Robust Weighting and Matching Techniques for Causal Inference in Observational Studies with Continuous Treatment

Universität Stuttgart

Ioan Scheffel

October 29, 2022

Contents

Chapter 1

Introduction

Researchers are often left with observational studies to answer questions about causality. When confounders are present the task of infering causality can become arbitrarily complex. Propensity score methods [6], e.g. inverse probability weighting or matching, are popular methods to adjust for confounders. Usually these methods rely heavily on estimates of the true propensity score, which are known to suffer from model dependencies and misspecification [4]. This issue becomes more pressing when moving from binary to continuous treatment [3]. Therefore methods have been developed to directly target imbalances in the data [1] [2] [11]. We take a closer look at [10] and extend the analysis to settings with continuous treatment [9] [8].

Chapter 2

Balancing Weights

Assumption 1. *Assume, the following conditions hold:*

1.1. *The minimizer* $\lambda_0 = \arg\min_{\lambda\in\Theta} \mathbb{E}\left[-Tn\rho\left(B(X)^T\lambda\right) + B(X)^T\lambda\right]$ *is unique, where* $\Theta \subseteq \mathbb{R}^n$ *is the parameter space for* λ.

1.2. *The parameter space* $\Theta \subseteq \mathbb{R}^n$ *is compact.*

1.3. $\lambda_0 \in int(\Theta)$, *where* $int(\cdot)$ *stands for the interior of a set.*

1.4. *There exists* $\lambda_1^* \in \Theta$ *such that* $\left\|m^*(\cdot) - B(\cdot)^T\lambda_1^*\right\|_\infty \leq \varphi_{m^*}$, *where* $m^*(\cdot) := \left(\rho'\right)^{-1}\left(\frac{1}{n\pi(\cdot)}\right)$.

1.5. *There exists a constant* $\varphi_\pi \in \left(0, \frac{1}{2}\right)$ *such that* $\pi(x) \in (\varphi_\pi, 1-\varphi_\pi)$ *for all* $x \in \mathcal{X}$

1.6. *There exists* $\varphi_{\rho''} > 0$ *such that* $-\rho'' \geq \varphi_{\rho''} > 0$

1.7. *There exists* $\varphi_{B(x)B(x)^T} > 0$ *such that* $B(x)B(x)^T \succcurlyeq \varphi_{B(x)B(x)^T} I$

1.8. *There exists* $\varphi_{\|B\|} > 0$ *such that* $\sup_{x\in\mathcal{X}} \|B(x)\|_2 \leq \varphi_{\|B\|}$.

1.9. *The number of basis functions satisfies* $K = o(n)$.

We study the following problem:

$$\begin{aligned} &\underset{w\in\mathbb{R}^n}{\text{minimize}} \quad && \sum_{i=1}^n T_i f(w_i) \\ &\text{subject to} && \left|\sum_{i=1}^n w_i T_i B_k(X_i) - \frac{1}{n}\sum_{i=1}^n B_k(X_i)\right| \leq \delta_k,\ k = 1, \ldots, K \end{aligned} \tag{2.1}$$

We aim to prove that the solution to Problem (2.1) is asymptotical consistent with the propensity score, i.e.

Theorem 2.1. *Under some (non-optimal) Assumptions, there exist constants* $c_1, c_2 > 0$ *and decreasing sequences* $(\varepsilon_n^1), (\varepsilon_n^2) \subset (0,1]$ *that converge to 0 such that for all* $\tau \in (0,1]$ *there exists a constant* $c_\tau \in [0, \infty)$ *only depending on* τ *such that for all* $n \geq 1$ *and* $\tau \in (0,1]$ *it holds*

$$\mathbb{P}\left(\left\| w_i^* - \frac{1}{n\pi(X_i)} \right\|_\infty \leq c_1 c_\tau \varepsilon_n^1 \right) \geq 1 - \tau,$$
$$\left\| w_i^* - \frac{1}{n\pi(X_i)} \right\|_{\mathbb{P},2} \leq c_2 \varepsilon_n^2,$$

where w^* *is the solution to Problem* (2.1).

Plan of Proof

It is easier to study the dual of Problem (2.1). Thus we employ results from convex analysis [5] to establish

Proposition 2.1. *The dual of Problem* (2.1) *is equivalent to the unconstrained optimization problem*

$$\underset{\lambda \in \mathbb{R}^K}{\textit{minimize}} \quad \frac{1}{n} \sum_{j=1}^{n} \left[-T_j n \rho \left(B(X_j)^T \lambda \right) + B(X_j)^T \lambda \right] + |\lambda|^T \delta, \tag{2.2}$$

where $B(X_j) = (B_k(X_j))_{1 \leq k \leq K}$ *denotes the* K *basis functions of the covariates,* $\rho(t) := \frac{t}{n} - t(h^{'})^{-1}(t) + h((h^{'})^{-1}(t))$ *with* $h(x) := f\left(\frac{1}{n} - x\right)$ *and* $|\lambda| := (|\lambda_k|)_{1 \leq k \leq K}$*. Moreover, the primal solution* w_j^* *satisfies*

$$w_j^* = \rho^{'} \left(B(X_j)^T \lambda^\dagger \right) \tag{2.3}$$

for $j = 1, \ldots, n$*, where* $\lambda^\dagger$ *is the solution to the dual optimization problem.*

The core of the subsequent analysis is based on Assumption 1.4, i.e. the existence of an oracle parameter λ_1^* in a sieve estimate of the true propensity score (or a transformation). It is then natural to enquire about the convergence of the dual solution $\lambda^\dagger$ to λ_1^*. Making certain assumptions and employing matrix concentration inequalitys [7] we can establish

Proposition 2.2. *Under some (non-optimal) Assumptions, there exists a constant* $c_3 > 0$ *and a decreasing sequence* $(\varepsilon_n^3) \subset (0, 1]$ *that converges to 0 such that for all* $\tau \in (0, 1]$ *there exists a constant* $\tilde{c_\tau} \in [0, \infty)$ *only depending on* τ *such that for all* $n \geq 1$ *and* $\tau \in (0, 1]$ *it holds*

$$\mathbb{P}\left(\left\|\lambda^\dagger - \lambda_1^*\right\|_2 \leq c^3 \tilde{c_\tau}(\varepsilon_n^3)\right) \geq 1 - \tau. \tag{2.4}$$

It is then straightforward to prove a more general result then Theorem 2.1.

Theorem 2.2. *Under some (non-optimal) Assumptions, there exist constants* $c_1, c_2 > 0$ *and decreasing sequences* $(\varepsilon_n^1), (\varepsilon_n^2) \subset (0, 1]$ *that converge to 0 such that for all* $\tau \in (0, 1]$ *there exists a constant* $c_\tau \in [0, \infty)$ *only depending on* τ *such that for all* $n \geq 1$ *and* $\tau \in (0, 1]$ *it holds*

$$\mathbb{P}\left(\left\|w^*(\cdot) - \frac{1}{n\pi(\cdot)}\right\|_\infty \leq c_1 c_\tau \varepsilon_n^1\right) \geq 1 - \tau,$$
$$\left\|w^*(X) - \frac{1}{n\pi(X)}\right\|_{\mathbb{P},2} \leq c_2 \varepsilon_n^2,$$

where $w^*(X)$ *is as in* (2.3) *without the index.*

Proof of theorem 2.2

Proof. Motivated by Proposition 5.1 we set $\|\Delta\|_2 = C$ and consider

$$G(\lambda) := \frac{1}{n} \sum_{j=1}^{n} \left[-T_j n\rho\left(B(X_j)^T \lambda\right) + B(X_j)^T \lambda\right] + |\lambda|^T \delta. \tag{2.5}$$

Since $\rho \in C^2(\mathbb{R})$ we can employ Proposition 5.1, Corollary 5.1.1 and Proposition 5.2 to get

$$
\begin{aligned}
& G(\lambda_1^* + \Delta) - G(\lambda_1^*) \\
& \geq \frac{1}{n} \sum_{j=1}^{n} \left[-T_j n \rho' \left(B(X_j)^T \lambda_1^* \right) + 1 \right] \Delta^T B(X_j) \\
& + \frac{1}{2} \sum_{j=1}^{n} -T_j \rho'' \left(B(X_j)^T (\lambda_1^* + \xi \Delta) \right) \Delta^T \left(B(X_j) B(X_j)^T \right) \Delta \\
& - |\Delta|^T \delta \\
& \geq - \|\Delta\|_2 \left(\left\| \frac{1}{n} \sum_{j=1}^{n} \left[-T_j n \rho' \left(B(X_j)^T \lambda_1^* \right) + 1 \right] B(X_j) \right\|_2 + \|\delta\|_2 \right) \\
& + n \|\Delta\|_2^2 \varphi_{\rho''} \underline{\varphi_{BB^T}} \\
& := - \|\Delta\|_2 (I_1 + \|\delta\|_2) + \|\Delta\|_2^2 I_2 .
\end{aligned}
\tag{2.6}
$$

The second inequality is due to the Cauchy-Schwarz-Inequality and Assumptions 1.6 and 1.7 . We want to establish probabilistic upper bounds of the factor associated with $-\|\Delta\|_2$. This will be done with appropriate assumptions on $\|\delta\|_2$ and a thorough analysis of I_1. If we then restrict lower bounds of I_2 to approprietely slow convergence to 0, e.g. by assumptions on $\varphi_{\rho''}$ and $\underline{\varphi_{BB^T}}$, we can choose $\|\Delta\|_2$ large enough, such that (2.6) yields $G(\lambda_1^*+\Delta)-G(\lambda_1^*) > 0$ with arbitrarily large probabitity for n large enough. With Proposition 5.1 it follows then immediately Proposition 2.2 .

Analysis of I_1

We want to use Assumption 1.3. Thus we perform the following split:

$$
\begin{aligned}
I_1 \leq & \left\| \sum_{j=1}^{n} T_j \left[\rho' \left(B(X_j)^T \lambda_1^* \right) - \frac{1}{n\pi(X_j)} \right] B(X_j) \right\|_2 \\
& + \left\| \frac{1}{n} \sum_{j=1}^{n} \left[\frac{T_j}{\pi(X_j)} - 1 \right] B(X_j) \right\|_2 \\
& =: J_1 + J_2
\end{aligned}
\tag{2.7}
$$

Analysis of J_1

By the Lipschitz-continuity of $\rho^{'}$, Assumption 1.8 and Assumption 1.4, $T \in \{0,1\}$ and the triangle inequality we have

$$J_1 \leq nL_{\rho'} \varphi_{\|B(x)\|} \varphi_{m^*} \tag{2.8}$$

Analysis of J_2

We want to employ Theorem 4.1. To this end we define the independent random matrices

$$\begin{aligned} A_j &:= \frac{1}{n}\left[\frac{T_j}{\pi(X_j)} - 1\right] B(X_j)\,, \quad j = 1, \ldots, n, \\ S &:= \sum_{j=1}^{n} A_j \end{aligned} \tag{2.9}$$

and check conditions (4.1) and (4.2). Note that $\|S\|_2 = J_2$. By the properties of conditional expectation it holds

$$\mathbb{E}\left[\frac{T_j}{\pi(X_j)} B(X_j)\right] = \mathbb{E}\left[\mathbb{E}\left[T_j \mid X_j\right] \frac{1}{\pi(X_j)} B(X_j)\right] = \mathbb{E}[B(X_j)]. \tag{2.10}$$

Taking the expectation in (2.9) and using (2.10) we get $\mathbb{E}[A_j] = 0$ for all $j = 1, \ldots, n$. Since

$$\left|\frac{T_j}{\pi(X_j)} - 1\right| \leq 1 + \frac{1 - \varphi_\pi}{\varphi_\pi} = \frac{1}{\varphi_\pi} \tag{2.11}$$

by Assumption 1.5, we can employ Assumption 1.8 together with (2.11) and (2.9) to get

$$\|A_j\|_2 \leq \frac{\varphi_{\|B\|}}{n\varphi_\pi} =: L. \tag{2.12}$$

Thus, condition (4.1) is satisfied. Next we turn to the matrix variance statistic $v(S)$ (4.2). By (2.9) and (2.11) we have

$$\mathbb{E}\left[A_j A_j^T\right] \leq \left(\frac{1}{n\varphi_\pi}\right)^2 \mathbb{E}\left[B(X)B(X)^T\right] \tag{2.13}$$

and by (2.12)

$$\mathbb{E}\left[A_j^T A_j\right] \leq L^2. \tag{2.14}$$

Since $\max\{a,b\} \leq |a| + |b|$ we can use (2.13) and (2.14) to get

$$v(S) \leq \frac{1}{n}\frac{\lambda_{\max}}{\varphi_\pi^2} + nL^2, \tag{2.15}$$

where $\lambda_{\max}$ is the maximal eigenvalue of the symmatric (non-random) matrix $\mathbb{E}\left[B(X)B(X)^T\right]$. Having dealt with (4.1) and (4.2) we can establish the expectation bound (4.3) of Theorem 4.1. Together with (2.12) and (2.15) we get

$$\begin{aligned}
&\mathbb{E}[J_2] \\
&\leq \sqrt{\frac{2\log(K+1)\left(\lambda_{\max} + \varphi_{\|B\|}^2\right)}{n\varphi_\pi^2}} + \frac{\log(K+1)\varphi_{\|B\|}}{3n\varphi_\pi} \\
&\leq \frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(\sqrt{2} + \frac{1}{3}\sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{2\lambda_{\max}}\right].
\end{aligned} \tag{2.16}$$

Since $K = o(n)$ by Assumption 1.9 we can discuss the other influences on the quality of the bound (2.16). On a high-level it is readily clear that appropriate bounds on $\varphi_\pi, \varphi_{\|B\|}$ and $\lambda_{\max}$ will shrink $\mathbb{E}[J_2]$ to 0 and will assist in establishing learning rates.

We could also have invoked the probabitity bound (4.4) of Theorem 4.1. But for the sake of simplicity we prefer the combination of the expectation bound (2.16) and the Markov inequality. With the latter we get

$$J_2 \leq \frac{1}{\tau}\frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(\sqrt{2} + \frac{1}{3}\sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{2\lambda_{\max}}\right] \tag{2.17}$$

with probabitity $\geq 1-\tau$.

If we choose $\|\Delta\|_2$ to be

$$\begin{aligned}
&\left(\sqrt{2}\frac{1}{\tau}\frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(1 + \sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{\lambda_{\max}}\right]\right. \\
&\qquad \left. + L_{\rho'}\varphi_{\|B\|}\varphi_{m^*} + \frac{\|\delta\|_2}{n}\right)\frac{1}{\varphi_{\rho''}\underline{\varphi_{BB^T}}}
\end{aligned} \tag{2.18}$$

we get by (2.6), (2.7), (2.8), (2.17) and Proposition 5.1

$$\begin{aligned}\mathbb{P}\left(\left\|\lambda^\dagger - \lambda_1^*\right\|_2 \leq C\right) &= \mathbb{P}\left(\inf_{\|\Delta\|_2 = C} G(\lambda_1^* + \Delta) - G(\lambda_1^*) > 0\right) \\ &\geq 1 - \tau,\end{aligned}$$

where C is as in (2.18). With appropriate Assumptions (as discussed before) we can then establish Proposition 2.2.

Finish 2

$$\begin{aligned}\left\|w^*(X) - \frac{1}{n\pi(X)}\right\|_{\mathbb{P},2} &\leq L_{\rho'}\left[\left\|B(X)^T\left(\lambda^\dagger - \lambda_1^*\right)\right\|_{\mathbb{P},2}\right. \\ &\quad \left. + \left\|m^*(X) - B(X)^T\lambda_1^*\right\|_{\mathbb{P},2}\right] \\ &\leq L_{\rho'}\left(\varphi_{\|B\|}\sqrt{C^2(1-\tau) + \mathrm{diam}(\Theta)^2\tau} + \varphi_{m^*}\right)\end{aligned}$$

$$\begin{aligned}\left\|w^*(\cdot) - \frac{1}{n\pi(\cdot)}\right\|_\infty &\leq L_{\rho'}\left[\left\|B(\cdot)^T\left(\lambda^\dagger - \lambda_1^*\right)\right\|_\infty\right. \\ &\quad \left. + \left\|m^*(\cdot) - B(\cdot)^T\lambda_1^*\right\|_\infty\right] \\ &\leq L_{\rho'}\left(\varphi_{\|B\|}C + \varphi_{m^*}\right)\end{aligned}$$

with probabitity greater than $1 - \tau$. □

The next step consists of strenghtening the Assumptions to get concrete learning rates. This can be done in a series of examples.

Chapter 3

Convex Analysis

We begin by defining convex sets

Definition 3.1. *A subset* $\Omega \subseteq \mathbb{R}^n$ *is called CONVEX if we have* $\lambda x + (1 - \lambda)y \in \Omega$ *for all* $x, y \in \Omega$ *and* $\lambda \in (0, 1)$.

Clearly, the line segment $[a, b] := \{\lambda a + (1 - \lambda)b \mid \lambda \in [0, 1]\}$ is contained in Ω for all $a, b \in \Omega$ if and only if Ω is a convex set.

Next we define convex functions.

The concept of convex functions is closely related to convex sets.

The line segment between two points on the graph of a convex function lies on or above and does not intersect the graph.

In other words: The area above the graph of a convex function f is a convex set, i.e. the *epigraph* $\operatorname{epi}(f) := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$ is a convex set in $\mathbb{R}^{n+1}$.

Often an equivalent characterisation of convex functions is more useful.

Theorem 3.1. *The convexity of a function* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *on* $\mathbb{R}^n$ *is equivalent to the following statement:*

For all $x, y \in \mathbb{R}^n$ *and* $\lambda \in (0, 1)$ *we have*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \tag{3.1}$$

Definition 3.2. proper convex function

Definition 3.3. convex conjugate

Given proper convex functions $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a matrix $A \in \mathbb{R}^{n\times n}$, we define the primal minimization problem as follows:

$$\text{minimize} \quad f(x) + g(Ax) \quad \text{subject to} \quad x \in \mathbb{R}^n. \tag{3.2}$$

The Fenchel dual problem is then

$$\text{maximize} \quad -f^*\left(A^T y\right) - g^*(-y) \quad \text{subject to} \quad y \in \mathbb{R}^n. \tag{3.3}$$

Theorem 3.2. *Let $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper convex functions and $0 \in ri(dom(g) - A(dom(f)))$. Then the optimal values of* (3.2) *and* (3.3) *are equal, i.e.*

$$\inf_{x\in\mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y\in\mathbb{R}^n} \left\{-f^*\left(A^T y\right) - g^*(-y)\right\}. \tag{3.4}$$

Chapter 4

Matrix Concentration Inequalities

Theorem 4.1. (Matrix Bernstein Inequality) *Let* $(A_k)_{1\le k\le n} \subseteq \mathbb{R}^{d_1\times d_2}$ *be a finite sequence of independent, random matrices. Assume that*

$$\mathbb{E}(A_k) = 0 \quad \text{and} \quad \|A_k\| \le L \quad \text{for each} \quad k \in \{1,\dots,n\}. \tag{4.1}$$

Introduce the random matrix

$$S := \sum_{k=1}^{n} A_k.$$

Let $v(S)$ *be the matrix variance statistic of the sum:*

$$\begin{aligned} v(S) :&= \max\left\{\left\|\mathbb{E}(SS^T)\right\|, \left\|\mathbb{E}(S^TS)\right\|\right\} \\ &= \max\left\{\left\|\sum_{k=1}^{n}\mathbb{E}(A_kA_k^T)\right\|, \left\|\sum_{k=1}^{n}\mathbb{E}(A_k^TA_k)\right\|\right\}. \end{aligned} \tag{4.2}$$

Then

$$\mathbb{E}\,\|S\| \le \sqrt{2v(S)\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2). \tag{4.3}$$

Furthermore, for all $t \ge 0$*,*

$$\mathbb{P}(\|S\| \ge t) \ge (d_1+d_2)\exp\left(\frac{-t^2/2}{v(S)+Lt/3}\right). \tag{4.4}$$

Chapter 5

Simple yet useful Calculations

Proposition 5.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous such that a minimum x^* exists and is unique. Then for all $y \in \mathbb{R}^n$ and $C > 0$ it follows*

$$\inf_{\|\Delta\|=C} f(y+\Delta) - f(y) > 0 \qquad \Rightarrow \qquad \|x^* - y\| \leq C. \tag{5.1}$$

Proof. Since $\mathcal{C} := \{\|\Delta\| \leq C\}$ is compact and

$$f(x^*) \leq f(y) < \inf_{\|\Delta\|=C} f(y+\Delta),$$

the continious function $f(y + \,\cdot\,)$ has a minimum in $\mathrm{int}(\mathcal{C}) := \{\|\Delta\| < C\}$. Since x^* is the unique minimum of f there exists $\Delta^* \in \mathrm{int}(\mathcal{C})$ such that $x^* - y = \Delta^*$. We conclude that $\|x^* - y\| \leq C$. □

Theorem 5.1. (Multivariate Taylor Theorem) *Let $f \in C^2(\mathbb{R}^n, \mathbb{R})$. Then for all $x, \Delta \in \mathbb{R}^n$ there exists $\xi \in [0,1]$ such that it holds*

$$\begin{aligned} f(x+\Delta) = f(x) &+ \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\Delta_i + \sum_{\substack{i,j=1 \\ i \neq j}} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i \partial x_j}\Delta_i \Delta_j \\ &+ \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i^2}\Delta_i^2 \end{aligned} \tag{5.2}$$

Corollary 5.1.1. *Let $f \in C^2(\mathbb{R})$. Then for all $a, x, \Delta \in \mathbb{R}^n$ there exist $\xi \in [0,1]$ such that it holds*

$$f(a^T(x+\Delta)) - f(a^T x) = f^{'}(a^T x)\,\Delta^T a + \frac{1}{2} f^{''}(a^T(x+\xi\Delta))\,\Delta^T A\ \Delta, \tag{5.3}$$

where $A := aa^T \in \mathbb{R}^{n \times n}$.

Proof. By the chain rule we have for all $a, x, \Delta \in \mathbb{R}^n$ and $\xi \in [0, 1]$

$$\frac{\partial^2 f(a^T(x+\xi\Delta))}{\partial x_i \partial x_j} = f^{''}(a^T(x+\xi\Delta))\, a_i a_j. \tag{5.4}$$

Since $A := aa^T$ is symmetric we have

$$\Delta^T A\ \Delta = 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_i a_j \Delta_i \Delta_j + \sum_{i=1}^{n} a_i^2 \Delta_i^2. \tag{5.5}$$

Plugging (5.4) and (5.5) into (5.2) yields (5.3). □

Proposition 5.2. *For all $x, y \in \mathbb{R}$ it holds*

$$|x+y| - |x| \geq -|y| \tag{5.6}$$

Proof. Checking all 6 combinations of $x+y$, x, y being nonnegative or negative yields the result. □

Bibliography

[1] Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, March 2018.

[2] Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012.

[3] Keisuke Hirano and Guido W. Imbens. The Propensity Score with Continuous Treatments. In Andrew Gelman and Xiao-Li Meng, editors, *Wiley Series in Probability and Statistics*, pages 73–84. John Wiley & Sons, Ltd, Chichester, UK, July 2005.

[4] Joseph D. Y. Kang and Joseph L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539, November 2007.

[5] Boris S. Mordukhovich and Nguyen Mau Nam. ENHANCED CALCULUS AND FENCHEL DUALITY. In Boris S. Mordukhovich and Nguyen Mau Nam, editors, *Convex Analysis and Beyond: Volume I: Basic Theory*, Springer Series in Operations Research and Financial Engineering, pages 255–310. Springer International Publishing, Cham, 2022.

[6] Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

[7] Joel A. Tropp. An Introduction to Matrix Concentration Inequalities, January 2015.

[8] Stefan Tübbicke. Entropy Balancing for Continuous Treatments, May 2020.

[9] Brian G. Vegetabile, Beth Ann Griffin, Donna L. Coffman, Matthew Cefalu, and Daniel F. McCaffrey. Nonparametric Estimation of Population Average Dose-Response Curves using Entropy Balancing Weights for Continuous Exposures, March 2020.

[10] Yixin Wang and José R. Zubizarreta. Minimal Dispersion Approximately Balancing Weights: Asymptotic Properties and Practical Considerations. *Biometrika*, page asz050, October 2019.

[11] José R. Zubizarreta. Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association*, 110(511):910–922, July 2015.