-r_\pi}).$$

Proof. Write $A_j = B(X_j) = \{B_1(X_j), ..., B_K(X_j)\}$. Recall that the optimization objective is

$$G(\lambda) := \frac{1}{n}\sum_{j=1}^{n}\{-Z_j n\rho(A_j^\top\lambda) + A_j^\top\lambda\} + |\lambda|^\top\delta,$$

where $G(\cdot)$ is convex in λ by the concavity of $\rho(\cdot)$. To show that a minimizer Δ^* of $G(\lambda_1^* + \Delta)$ exists in $\mathcal{C} = \{\Delta \in \mathbb{R}^K : ||\Delta||_2 \leq CK^{1/2}(\log K)/n + K^{1/2-r_\pi}\}$ for some constant C, it suffices to show that

$$E\{\inf_{\Delta\in\mathcal{C}} G(\lambda_1^* + \Delta) - G(\lambda_1^*) > 0\} \to 1, \text{ as } n \to \infty, (*)$$

by the continuity of $G(\cdot)$.

To show $(*)$, we use mean value theorem: for some $\tilde{\lambda}$ between $\lambda^\dagger$ and λ_1^*,

$$G(\lambda_1^* + \Delta) - G(\lambda_1^*) \tag{1}$$

$$\geq \Delta \cdot \frac{1}{n}\sum_{j=1}^{n}\{-Z_j n\rho'(A_j^\top \lambda_1^*)A_j + A_j\} + \frac{1}{2}\Delta^\top \cdot \{\sum_{j=1}^{n} -Z_j\rho''(A_j^\top \tilde{\lambda})A_j^\top A_j\} \cdot \Delta - |\Delta|^\top \delta \tag{2}$$

$$\geq -||\Delta||_2 \cdot ||\frac{1}{n}\sum_{j=1}^{n} -Z_j n\rho'(A_j^\top \lambda_1^*)A_j + A_j||_2 + \frac{1}{2}\Delta^\top \cdot \{\sum_{j=1}^{n} -Z_j\rho''(A_j^\top \tilde{\lambda})A_j^\top A_j\} \cdot \Delta - ||\Delta||_2||\delta||_2 \tag{3}$$

$$\geq -||\Delta||_2 \cdot ||\frac{1}{n}\sum_{j=1}^{n} -Z_j n\rho'(A_j^\top \lambda_1^*)A_j + A_j||_2 - ||\Delta||_2||\delta||_2. \tag{4}$$

The first inequality is due to the triangle inequality, $|\lambda_1^* + \Delta| - |\lambda_1^*| \geq -|\Delta|$. The second inequality follows from Cauchy-Schwarz inequality. The third inequality is due to the positivity of $\frac{1}{2}\Delta^\top \cdot \{\sum_{j=1}^{n} -Z_j\rho''(A_j^\top \tilde{\lambda})A_j^\top A_j\} \cdot \Delta$ by Assumption 1.3.

Next we notice that

$$\begin{aligned}
&||\frac{1}{n}\sum_{j=1}^{n}\{-Z_j n\rho'(A_j^\top \lambda_1^*)A_j + A_j\}||_2 \\
\leq &||\frac{1}{n}\sum_{j=1}^{n}(-Z_j\frac{1}{\pi_j}A_j + A_j)||_2 + ||\frac{1}{n}\sum_{j=1}^{n} -Z_j\{\frac{1}{\pi_j} - n\rho'(A_j^\top \lambda_1^*)\}A_j||_2 \\
\leq &||\frac{1}{n}\sum_{j=1}^{n}(1 - \frac{Z_j}{\pi_j})A_j||_2 + \frac{1}{n}\sum_{j=1}^{n}||A_j||_2 O(K^{-r_\pi}).
\end{aligned}$$

The first inequality is due to the triangle inequality. The second inequality is due to Assumption 1.3 and 1.6.

We first use the Bernstein's inequality to bound both terms.

Recall that the Bernstein's inequality for random matrices in Tropp et al. (2015) says the following. Let $\{Z_k\}$ be a sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume that $EZ_k = 0$ and $||Z_k||_2 \leq R_n$ almost surely. Define

$$\sigma_n^2 = \max\{||\sum_{k=1}^{n} E(Z_k Z_k^\top)||_2, ||\sum_{k=1}^{n} E(Z_k^\top Z_k)||_2\}.$$

Then for all $t \geq 0$,

$$pr(||\sum_{k=1}^{n} Z_k||_2 \geq t) \leq (d_1 + d_2)\exp(-\frac{t^2/2}{\sigma_n^2 + R_n t/3}).$$

For the first term $||\frac{1}{n}\sum_{j=1}^{n}(1 - \frac{Z_j}{\pi_j})A_j||_2$, we notice that

$$E\{\frac{1}{n}(1 - \frac{Z_j}{\pi_j})A_j\} = E[E\{\frac{1}{n}(1 - \frac{Z_j}{\pi_j})A_j \mid X_j\}] = 0. \tag{5}$$

The last equality is because $E(Z_j) = \pi_j$.

Then for $||\frac{1}{n}\sum_{j=1}^{n}(1-Z_j/\pi_j)A_j||_2$, we have

$$||\frac{1}{n}(1-\frac{Z_j}{\pi_j})A_j||_2 \tag{6}$$
$$\leq\frac{1}{n}||(1-\frac{Z_j}{\pi_j})||_2||A_j||_2 \tag{7}$$
$$\leq\frac{1}{n}(\frac{1-\pi_j}{\pi_j})CK^{1/2} \tag{8}$$
$$=\frac{1}{n}\{n\rho'(A_j^\top\lambda_1^*)-1\}CK^{1/2} \tag{9}$$
$$\leq C'\frac{K^{1/2}}{n}. \tag{10}$$

The first inequality is due to Cauchy-Schwarz inequality. The second inequality is due to Assumption 1.4 and $E(1-Z_j/\pi_j)^2=\text{var}(1-Z_j/\pi_j)=\pi_j(1-\pi_j)/\pi_j^2=(1-\pi_j)/\pi_j$. The third equality is due to $\pi_j=\{n\rho'(A_j^\top\lambda_1^*)\}^{-1}$. The fourth inequality is due to Assumption 1.3.

Finally, for $||\sum_{k=1}^{n}E\{\frac{1}{n^2}(1-\frac{Z_j}{\pi_j})^2A_jA_j^\top\}||_2$, we have

$$||\sum_{k=1}^{n}E\{\frac{1}{n^2}(1-\frac{Z_j}{\pi_j})^2A_jA_j^\top\}||_2 \tag{11}$$
$$\leq\frac{1}{n}\sup_j(1-\frac{Z_j}{\pi_j})^2||E(A_jA_j^\top)||_2 \tag{12}$$
$$\leq\frac{C''}{n}. \tag{13}$$

The first inequality is taking the sup over $(1-\frac{Z_j}{\pi_j})^2$. The second inequality is due to Assumption 1.3, 1.4, and $\pi_j=\{n\rho'(A_j^\top\lambda_1^*)\}^{-1}$.

Equation (5), Equation (10), and Equation (13), together with the Bernstein's inequality, imply

$$\text{pr}\{||\frac{1}{n}\sum_{j=1}^{n}(1-\frac{Z_j}{\pi_j})A_j||_2\geq t\}\leq(K+1)\exp(-\frac{t^2/2}{\frac{C''}{n}+C'\frac{K^{1/2}}{n}\cdot t/3}). \tag{14}$$

The right side goes to zero as $K\to\infty$ when

$$\frac{t^2/2}{\frac{C''}{n}+C'\frac{K^{1/2}}{n}\cdot t/3}\geq\log K.$$

It suffices when $t=O_p\{K^{1/2}(\log K)/n\}$.

Therefore, we have

$$||\frac{1}{n}\sum_{j=1}^{n}(1-\frac{Z_j}{\pi_j})A_j||_2=O_p\{K^{1/2}(\log K)/n\}. \tag{15}$$

Now we work on the second term $\frac{1}{n}\sum_{j=1}^{n}||A_j||_2O(K^{-r_\pi})$. We have

$$\frac{1}{n}\sum_{j=1}^{n}||A_j||_2O(K^{-r_\pi})\leq CK^{1/2-r_\pi}. \tag{16}$$

This inequality is due to Assumption 1.4.

Combining Equation (15), Equation (16), and Assumption 1.7, we have

$$\begin{aligned}
&G(\lambda_1^* + \Delta) - G(\lambda_1^*) \\
=&- ||\Delta||_2 \cdot O_p(\frac{K^{1/2}\log K}{n} + K^{1/2-r_\pi}) + \frac{1}{2}||\Delta||_2^2||\delta||_2 \\
\geq& 0
\end{aligned}$$

for $\Delta = C\frac{K^{1/2}\log K}{n} + K^{1/2-r_\pi}$ with large enough constant $C > 0$.

$(*)$ is thus proved. □

Now we prove Theorem 2.

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} |nw^*(x) - \frac{1}{\pi(x)}| \\
=&\sup_{x\in\mathcal{X}} |n\rho'\{B(x)^\top\lambda^\dagger\} - n\rho'\{m^*(x)\}| \\
\leq&\sup_{x\in\mathcal{X}} |n\rho'\{B(x)^\top\lambda^\dagger\} - n\rho'\{B(x)^\top\lambda_1^*\}| + \sup_{x\in\mathcal{X}} |n\rho'\{B(x)^\top\lambda_1^*\} - n\rho'\{m^*(x)\}| \\
=&O\{\sup_{x\in\mathcal{X}} |B(x)^\top\lambda^\dagger - B(x)^\top\lambda_1^*|\} + O(K^{-r_\pi}) \\
\leq&O\{\sup_{x\in\mathcal{X}} ||B(x)||_2||\lambda^\dagger - \lambda_1^*||_2\} + O(K^{-r_\pi}) \\
=&O_p\{K(\frac{\log K}{n} + K^{-r_\pi})\} + O(K^{-r_\pi}) \\
=&O_p(\frac{K\log K}{n} + K^{1-r_\pi}) \\
=&o_p(1)
\end{aligned}$$

The first equality rewrites $\pi(x) = \{n\rho'(B(x)^\top\lambda_1^*)\}^{-1}$. The second inequality is due to the triangle inequality. The third inequality is due to Assumptions 1.3 and 1.6. The fourth inequality is due to the Cauchy-Schwarz inequality. The fifth equality is due to Lemma 2 and Assumption 1.4. The sixth equality holds because the first term dominates the second. The seventh equality is due to Assumptions 1.5 and 1.6.

Also, we have

$$\begin{aligned}
&||nw^*(x) - \frac{1}{\pi(x)}||_{P,2} \\
=&||n\rho'\{\lambda^{\dagger\top}B(X)\} - \frac{1}{\pi(x)}||_{P,2} \\
\lesssim&||n\rho'\{\lambda^{\dagger\top}B(X)\} - n\rho'\{\lambda_1^{*\top}B(X)\}||_{P,2} + ||\frac{1}{\pi(x)} - n\rho'\{\lambda_1^{*\top}B(X)\}||_{P,2} \\
\lesssim&||(\lambda^\dagger - \lambda_1^*)^\top B(X)||_{P,2} + \sup_{x\in\mathcal{X}} |m^*(x) - \lambda_1^{*\top}B(x)| \\
=&O_p\{K^{1/2}(\frac{\log K}{n} + K^{-r_\pi})\} + O(K^{-r_\pi}) \\
=&O_p(\frac{K^{1/2}\log K}{n} + K^{1/2-r_\pi}) \\
=&o_p(1)
\end{aligned}$$

The first equality rewrites $\pi(x) = [n\rho'\{B(x)^\top\lambda_1^*\}]^{-1}$. The second inequality is due to the triangle inequality. The third inequality is due to Assumption 1.3. The fourth inequality is due to Lemma 2, As-

sumption 1.4 and Assumption 1.6. The fifth equality is due to the first term dominates the second. The sixth equality is due to Assumption 1.5 and Assumption 1.6.

Proof of Theorem 3

Proof. The proof utilizes empirical processes techniques as in Fan et al. (2016).

We first decompose $\hat{Y}_{w^*} - \bar{Y}$ into several residual terms.

$$\begin{aligned}
\hat{Y}_{w^*} - \bar{Y} &= \sum_{i=1}^{n} Z_i w_i^* Y_i - \bar{Y} \\
&= \sum_{i=1}^{n} Z_i w_i^* \{Y_i - Y(X_i)\} + \sum_{i=1}^{n} (Z_i w_i^* - \frac{1}{n}) Y(X_i) + \{\frac{1}{n}\sum_{i=1}^{n} Y(X_i) - \bar{Y}\} \\
&= \sum_{i=1}^{n} Z_i w_i^* \{Y_i - Y(X_i)\} + \sum_{i=1}^{n} (Z_i w_i^* - \frac{1}{n}) \{Y(X_i) - \lambda_2^{*\top} B(X_i)\} \\
&+ \sum_{i=1}^{n} (Z_i w_i^* - \frac{1}{n}) \lambda_2^{*\top} B(X_i) + \{\frac{1}{n}\sum_{i=1}^{n} Y(X_i) - \bar{Y}\} \\
&= \frac{1}{n}\sum_{i=1}^{n} S_i + R_0 + R_1 + R_2,
\end{aligned}$$

where

$$\begin{aligned}
S_i &= \frac{Z_i}{\pi_i}\{Y_i - Y(X_i)\} + \{Y(X_i) - \bar{Y}\}, \\
R_0 &= \sum_{i=1}^{n} (w_i^* - \frac{1}{n\pi_i}) Z_i \{Y_i - Y(X_i)\}, \\
R_1 &= \sum_{i=1}^{n} (Z_i w_i^* - \frac{1}{n}) \{Y(X_i) - \lambda_2^{*\top} B(X_i)\}, \\
R_2 &= \sum_{i=1}^{n} (Z_i w_i^* - \frac{1}{n}) \{\lambda_2^{*\top} B(X_i)\}.
\end{aligned}$$

Below we show $R_j = o_p(n^{-1/2}), 0 \le j \le 2$. The conclusion follows from S_i taking the same form as the efficient score (Hahn, 1998). $\hat{Y}_{w^*}$ is thus asymptotically normal and semiparametrically efficient.

We first study $R_0 = \sum_{i=1}^{n} (n w_i^* - 1/\pi_i) Z_i \{Y_i - Y(X_i)\}/n$. Consider an empirical process $\mathbb{G}_n(f_0) = n^{1/2}[\sum_{i=1}^{n} f_0(Z_i, Y_i, X_i)/n - E\{f_0(Z, Y, X)\}]$, where

$$f_0(Z, Y, X) = Z\{Y - Y(X)\}\left[n\rho'\{m(X)\} - \frac{1}{\pi(x)}\right].$$

By the missing at random assumption, we have $E f_0\{Z, Y(1), X\} = 0$.

By Theorem 2, we have

$$\sup_{x \in \mathcal{X}} |\rho'\{B(x)^\top \lambda^\dagger\} - \frac{1}{n\pi(x)}| = O_p(\frac{K \log K}{n} + K^{1-r_\pi}) = o_p(1).$$

By Markov's inequality and maximal inequality, we have

$$n^{1/2} R_0 \le \sup_{f_0 \in \mathcal{F}} \mathbb{G}_n(f_0) \lesssim E \sup_{f_0 \in \mathcal{F}} \mathbb{G}_n(f_0) \lesssim J_{[]}\{||F_0||_{P,2}, \mathcal{F}, L_2(P)\},$$

where the set of functions is $\mathcal{F} = \{f_0 : ||m - m^*||_\infty \le \delta_0\}$, where $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ and $\delta_0 = C\{K(\log K)/n + K^{1-r_\pi}\}$ for some constant $C > 0$.

The second inequality is due to Markov's inequality. $J_{[]}\{||F_0||_{P,2}, \mathcal{F}, L_2(P)\}$ is the bracketing integral. $F_0 := \delta_0|Y - Y(X)| \gtrsim |f_0(Z, Y, X)|$ is the envelop function. We also have $||F_0||_{P,2} = (EF_0^2)^{1/2} \lesssim \delta_0$ by $E|Y - Y(X)| < \infty$.

Next we bound $J_{[]}\{||F_0||_{P,2}, \mathcal{F}, L_2(P)\}$ by $n_{[]}\{\varepsilon, \mathcal{F}, L_2(P)\}$:

$$J_{[]}\{||F_0||_{P,2}, \mathcal{F}, L_2(P)\} \lesssim \int_0^\delta [n_{[]}\{\varepsilon, \mathcal{F}, L_2(P)\}]^{1/2} \mathrm{d}\varepsilon.$$

Define a new set of functions $\mathcal{F}_0 = \{f_0 : ||m - m^*||_\infty \leq C\}$ for some constant $C > 0$, because need a constant different than δ_0 as $\delta_0 \to 0$ can change. Then,

$$\begin{aligned}
\log n_{[]}\{\varepsilon, \mathcal{F}, L_2(P)\} &\lesssim \log n_{[]}\{\varepsilon, \mathcal{F}_0\delta_0, L_2(P)\} \\
&= \log n_{[]}\{\varepsilon/\delta_0, \mathcal{F}_0, L_2(P)\} \\
&\lesssim \log n_{[]}\{\varepsilon/\delta_0, \mathcal{M}, L_2(P)\} \\
&\lesssim (\delta_0/\varepsilon)^{(1/k_1)}.
\end{aligned}$$

The first inequality is due to the fact that $\rho'(\cdot)$ bounded away from 0 and Lipschitz. The last inequality is due to Assumption 2.2.

Therefore, we have

$$J_{[]}\{||F_0||_{P,2}, \mathcal{F}, L_2(P)\} \lesssim \int_0^\delta [\log n_{[]}\{\varepsilon, \mathcal{F}, L_2(P)\}]^{1/2} \mathrm{d}\varepsilon \lesssim \int_0^\delta (\delta_0/\varepsilon)^{(1/2k_1)} \mathrm{d}\varepsilon.$$

This goes to 0 as δ goes to 0 by $2k_1 > 1$ and the integral converges. Thus, this shows that $n^{1/2}R_0 = o_p(1)$.

Next, we consider $R_1 = \sum_{i=1}^n (nZ_i w_i^* - 1)\{Y(X_i) - \lambda_2^{*\top} B(X_i)\}/n$. Define the empirical process $\mathbb{G}_n(f_1) = n^{1/2}[\sum_{i=1}^n f_1(Z_i, X_i)/n - E\{f_1(Z, X)\}]$, where $f_1(Z, X) = [nZ\rho'\{m(x)\} - 1]\{Y(X) - \lambda_2^{*\top} B(X)\}$.

Write $\Delta(X) := Y(X) - \lambda_2^{*\top} B(X)$. By Assumption 2.3, we have $||\Delta||_\infty \lesssim K^{-r_y}$.

By Theorem 2, we have

$$||n\rho'\{\lambda^{\dagger\top} B(X)\} - \frac{1}{\pi(x)}||_{P,2} = O_p(\frac{K \log K}{n} + K^{1-r_\pi}).$$

Therefore, we have

$$\begin{aligned}
n^{1/2}R_1 &= \mathbb{G}_n(f_1) + n^{1/2} E f_1(Z, X) \\
&\leq \sup_{f_1 \in \mathcal{F}_1} \mathbb{G}_n(f_1) + n^{1/2} \sup_{f_1 \in \mathcal{F}_1} E f_1,
\end{aligned}$$

where $\mathcal{F}_1 = \{f_1 : ||m - m^*||_{P,2} \leq \delta_1, ||\Delta||_\infty \leq \delta_2\}$,$\delta_1 = C\{K^{1/2}(\log K)/n + K^{1/2-r_\pi}\}$, $\delta_2 = CK^{-r_y}$ for some constant $C > 0$.

Again, by Markov's inequality and the maximal inequality,

$$\sup_{f_1 \in \mathcal{F}_1} \mathbb{G}_n(f_1) \lesssim E \sup_{f_1 \in \mathcal{F}_1} \mathbb{G}_n(f_1) \lesssim J_{[]}\{||F_1||_{P,2}, \mathcal{F}, L_2(P)\},$$

where $F_1 := C\delta_2$ for some constant $C > 0$ so that $||F_1||_{P,2} \lesssim \delta_2$.

Similar to characterizing R_1, we we bound $J_{[]}(||F_1||_{P,2}, \mathcal{F}_1, L_2(P)$ by $n_{[]}(\varepsilon, \mathcal{F}_1, L_2(P))$:

$$J_{[]}(||F_1||_{P,2}, \mathcal{F}_1, L_2(P) \lesssim \int_0^\delta \{n_{[]}(\varepsilon, \mathcal{F}_1, L_2(P))\}^{1/2} \mathrm{d}\varepsilon.$$

Then, we bound $n_{[]}(\varepsilon, \mathcal{F}_1, L_2(P))$:

$$\begin{aligned}\log n_{[]}\{\varepsilon, \mathcal{F}_1, L_2(P)\} &\lesssim \log n_{[]}\{\varepsilon/\delta_2, \mathcal{F}_0, L_2(P)\} \\ &\lesssim \log n_{[]}\{\varepsilon/\delta_2, G_{10}, L_2(P)\} + \log n_{[]}\{\varepsilon/\delta_2, G_{20}, L_2(P)\} \\ &\lesssim \log n_{[]}\{\varepsilon/\delta_2, \mathcal{M}, L_2(P)\} + \log n_{[]}\{\varepsilon/\delta_2, \mathcal{H}, L_2(P)\} \\ &\lesssim (\delta_1/\varepsilon)^{1/k_1} + (\delta_2/\varepsilon)^{1/k_2}.\end{aligned}$$

where

$$\mathcal{F}_0 = \{f_1 : ||m - m^*||_{P,2} \leq C, ||\Delta||_{P,2} \leq 1\},$$

$$\mathcal{G}_{10} = \{m \in \mathcal{M} + m^* : ||m||_{P,2} \leq C\},$$

$$\mathcal{G}_{20} = \{\Delta \in \mathcal{H} - \lambda_2^{*\top} B(x) : ||\Delta||_{P,2} \leq 1\}.$$

The second inequality is due to ρ' is Lipschitz and bounded away from 0.
Therefore we have

$$J_{[]}\{||F_1||_{P,2}, \mathcal{F}_1, L_2(P)\} \lesssim \int_0^\delta (\delta_1/\varepsilon)^{(1/2k_1)} d\varepsilon + \int_0^\delta (\delta_2/\varepsilon)^{(1/2k_2)} d\varepsilon.$$

By $2k_1 > 1$ and $2k_2 > 1$, we have $J_{[]}\{||f_1||_{P,2}, \mathcal{F}, L_2(P)\} = o(1)$. This gives $\sup_{f_1 \in \mathcal{F}_1} \mathbb{G}_n(f_1) = o_p(1)$.
Now we look at $n^{1/2} \sup_{f_1 \in \mathcal{F}_1} Ef_1$,

$$\begin{aligned}n^{1/2} \sup_{f_1 \in \mathcal{F}} Ef_1 &= n^{1/2} \sup_{m \in \mathcal{G}_1, \Delta \in \mathcal{G}_2} E\{\pi(X)[n\rho'\{m(X)\} - 1]\Delta(X)\} \\ &= n^{1/2} \sup_{m \in \mathcal{G}_1, \Delta \in \mathcal{G}_2} E([n\rho'\{m(x)\} - \frac{1}{\pi(x)}]\pi(x)\Delta(x)) \\ &\lesssim n^{1/2} \sup_{m \in \mathcal{G}_1} ||n\rho'\{m(x)\} - \frac{1}{\pi(x)}||_{P,2} \sup_{\Delta \in \mathcal{G}_2} ||\Delta(x)||_{P,2} \\ &\lesssim n^{1/2}\delta_1\delta_2 = o_p(1),\end{aligned}$$

where $\mathcal{G}_1 = \{m \in \mathcal{M} : ||m - m^*||_{P,2} \leq \delta_1\}, \mathcal{G}_2 = \{\Delta \in \mathcal{H} - \lambda_2^{*\top} B(x) : ||\Delta||_\infty \leq \delta_2\}$.
The last equality is due to the assumption $n^{1/2} \lesssim K^{r_\pi + r_y - 1/2}$.
Therefore, we can conclude that $n^{1/2} R_1 = o_p(1)$.
Lastly, $R_2 = \lambda_2^{*\top}\{\sum_{i=1}^n (Z_i w_i^* - 1/n) B(X_i)\} = o_p(1)$ by $\sum_{i=1}^n (Z_i w_i^* - 1/n) B(X_i) \leq ||\delta||^2 = o_p(1)$ due to the constraints posited in the optimization problem for minimal weights.
We finally prove the consistency of the variance estimator. We need a stronger smoothness assumption, i.e. $r_y > 1$.
Under assumptions 1 and 2, we construct a variance estimator based on a direct approximation of the efficient influence function. Recall that the efficient influence function determines the semiparametric efficiency bound (Hahn, 1998):

$$\begin{aligned}V_{opt} &:= var(Y(X_i)) + E\{var(Y|X_i)/\pi(X_i)\} \\ &= E\left\{\left(\frac{Z_i Y_i}{\pi(X_i)} - \bar{Y} - Y(X_i)(\frac{Z_i}{\pi(X_i)} - 1)\right)^2\right\}.\end{aligned}$$

We estimate V_{opt} with $\hat{V}_K$:

$$\hat{V}_K = \frac{1}{n}\sum_{i=1}^{n}\Bigg[nZ_i w_i Y_i - \sum_{i=1}^{n} w_i Y_i - B(X_i)^\top \left\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\right\}^{-1} \left\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top Y_i\right\}(nZ_i w_i - 1)\Bigg]^2.$$

In particular, $\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top Y_i\}$ is a least square estimator of $Y(X_i)$.

To show $\hat{V}_K$ is consistent with V_{opt}, it is sufficient to show

$$|B(X_i)^\top\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top Y_i\} - Y(X_i)| \xrightarrow{a.s.} 0. \qquad (**)$$

This is because nw_i is a consistent estimator of $1/\pi(X_i)$ by Theorem 2 and $\sum_{i=1}^{n} w_i Y_i$ is a consistent estimator of $\bar{Y}$ by Theorem 3.

Below we prove $(**)$.

We first rewrite Y_i as $Y_i = B(X_i)^\top \lambda_2^* + \gamma + \epsilon_i$, where $\gamma = O(K^{-r_y})$ from Assumption 2.3, and ϵ_i is some iid zero mean error with variance $\sigma^2 = var(Y|X_i)$. Therefore,

$$\begin{aligned}
&\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top Y_i\}\\
=&\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}[\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top \{B(X_i)^\top \lambda_2^* + \gamma + \epsilon_i\}]\\
=&\lambda_2^* + \{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top\}\gamma\\
&+ \{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top \epsilon_i\}\\
=&\lambda_2^* + E\{Z_i w_i B(X_i)^\top B(X_i)\}^{-1} E\{Z_i w_i B(X_i)^\top\}\{\gamma + E(\epsilon_i)\} + O_p(n^{-1/2})\\
=&\lambda_2^* + O_p(K^{-r_y+1/2}).
\end{aligned}$$

The last equality is due to assumptions 1.4 and 2.3 and the law of large numbers.

Finally we have

$$\begin{aligned}
&B(X_i)^\top\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top B(X_i)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} Z_i w_i B(X_i)^\top Y_i\}\\
=&B(X_i)^\top \lambda_2^* + B(X_i)\cdot O_p(K^{-r_y+1/2})\\
=&Y(X_i) + B(X_i)\cdot O_p(K^{-r_y+1/2}) + O_p(K^{-r_y})\\
=&Y(X_i) + o_p(1)
\end{aligned}$$

The last equality is due to assumption 1.4 and the additional assumption $r_y > 1$. □

C. THEOREM 4 EXPLAINED

Due to the connection to shrinkage estimation, for each basis function that we balance, we implicitly include a corresponding term in the inverse propensity score model. In practice, we are often concerned about the estimation loss due to fitting an overly complex model. In the context of minimal weights, an overly complex model corresponds to balancing more terms than are needed.

Theorem 4 is an oracle inequality that bounds this loss and states that approximate balancing—as opposed to exact balancing—mimics the act of upper bounding the number of effective balancing constraints. Hence, minimal weights do not suffer much from excessive balancing when few constraints are active. We also remark that this sparsity assumption on the balancing constraints is commonly satisfied in real data sets. This is exemplified by the sparsity of the shadow prices in the 2010 Chilean post earthquake survey data; see Figure 1 of Zubizarreta (2015).

The oracle inequality we proved in Section 4 leverages an oracle inequality for lasso in the high dimensional generalized linear model literature (Van de Geer, 2008). The original oracle inequality says the lasso estimator (with ℓ^1 penalty) under general Lipschitz losses behaves similarly to the estimator with ℓ^0 penalty, if the true generalized linear model is sparse.

Recall that the minimal weights compute

$$\lambda^{\dagger} := \arg\min G(\lambda) = \arg\min \sum_{j=1}^{n} \left\{ -Z_j \rho(A_j^{\top}\lambda) + A_j^{\top}\lambda \cdot \frac{1}{n} \right\} + |\lambda|^{\top}\delta.$$

This is a lasso estimator under the loss function

$$L_w(x,z) = -z \cdot n(\rho \circ (\rho')^{-1} \circ w)(x) + ((\rho')^{-1} \circ w)(x),$$

where the fit for w is $\hat{w}(x) = \rho'(B(x)^{\top}\hat{\lambda})$. This loss function is the same loss function as in Equation (4) but written as a function of w. Correspondingly, the empirical loss is

$$\sum_{n}^{i=1} L_w(X_i, Z_i) = \frac{1}{n} \sum_{i=1}^{n} \left\{ -Z_i n \cdot (\rho \circ (\rho')^{-1} \circ w)(X_i) + ((\rho')^{-1} \circ w)(X_i) \right\},$$

and the theoretical risk is

$$\begin{aligned} EL_w(X,Z) =& \frac{1}{n} \sum_{i=1}^{n} E\left\{ -Z_i n \cdot (\rho \circ (\rho')^{-1} \circ w)(X_i) + ((\rho')^{-1} \circ w)(X_i) \right\} \\ =& \frac{1}{n} \sum_{i=1}^{n} \left\{ -\pi(X_i) \cdot n(\rho \circ (\rho')^{-1} \circ w)(X_i) + ((\rho')^{-1} \circ w)(X_i) \right\}. \end{aligned}$$

We define the target w^0 as the minimizer of the theoretical risk

$$w^0(x) := \arg\min EL_w(X,Z) = \frac{1}{n\pi(x)}.$$

The last equality is due to setting $\partial EL_w(X,Z)/\partial w = 0$. This is the true inverse propensity score function used for inverse probability weights. We are interested in studying the excess risk of estimators

$$\mathcal{E}(w) := E\{L_w(X,Z) - L_{w^0}(X,Z)\}.$$

For simplicity of notation, we write $w_\lambda(x) = \rho'(B(x)^{\top}\lambda), \lambda \in \mathbb{R}^K$. Approximate balancing weights thus perform the empirical risk minimization of

$$\lambda^{\dagger} := \arg\min_{\lambda} \{ \frac{1}{n} \sum_{i=1}^{n} L_{w_\lambda}(X_i, Z_i) + |\lambda|^{\top}\delta \}.$$

We look at the case of $\delta = \delta^{+}(\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_K)$, for some $\delta^{+} > 0$, where $\hat{\sigma}_k$ is the (sample) standard error of $B_k(X), k = 1, ..., K$. This aligns closely with the common way of setting δ; we specify approximate balancing constraints in units of the standard error of each covariate.

We consider the following oracle estimator

$$\lambda^{*} := \arg\min_{\lambda} \{ EL_{w_\lambda}(X,Z) + ||\lambda||_0 \cdot C_0 \},$$

for some constant $C_0 > 0$. λ^* can also be seen as the minimizer of $\mathbb{P}L_{w_\lambda}$ under the constraint that $||\lambda||_0 \leq C_1$ for some $C_1 > 0$.

$||\lambda||_0$ is the number of nonzero entries of λ. This is also the number of active or effective covariate balancing constraints in the optimization problem (1). In this sense, the oracle estimator roughly performs the same covariate balancing exactly as its approximate counterpart $\lambda^\dagger$ but the number of effective constraints being capped by some constant.

We now assume the following conditions hold and present the oracle inequality.

Assumption 3. The following conditions hold:

1. There exist constants $0 < c_0 < 1/2$, such that $c_0 \leq n\rho'(v) \leq 1 - c_0$ for any $v = B(x)^\top \lambda$ with $\lambda \in int(\Theta)$. Also, there exist constants $c_1 < c_2 < 0$, such that $c_1 \leq n\rho''(v) \leq c_2 < 0$ in some small neighborhood $\mathcal{B}$ of $v^* = B(x)^\top \lambda^\dagger$.
2. $\epsilon_0 < \pi(x) < 1 - \epsilon_0, \forall x$, for some constant $0 < \epsilon_0 < 1$,
3. $M := \max ||B_k(x)||_\infty / \sigma_k < \infty$, where σ_k is the (population) standard deviation of $B_k(X), k = 1, ..., K$.

Commenting on the previous assumptions, Assumption 3.1 is similar to Assumption 1.3. Assumption 3.2 is similar to the overlap condition of propensity scores. Both of them ensure the quadratic margin condition required by the lasso oracle inequality (the quadratic margin condition says in the ℓ^∞ neighborhood of w^0 the excess risk $\mathcal{E}$ is bounded from below by a quadratic function). Assumption 3.3 is similar to Assumption 1.4. It ensures the existence of the constant $\bar{\lambda} > 0$ in the theorem.

We further assume the following technical conditions.

Assumption 4. Assume the following technical conditions hold.

1. There exists $\eta > 0$ such that $n||w_{\lambda^*} - w^0||_\infty \leq \eta$ and $n||w_{\tilde{\lambda} - w^0}||_\infty \leq \eta$, where $\tilde{\lambda} = \arg\min_{\lambda \in \Theta: \sum_k \sigma_k |\lambda - \lambda^*| \leq 9\mathcal{E}(w_{\lambda^*}) + 675\bar{\lambda}^2 ||\lambda^*||_0} \{\mathcal{E}(w_\lambda) - 15\bar{\lambda} \sum_{k:\lambda^* \neq 0} \sigma_k |\lambda - \lambda^*|\}$,
2. $\{\log(2K)\}^{1/2} n^{-1/2} M \leq 0.13$,
3. $a_n := \{2\log(2K)\}^{1/2} M n^{-1/2} + \log(2K) M n^{-1/2}$,
4. For some $t > 0$ we are free to set, $\bar{\lambda} := 4a_n(1 + t\{2(1 + 8a_n M)\}^{1/2} + 8t^2 a_n M/3) > 6.4\{\log(2K)\}^{1/2} n^{-1/2}$,
5. $s > 0$ solves $a_n(1 + s\{2(1 + 2a_n M)\}^{1/2} + 2t^2 a_n M/3) = 9/5$,
6. $\alpha = \exp(-n a_n^2 s^2) + 7\exp(-4n a_n t^2)$.

The technical assumptions are inherited from Theorem 2.2 of Van de Geer (2008).

The first technical assumption is needed because the quadratic margin condition $\mathcal{E}(nw_\lambda) \geq cn||w_\lambda - w^0||^2$ only holds locally for w_η within the η neighborhood of w^0, $||w_\lambda - w^0||_\infty \leq \eta/n$. The estimator $\tilde{\lambda}$ strikes the balance between how much excess risk it incurs and how different it is from the oracle estimator λ^* in the ℓ^1 neighborhood of λ^*.

The second technical assumption is to ensure the applicability of Bousquet's inequality to the empirical process induced by Z conditional on X. The constant 0.13 is rather arbitrary; it could be replaced by any constant smaller that $(\sqrt{6} - \sqrt{2})/2$ if other constants are adjusted accordingly.

The third technical assumption on a_n is due to the usual rate of decay in probability for Gaussian linear model with orthogonal design, resulting from a symmetrization inequality and a contraction inequality.

The fourth technical assumption on $\bar{\lambda}$ is setting a lower bound for the smoothing parameter. It follows from the Bousquet's inequality. t is a parameter to be set by users; we need to strike the balance between small excess risk due to small t and large confidence in the upper bound for excess risk due to large t.

The fifth technical condition on s is due to the contraction inequality for the additional randomness in standard error of covariates $\hat{\sigma}_k$ relative to the true standard deviation σ_k.

The sixth technical condition on α defines "with high probability" as with probability $1 - \alpha$ where α decays exponentially in n.

With these assumptions, we have the following theorem.

THEOREM 5. *Under Assumption* 3 *and Assumption* 4*, with probability at least* $1-\alpha$*, we have*

$$\mathcal{E}(w_{\lambda^\dagger}) \leq 3\mathcal{E}(nw_{\lambda^*}) + 225\bar{\lambda}^2||\lambda^*||_0,$$

and

$$\sum_k \sigma_k|\lambda_k^\dagger - \lambda_k^*| \leq \frac{21}{4}\bar{\lambda}\mathcal{E}(nw_{\lambda^*}) + \frac{1575\bar{\lambda}}{4}||\lambda^*||_0,$$

where $\bar{\lambda} > 0$ *is a constant that depends on* K.

Theorem 4 in Section 4·1 is a consequence of Theorem 5 and Assumption 1.

Theorem 5 is a consequence of Theorem 2.2 in Van de Geer (2008) where the oracle properties for lasso estimators are established under general convex loss. We only need to show that the assumptions for Theorem 5.1 imply the assumptions of Theorem 2.2 in Van de Geer (2008) so that their conclusion applies.

When there are few active covariate balancing constraints, $||\lambda^*||_0$ will be small. The theorem then says that the excess risk of minimal weights is of the same order as the oracle estimator. Therefore, minimal weights mimics the exact balancing weights under a capped number of effective constraints. In other words, resorting to approximation in covariate balancing enjoys a similar effect of capping the number of effective balancing constraints. Hence, minimal weights is immune to the loss of excessive balancing.

An important practical question is how many covariates we should balance. Exact balancing weights can only balance a few covariates, because otherwise the problem does not admit a solution. Minimal weights relieve this problem: we can balance much more covariates with δ appropriately set. This oracle inequality says that we do not need to worry about excessive balancing. We only need to find a sweet spot between balancing many covariates loosely and balancing a few covariates strictly. This amounts to setting δ appropriately, which we address in Section 4.

Below we prove Theorem 5.

Proof. We only need to show assumptions L, B, and C in Theorem 2.2 of Van de Geer (2008) so that their oracle inequality applies to minimal weights.

First we show assumption L: the loss function is convex and Lipschitz. The loss function writes $L_w(x,z) = -z \cdot (\rho \circ (\rho')^{-1} \circ nw)(x) + ((\rho')^{-1} \circ w)(x)$. Fixing z, we have

$$\frac{\partial L_w(x,z)}{\partial w} = \{-z \cdot nw(x) + 1\}\{-\frac{\rho''}{n(\rho')^2}(nw(x))\}.$$

This is bounded due to assumptions 3.1 and 3.3, implying the Lipschitz property: derivatives of ρ and bounded, z is bounded by [0,1] and nw is bounded due to $n\rho'$ is bounded. The convexity of the loss is shown in Appendix B of Chan et al. (2016).

We then show assumption B: the quadratic marginal condition. We compute the second derivative of EL_w:

$$\begin{aligned}\frac{\partial^2 EL_w(X,Z)}{\partial w^2} =& -\pi(x)\cdot((n\rho')^{-1})'nw(x) + \{-\pi(x)nw(x)+1\}\{((n\rho')^{-1})''nw(x)\} \\ \geq& \pi(x)\frac{\rho''}{(\rho')^2}(\frac{1}{\pi(x)}) + |\eta|\cdot\{((\rho')^{-1})''nw(x)\}.\end{aligned}$$

This is lower bounded by a positive constant when $\eta > 0$ is small enough. This is ensured again by Assumption 3.1, in particular the concavity of ρ. The last step is due to a Taylor expansion around $nw(x) = 1/\pi(x)$ in its η-neighborhood.

Lastly we show assumption C: $\sum_{k\in\mathcal{K}}\sigma_k|\lambda_k - \tilde{\lambda}_k| \leq |\mathcal{K}|\cdot||w_\lambda - w_{\tilde{\lambda}}||$. This is again ensured by Assumption 3.1, in particular the boundedness of the first and second derivative.

The theorem then follows from Theorem 2.2 of Van de Geer (2008) where $H = cu^2/2$ and $G = u^2/(2c)$ for some constant $c > 0$ due to the quadratic margin condition. □

D. DETAILS ON EMPIRICAL STUDIES

D·1. *A Remark on the Right Heart Catheterization Study*

A remark on Table 1(b) is that the optimal error of the weighting estimator for the average treatment effect on the treated is sometimes smaller under bad overlap than under good overlap. This may be counterintuitive, but is a result of the estimand changing under good and bad overlap when estimating the average treatment effect on the treated. Specifically, the treated population is different in the simulated data sets with good and bad overlap, so the estimand is different. This phenomenon is absent when estimating the average treatment effect, where the estimand is the same under good and bad overlap (see Table 1(a)).

D·2. *The Kang and Schafer Example*

The Kang and Schafer example (Kang & Schafer, 2007) consists of four unobserved covariates $U_i \overset{iid}{\sim} N(0, I_4),\ i = 1, ..., n$. They are used to generate four covariates X_i that are observed by the investigator: $X_{i1} = \exp(U_{i1}/2),\ X_{i2} = U_{i2}/\{1 + \exp(U_{i1})\} + 10,\ X_{i3} = (U_{i1}U_{i3} + 0.6)^3$, and $X_{i4} = (U_{i2} + U_{i4} + 20)^2$. There is an outcome variable Y_i generated by $Y_i = 210 + 27.4U_{i1} + 13.72U_{i2} + 13.7U_{i3} + 13.7U_{i4} + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, 1)$, and an incomplete outcome indicator Z_i generated as a Bernoulli random variable with parameter $p_i = \exp(-U_{i1} - 2U_{i2} - 0.25U_{i3} - 0.1U_{i4})$. This incomplete outcome indicator denotes whether the outcome is observed ($Z_i = 1$) or not ($Z_i = 0$).

Using this data generation mechanism, the mean difference of the observed covariates between the complete and incomplete outcome data is of $(-0.4, -0.2, 0.1, -0.1)$ standard deviations. We consider this the "good overlap" case. We also consider another case where the generating mechanism of p_i is slightly different: $p_i = \exp(-U_{i1} - 0.5U_{i2} - 0.25U_{i3} - 0.1U_{i4})$. This makes covariate balance slightly worse, resulting in slightly larger mean differences of $(-0.3, -0.5, -0.1, -0.4)$ standard deviations. We consider this the "bad overlap" case.

Tables 2 presents the root mean squared error of the weighting estimates. Approximate balance outperforms exact balance in the bad overlap case. The improvement is not as marked as we documented in the RHC study because the good and bad overlap cases do not differ much: the mean difference goes from $(-0.4, -0.2, 0.1, -0.1)$ in the good overlap to $(-0.3, -0.5, -0.1, -0.4)$ in the bad overlap case. With this relatively small change in covariate balance, minimal weights immediately outperform the exact balancing weights in the bad overlap cases. This gives us an understanding of when we should use minimal weights. We also observe that minimal weights can sometimes outperform the exact balancing weights in the good overlap case.

D·3. *The LaLonde Data Set*

We next study the performance of minimal weights in the LaLonde data set (LaLonde, 1986). This data set has two components: an experimental part from a randomized experiment evaluating a large scale job training program (the National Supported Work Demonstration, NSW) on 185 participants; and an observational part, where the experimental control group from the randomized experiment is replaced by a control group of 15992 of nonparticipants drawn from the Current Population Survey (CPS). The experimental part provides a benchmark for the effect of the job training program to be recovered from observational part of the data set. This benchmark is \$1794 for the average treatment effect on the treated with a 95% confidence interval of $[551, 3038]$.

Table 3 presents the average treatment effect on the treated estimates and their 95% confidence intervals using minimal weights and its exact balancing counterpart. We use δ·sd for different levels of approximate balancing. Minimal weights together with the tuning algorithm produces more efficient mean average treatment effect on the treated estimates while remaining close to the experimental target \$1794. The 95% confidence intervals all contain the experimental 95% confidence interval and they become more efficient as δ increases. When δ grows to as large as 1 sd, the average treatment effect on the treated estimates starts to shift away from the target. This is intuitive as overly large δ would imply we are no longer balancing the covariates. In this regard, we conclude minimal weights produce more efficient average treatment effect on the treated estimates while being faithful to the truth (experimental target).

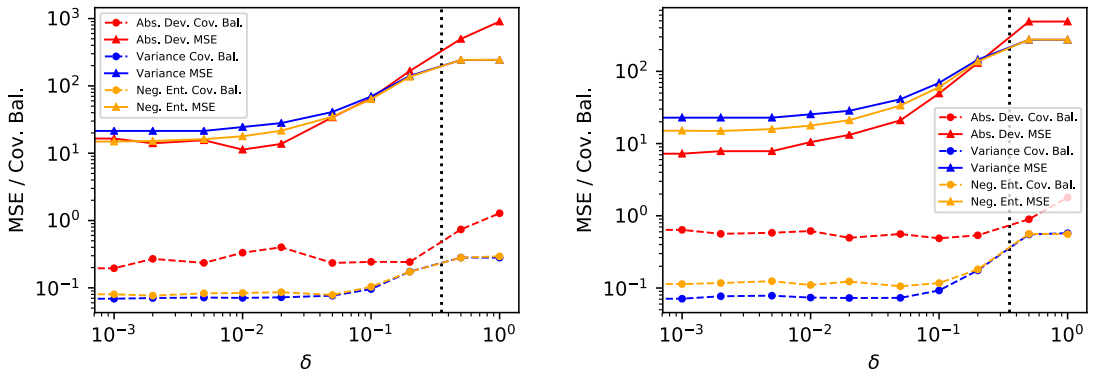
Minimize	Good Overlap Exact	Good Overlap Approx.	Bad Overlap Exact	Bad Overlap Approx.
Absolute Deviation	**6.38**	**6.38**	7.83	**7.20**
Variance	**5.71**	5.79	5.99	**5.65**
Negative Entropy	**5.55**	5.99	5.75	**5.30**

(a) Mean unobserved outcome

Minimize	Good Overlap Exact	Good Overlap Approx.	Bad Overlap Exact	Bad Overlap Approx.
Absolute Deviation	6.38	**5.01**	4.87	**4.80**
Variance	**4.50**	4.59	4.98	**4.85**
Negative Entropy	**3.70**	3.85	4.97	**4.87**

(b) Mean outcome

Table 2: Root mean squared error in the Kang-Schafer study. With bad overlap, approximate balancing can help reduce the estimation error.



(a) Good overlap

(b) Bad overlap

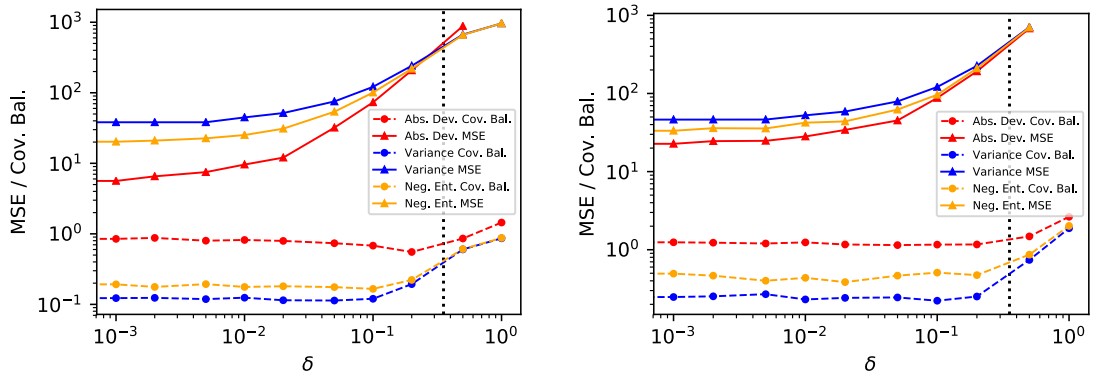
Fig. 2: Bootstrapped covariate balance C_S and mean squared error for different values of δ for the average treatment effect on the treated in the Kang and Shafer study. Using C_S to select δ as in Algorithm 1 coincides with or neighbors the optimal δ with the smallest error. (The horizontal axis start from $\delta = 0$. The vertical dotted line indicates $\delta = K^{-1/2}$, where K is the number of covariates being balanced. We recommend not choosing δ's bigger than $K^{-1/2}$ because they likely break the assumptions required by the asymptotics.)

D·4. *The Wong and Chan Simulation*

We finally study the minimal weights in the Wong & Chan (2018) simulation. It starts with a ten-dimensional multivariate standard Gaussian random vector $Z = (Z_1, \ldots, Z_{10})^\top$ for each observation. Then it generates ten observed covariates $X = (X_1, \ldots, X_{10})^\top$, where

$$X_1 = \exp(Z_1/2),$$

$$X_2 = Z_2/\{1 + \exp(Z_1)\},$$



(a) Good overlap

(b) Bad overlap

Fig. 3: Bootstrapped covariate balance C_S and mean squared error for different values of δ for the average treatment effect on the treated in the Kang and Shafer study. Using C_S to select δ as in Algorithm 1 coincides with or neighbors the optimal δ with the smallest error. (The horizontal axis start from $\delta = 0$. The vertical dotted line indicates $\delta = K^{-1/2}$, where K is the number of covariates being balanced. We recommend not choosing δ's bigger than $K^{-1/2}$ because they likely break the assumptions required by the asymptotics.)

Minimize	Exact	Approx.
Absolute Deviation	712 (2602)	**744 (1257)**
Variance	1668 (1076)	**1387 (886)**
Negative Entropy	1706 (958)	**1382 (1078)**

Table 3: Average treatment effect on the treated estimates in the Lalonde study. (We present the estimates as mean(sd).) Minimal weights produce more efficient estimates while being faithful to the truth.

$$X_3 = (Z_1 Z_3/25 + 0.6)^3,$$

$$X_4 = (Z_2 + Z_4 + 20)^2,$$

$$X_j = Z_j, j = 5, \ldots, 10.$$

The propensity score model is

$$\text{pr}(T = 1 \mid Z) = \exp(-Z_1 - 0.1Z_4)/\{1 + \exp(-Z_1 - 0.1Z_4)\}.$$

The study considers two outcome regression models. Model A is

$$Y = 210 + (1.5T - 0.5)(27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4) + \epsilon,$$

and model B is

$$Y = Z_1 Z_2^3 Z_3^2 Z_4 + Z_4 |Z_1|^{0.5} + \epsilon,$$

where $\epsilon \sim N(0, 1)$.

We generate a dataset of size $N = 5000$ and study both the average treatment effect and the average treatment effect on the treated estimates. (We take the size of the bootstrap samples as 1/10 of the original

Minimize	Outcome model A Exact	Outcome model A Approx.	Outcome model B Exact	Outcome model B Approx.
Absolute Deviation	0.67	**0.66**	**0.26**	**0.26**
Variance	**0.72**	0.79	0.26	**0.25**
Negative Entropy	**0.78**	0.89	**0.25**	**0.25**

(a) Average treatment effect on the treated

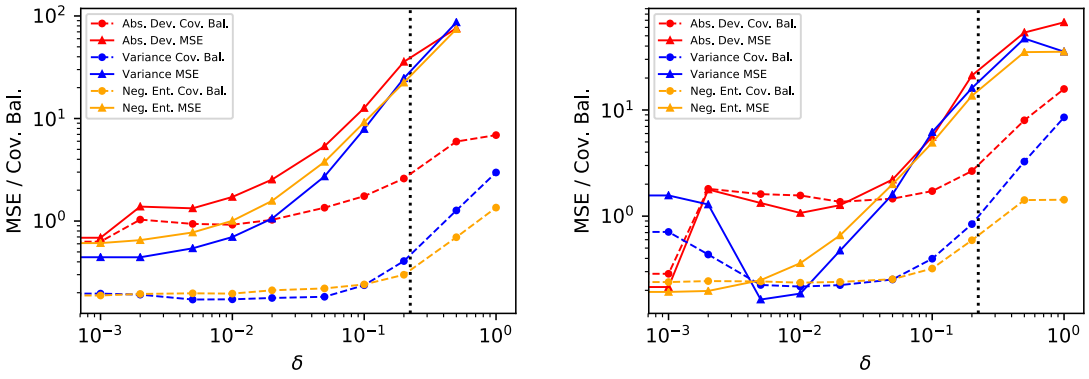
Minimize	Outcome model A Exact	Outcome model A Approx.	Outcome model B Exact	Outcome model B Approx.
Absolute Deviation	0.47	**0.45**	**0.23**	0.24
Variance	1.35	**0.51**	0.31	**0.21**
Negative Entropy	**0.44**	0.52	**0.21**	**0.21**

(b) Average treatment effect

Table 4: Root mean squared error in the Wong-Chan study. Approximate balancing often produce similar-or-better quality estimates than exact balancing.

sample size. We default to 10 bootstrap samples for covariate balance evaluation. We balance the first and second moments of the covariates.)

Tables 4 presents the root mean squared error of the weighting mean estimates. Approximate balancing with Algorithm 1 outperforms exact balancing in many cases, especially in estimating the average treatment effect. The performance is less stable with the outcome model A, where it could lead to suboptimal performance. When the treatment indicator interacts with potential confounders Z's, classical bootstrap agnostic to the treatment indicator does not serve as a good indicator of downstream estimation performance. Figure 4 shows the mean squared error versus bootstrapped covariate balance plot. The pattern of bootstrapped covariate balance roughly aligns with the mean squared error. This implies that selecting δ with Algorithm 1 (i.e. selecting according to the bootstrapped covariate balance) could often result in close-to-optimal error, especially in estimating the average treatment effect.



(a) Average treatment effect on the treated

(b) Average treatment effect

Fig. 4: Bootstrapped covariate balance C_S and mean squared error for different values of δ for the average treatment effect on the treated in the Wong and Chan study. Using C_S to select δ as in Algorithm 1 coincides with or neighbors the optimal δ with the smallest error, especially in estimating the average treatment effect. (The horizontal axis start from $\delta = 0$. The vertical dotted line indicates $\delta = K^{-1/2}$, where K is the number of covariates being balanced. We recommend not choosing δ's bigger than $K^{-1/2}$ because they likely break the assumptions required by the asymptotics.)