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

Universität Stuttgart

Universität Stuttgart

Ioan Scheffel

December 10, 2022

Contents

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] [10]. We take a closer look at [9] and extend the analysis to settings with continuous treatment [8] [7].

2 Balancing Weights

2.1 Introduction

We work in the Rubin Causal Model.

We assume a sample of n units which is drawn from a population distribution.

In i.i.d. fashion.

We observe $(\mathbf{X}_i, T_i, Y_i)$, where $\mathbf{X}$ are covariates, T is the indicator if treatment has been received and Y is the observed outcome.

In the Rubin Causal Model we assume that for each unit the potential outcome exist, i.e. (Y_i^0, Y_i^1) where Y^1 stands for the potential outcome had the unit received treatment and Y^0 for the potential outcome had the unit received **no** treatment.

It is clear that $Y_i = Y_i^{T_i}$ i.e. we can observe only one of the potential outcomes.

Thus there is a connection to missing data problems.

This is the dilemma of causal inference.

On the population level it is possible to estimate both.

Usually the means of the potential outcomes are compared against each other.

In randomized trials this is a valid approach to causal inference.

In observational studies however the treatment assignment is not known and direct comparison can lead to systematically wrong results.

This phenomenon is called **confounding**.

To address the issue of confounding many methods have been proposed.

An intuitive way to think about potential outcomes is to think of a stochastic process $Y(\cdot)$ indexed over $\{0, 1\}$. By observing Y_i we in fact sample from this process at random index T, i.e. from $Y(T)$. We have

$$\mathbf{E}[Y(T)] = \mathbf{E}[Y(1)|T = 1]\mathbf{P}[T = 1] + \mathbf{E}[Y(0)|T = 0]\mathbf{P}[T = 0]. \tag{2.1}$$

Suppose we observe $T = 1$. Clearly we have

$$\mathbf{E}[Y(T)|T = 1] = \mathbf{E}[Y(1)|T = 1] \tag{2.2}$$

2.2 Estimating the Population Mean of Potential Outcomes

2.3 Application of Matrix Concentration Inequalities

Analysis of $\mathbf{E}[\max_{i \le r} \|\mathbf{A}_i\|^2]$

We have

$$\mathbf{A}_i := \frac{1}{r}\left(\frac{1 - \pi_i}{\pi_i}\right)\mathbf{B}(X_i) \qquad \text{for } i \in \{1, \dots, r\}. \tag{2.3}$$

Since we take the maximum over a finite set it is attained for some $i^* \in \{1, \dots, r\}$:

$$\begin{aligned}\mathbf{E}[\max_{i \le r} \|\mathbf{A}_i\|^2] &= \mathbf{E}[\|\mathbf{A}_{i^*}\|^2] \\ &= \frac{1}{r^2}\mathbf{E}\left[\left(\frac{1-\pi_{i^*}}{\pi_{i^*}}\right)^2 \|\mathbf{B}(X_{i^*})\|^2\right] \le \frac{1}{r^2}\mathbf{E}\left[\left(\frac{1-\pi_{i^*}}{\pi_{i^*}}\right)^4\right]^{\frac{1}{2}} \mathbf{E}[\|\mathbf{B}(X_{i^*})\|^4]^{\frac{1}{2}} \\ &\le \frac{K}{r^2}\sqrt{C_\pi C_\mathbf{B}}\end{aligned} \tag{2.4}$$

In the last two steps we applied the Cauchy-Schwarz inequality and Assumption. Note that

$$\sum_{i=1}^{r} \mathbf{E}[\|\mathbf{A}_i\|^2] \le \frac{K}{r}\sqrt{C_\pi C_\mathbf{B}} \tag{2.5}$$

Assumption 2.1. *There exists* $C_\pi \ge 1$ *such that* $\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right] \le C_\pi$ *for all* $i \in \{1, \dots, r\}$.

Remark 2.1. *If we assume a logistic regression model for the propensity score it holds for some* $\theta \in \mathbb{R}^N$ *(*N *is the number of covariates)*

$$\frac{1-\pi(X)}{\pi(X)} = \exp(-\theta X) \qquad \textit{and} \qquad \mathbf{E}\left[\left(\frac{1-\pi(X)}{\pi(X)}\right)^4\right] = \mathbf{E}[\exp(-4\theta X)] = M_X(-4\theta), \tag{2.6}$$

where M_X *is the momement-generating function of* X*. While the first quantity in* (2.6) *may be unbounded when* X *has unbounded support, the latter quantity in* (2.6) *is still bounded for reasonable choices of* X*.* ♢

Assumption 2.2. *There exists* $C_\mathbf{B} \ge 1$ *such that* $\mathbf{E}[\mathbf{B}_k(X_i)^4] \le C_\mathbf{B}$ *for all* $(k, i) \in \{1, \dots, K\} \times \{1, \dots, r\}$.

Remark 2.2. *With Assumption we also get a bound on the fourth moment of* $\|\mathbf{B}(X_i)\|$*. Indeed, by the convexity of* $x \mapsto x^2$*, the monotonicity and linearity of the expectation it holds*

$$\begin{aligned}\mathbf{E}[\|\mathbf{B}(X_i)\|^4] &= \mathbf{E}\left[\left(\sum_{k=1}^{K}\mathbf{B}_k^2(X_i)\right)^2\right] = K^2\mathbf{E}\left[\left(\sum_{k=1}^{K}\frac{1}{K}\mathbf{B}_k^2(X_i)\right)^2\right] \le K^2\mathbf{E}\left[\sum_{k=1}^{K}\frac{1}{K}\mathbf{B}_k^4(X_i)\right] \\ &= K\sum_{k=1}^{K}\mathbf{E}\left[\mathbf{B}_k^4(X_i)\right] \le K^2 C_\mathbf{B}\end{aligned} \tag{2.7}$$

♢

Analysis of $v(\mathbf{S})$

We use the fact that $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ It holds

$$\sum_{i=1}^{r}\mathbf{E}[\mathbf{A}_i\mathbf{A}_i^\top] = \frac{1}{r^2}\sum_{i=1}^{r}\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 \mathbf{B}(X_i)\mathbf{B}(X_i)^\top\right] = \frac{1}{r^2}\left(\sum_{i=1}^{r}\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B_k(X_i)B_l(X_i)\right]\right)_{1\le k,l\le K}. \tag{2.8}$$

Thus

$$\begin{aligned}
&\left\|\sum_{i=1}^{r}\mathbf{E}[\mathbf{A}_i\mathbf{A}_i^\top]\right\|_2^2 \\
&\leq \left\|\sum_{i=1}^{r}\mathbf{E}[\mathbf{A}_i\mathbf{A}_i^\top]\right\|_F^2 = \frac{1}{r^4}\sum_{k,l=1}^{K}\left(\sum_{i=1}^{r}\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B_k(X_i)B_l(X_i)\right]\right)^2 \qquad (2.9)\\
&\leq \frac{1}{r^4}\sum_{k,l=1}^{K}\left(\sum_{i=1}^{r}\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right]^{\frac{1}{2}}\mathbf{E}[B_k(X_i)^4]^{\frac{1}{4}}\mathbf{E}[B_l(X_i)^4]^{\frac{1}{4}}\right)^2 \leq \left(\frac{K}{r}\right)^2 C_\pi C_B
\end{aligned}$$

On the other hand

$$\begin{aligned}
\left\|\sum_{i=1}^{r}\mathbf{E}[\mathbf{A}_i^\top\mathbf{A}_i]\right\|_2 &= \sum_{i=1}^{r}\mathbf{E}[\mathbf{A}_i^\top\mathbf{A}_i] = \frac{1}{r^2}\sum_{i=1}^{r}\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2\|\mathbf{B}(X_i)\|_2^2\right] \\
&\leq \frac{1}{r^2}\sum_{i=1}^{r}\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right]^{\frac{1}{2}}\mathbf{E}[\|\mathbf{B}(X_i)\|_2^4]^{\frac{1}{2}} \leq \frac{K}{r}\sqrt{C_\pi C_B}
\end{aligned} \qquad (2.10)$$

It follows

$$v(\mathbf{S}) \leq \frac{K}{r}\sqrt{C_\pi C_B} \qquad (2.11)$$

Thus we can apply Theorem 4.1 to get

$$\mathbf{E}[\|\mathbf{S}\|_2] \leq \sqrt{2e\frac{K}{r}\sqrt{C_\pi C_B}\log(K+1)} + 4e\frac{\sqrt{K}}{r}\sqrt[4]{C_\pi C_B}\log(K+1) \leq 14C_\pi C_B\sqrt{\frac{K\log(K+1)}{r}} \qquad (2.12)$$

3 Convex Analysis

In the following we do not expect the reader to be familiar with convex analysis. However, some very well known results will be stated without proof. The interested reader can study [5] for the bedrock analysis.

We begin by defining convex sets

Definition 3.1. (Convex Set) *A subset $\Omega \subseteq \mathbb{R}^n$ is called* ***convex*** *if we have*

$$\lambda x + (1-\lambda)y \in \Omega \quad \textit{for all } x, y \in \Omega \textit{ and } \lambda \in (0,1). \tag{3.1}$$

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* $\operatorname{epi}(f) := \{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \le \alpha\}$ 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

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y). \tag{3.2}$$

Definition 3.2. proper convex function

3.1 Conjugate Calculus

When studying different primal problems such as (**??**) we often turn to the dual instead. Therefore we need some reliable tools. Begin able to compute specific convex conjugates is one tool required.

Definition 3.3. (Convex conjugate) *Given a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the* ***convex conjugate*** *$f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ of f is defined as*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \tag{3.3}$$

Note that f in Definition 3.3 does not have to be convex. On the other hand, the convex conjugate is always convex:

Proposition 3.1. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper function. Then its convex conjugate $f^* : \mathbb{R}^n \to (-\infty, \infty]$ is convex.*

Definition 3.4. *Given a nonempty subset $\Omega \subseteq \mathbb{R}^n$ the **support function** $\sigma_\Omega : \mathbb{R}^n \to \overline{\mathbb{R}}$ of Ω is defined by*

$$\sigma_\Omega(x^*) := \sup_{x\in\Omega} \langle x^*, x\rangle \qquad \textit{for } x^* \in \mathbb{R}^n. \tag{3.4}$$

Lemma 3.1. *For any proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ we have*

$$f^*(x^*) = \sigma_{\mathrm{epi}(f)}(x^*, -1) \qquad \textit{for } x^* \in \mathbb{R}^n. \tag{3.5}$$

Proof. Let $x^* \in \mathbb{R}^n$ and $(x, \lambda) \in \mathrm{epi}(f)$. Then $x \in \mathrm{dom}(f)$ and $f(x) \leq \lambda$. Thus

$$\langle x^*, x\rangle - f(x) \geq \langle x^*, x\rangle - \lambda \qquad \text{for all } (x, \lambda) \in \mathrm{epi}(f). \tag{3.6}$$

On the other hand $(x, f(x)) \in \mathrm{epi}(f)$ for all $x \in \mathrm{dom}(f)$. It follows

$$\langle x^*, x\rangle - f(x) \leq \sup_{(x,\lambda)\in\mathrm{epi}(f)} \langle x^*, x\rangle - \lambda \qquad \text{for all } x \in \mathrm{dom}(f). \tag{3.7}$$

Taking the supremum in the last two displays yields

$$\begin{aligned} f^*(x^*) = \sup_{x\in\mathrm{dom}(f)} \langle x^*, x\rangle - f(x) &= \sup_{(x,\lambda)\in\mathrm{epi}(f)} \langle x^*, x\rangle - \lambda & (3.8)\\ &= \sup_{(x,\lambda)\in\mathrm{epi}(f)} \langle (x^*, -1), (x, \lambda)\rangle = \sigma_{\mathrm{epi}(f)}(x^*, -1). & (3.9)\end{aligned}$$

□

Proposition 3.2.

Theorem 3.2. (Conjugate Chain Rule) *Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear map (matrix) and $g : \mathbb{R}^n \to (-\infty, \infty]$ a proper convex function. If $Im(A) \cap ri(dom(g)) \neq \emptyset$ it follows*

$$(g \circ A)^*(x^*) = \inf_{y^*\in(A^*)^{-1}(x^*)} g^*(y^*). \tag{3.10}$$

Furthermore, for any $x^* \in dom(g \circ A)^*$ there exists $y^* \in (A^*)^{-1}(x^*)$ such that $(g \circ A)^*(x^*) = g^*(y^*)$.

Definition 3.5. (Infimal convolution) *Given functions $f_i : \mathbb{R}^n \to (-\infty, \infty]$ for $i = 1, \ldots, n$ the **infimal convolution** of these functions as defined as*

$$(f_1\Box \ldots \Box f_m)(x) := \inf_{\substack{x_i\in\mathbb{R}^n\\ \sum_{i=1}^m x_i = x}} \sum_{i=1}^m f_i(x_i) \tag{3.11}$$

Theorem 3.3. *Let $f, g : \mathbb{R}^n \to (-\infty, \infty]$ be proper convex functions and $ri(dom(f)) \cap ri(dom(g)) \neq \emptyset$. Then we have the conjugate sum rule*

$$(f + g)^*(x^*) = (f^*\Box g^*)(x^*) \tag{3.12}$$

for all $x^* \in \mathbb{R}^n$. Moreover, the infimum in $(f^*\Box g^*)(x^*)$ is attained, i.e., for any $x^* \in dom(f+g)^*$ there exists vectors x_1^*, x_2^* for which

$$(f + g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{3.13}$$

3.2 Fenchel Duality

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:

Problem 3.1. (Primal) *Given proper convex functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ and a matrix $A \in \mathbb{R}^{m\times n}$ we define the **primal optimization problem** to be*

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \qquad f(x) + g(Ax)$$

Remark 3.1. *Problem 3.1 appears in the unconstrained form. We can impose constraints by controling for the domains of f and g. To incorporate linear constraints $Ax \leq 0$ or more general constraints $x \in \Omega$, where Ω is a convex set, we can choose*

$$g(x) = \delta_\Omega(x) := \tag{3.14}$$

where $x \notin \Omega$ leads to $f(x) + g(x) = \infty$ and the optimization problem (if feasible) will exclude x from the solutions. ♢

Problem 3.2. (Dual) *Consider the same setting as in Problem 3.1. Using the convex conjugates of f, g and the transpose of A we define the **dual problem** of Problem 3.1 to be*

$$\underset{y^*\in\mathbb{R}^m}{\text{maximize}} \qquad -f^*(A^\top y^*) - g^*(y^*).$$

Proposition 3.3. *Consider the optimization problem 3.1 and its dual 3.2, where the functions f and g are not assumed to be convex. Define the **optimal values** of these problems by*

$$\widehat{p} := \inf_{x\in\mathbb{R}^n} f(x) + g(Ax) \quad \text{and} \quad \widehat{d} := \sup_{y\in\mathbb{R}^m} -f^*(A^\top y) - g^*(y).$$

Then we have the relationship $\widehat{d} \leq \widehat{p}$.

Proof. It holds

$$\begin{aligned}
-f^*(A^\top y^*) - g^*(y^*) &= -\sup_{x\in\mathbb{R}^n} \langle A^\top y^*, x\rangle - f(x) - \sup_{y\in\mathbb{R}^m} \langle -y^*, y\rangle - g(y) \\
&= \inf_{x\in\mathbb{R}^n} f(x) - \langle y^*, Ax\rangle + \inf_{y\in\mathbb{R}^m} g(y) + \langle y^*, y\rangle \\
&\leq \inf_{x\in\mathbb{R}^n} f(x) - \langle y^*, Ax\rangle + \inf_{x\in\mathbb{R}^n} g(Ax) + \langle y^*, Ax\rangle \\
&\leq \inf_{x\in\mathbb{R}^n} f(x) - \langle y^*, Ax\rangle + g(Ax) + \langle y^*, Ax\rangle \\
&= \inf_{x\in\mathbb{R}^n} f(x) + g(Ax) = \widehat{p}
\end{aligned}$$

The first equality is due to the definition of convex conjugates, the second equality due to $\langle A^\top y, x\rangle = \langle y, Ax\rangle$ and $\inf\{-B\} = -\sup\{B\}$ for all $B \subseteq \overline{\mathbb{R}}$ and the first inequality due to $\mathrm{Im}(A) \subseteq \mathbb{R}^m$. Taking the supremum with respect to all $y^* \in \mathbb{R}^m$ yields the result. □

Theorem 3.4. *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 (3.1) and (3.2) are equal, i.e.*

$$\inf_{x\in\mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y\in\mathbb{R}^n} \left\{-f^*\left(A^T y\right) - g^*(-y)\right\}. \tag{3.15}$$

Lemma 3.2. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be convex. Then for all $y \in \mathbb{R}^n$ and $C > 0$*

$$\inf_{\|\Delta\|=C} f(y+\Delta) - f(y) \geq 0 \quad \Longrightarrow \quad \exists y^* \in \mathbb{R}^n \colon y^* \text{ is global minimum of } f \text{ and } \|y^* - y\| \leq C. \tag{3.16}$$

***Proof*.** Since $\mathcal{C} := \{\|\Delta\| \leq C\}$ is convex f has a local minimum in $y{+}\mathcal{C} := \{y + \Delta \mid \|\Delta\| \leq C\}$. Suppose towards a contradiction that $y^* \in y{+}\mathcal{C}$ is a local minimum, but not a global minimum and the left-hand side of (3.16) is true. Then it holds

$$f(x) < f(y^*) \quad \text{for some } x \in \mathbb{R}^n \setminus y + \mathcal{C}. \tag{3.17}$$

Furthermore since $y + \mathcal{C}$ is compact and contains y^*, the line segment $\mathcal{L}[y^*, x]$ contains a point on the boundary of $y + \mathcal{C}$, i.e.

$$\theta x + (1 - \theta)y^* = y + \Delta_x \quad \text{for some } \theta \in (0, 1) \text{ and } \Delta_x \text{ with } \|\Delta_x\| = C. \tag{3.18}$$

It follows

$$\begin{aligned} f(y^*) \leq f(y) \leq f(y + \Delta_x) &= f(\theta x + (1 - \theta)y^*) \\ &\leq \theta f(x) + (1 - \theta)f(y^*) < f(y^*), \end{aligned} \tag{3.19}$$

which is a contradiction. Thus every local minimum of f in $y + \mathcal{C}$ is also a global minimum. The first inequality is due to y^* being a local minimum of f in $y + \mathcal{C}$, the second inequality is due to the left-hand side of (3.16) being true, the equality is due to (3.18), the third inequality is due to the convexity of f and the strict inequality is due to (3.17). □

4 Matrix Concentration Inequalities

Theorem 4.1. (Matrix Rosenthal-Pinelis) *Let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be independent, random matrices with dimension $d_1 \times d_2$. Introduce the random matrix*

$$\mathbf{S} := \sum_{k=1}^{n} \mathbf{A}_k.$$

Let $v(\mathbf{S})$ be the matrix variance statistic of the sum:

$$v(\mathbf{S}) := \left\|\mathbf{E}[\mathbf{S}\mathbf{S}^\top]\right\| \vee \left\|\mathbf{E}[\mathbf{S}^\top\mathbf{S}]\right\| = \left\|\sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k\mathbf{A}_k^\top]\right\| \vee \left\|\sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k^T\mathbf{A}_k]\right\|. \tag{4.1}$$

Then

$$\left(\mathbf{E}\left[\|\mathbf{S}\|^2\right]\right)^{\frac{1}{2}} \leq \sqrt{2ev(\mathbf{S})\log(d_1+d_2)} + 4e\left(\mathbf{E}[\max_{k\leq n}\|\mathbf{A}_k\|^2]\right)^{\frac{1}{2}}\log(d_1+d_2). \tag{4.2}$$

Remark 4.1. *Since $\mathbf{E}[\|S\|] \leq \mathbf{E}[\|S\|^2]^{\frac{1}{2}}$ by the Cauchy-Schwarz inequality, Theorem 4.1 also holds with $\mathbf{E}[\|S\|]$ on the left-hand side of* (4.2). *To obtain a tail bound we can employ the Markov inequality and Theorem 4.1:*

$$\begin{aligned}&\mathbf{P}[\,\|S\| \geq t] \\ &\quad \leq \frac{\mathbf{E}[\|S\|]}{t} \leq \frac{1}{t}\left(\sqrt{2ev(\mathbf{S})\log(d_1+d_2)} + 4e\left(\mathbf{E}[\max_{k\leq n}\|\mathbf{A}_k\|^2]\right)^{\frac{1}{2}}\log(d_1+d_2)\right) \quad \textit{for } t>0.\end{aligned} \tag{4.3}$$

It might be possible to improve the log *term employing an intrinsic dimension argument.* ♢

5 Empirical Processes

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $(\mathcal{X}, \Sigma)$ a measurable space. Let $X_j : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{X}, \Sigma), j = 1, \ldots, n$ be independent and identically-distributed (i.i.d.) random variables with probability distribution $\mathbf{P}_X$ and $\mathcal{F}$ a family of measurable functions $f : (\mathcal{X}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Consider the map

$$f \mapsto G_n f := \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{P}_X f \right), \tag{5.1}$$

where $\mathbf{P}_X f := \int_{\mathcal{X}} f \mathrm{d}\mathbf{P}_X$. We call $(G_n f)_{f \in \mathcal{F}}$ the empirical process indexed by $\mathcal{F}$. Furthermore

$$\|G_n f\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |G_n f| . \tag{5.2}$$

Lemma 5.1. (Bernstein Inequality for Empirical Processes) *For any bounded, measurable function f it holds for all $t > 0$*

$$\mathbf{P}\left(|G_n f| > t\right) \leq 2 \exp\left(-\frac{1}{4} \frac{t^2}{\mathbf{P}_X(f^2) + t \|f\|_\infty / \sqrt{n}} \right) \tag{5.3}$$

Proof. By the Markov inequality it holds for all $\lambda > 0$

$$\mathbf{P}\left(G_n f > t\right) \leq e^{-\lambda t} \mathbf{E} \exp\left(\lambda G_n f\right) \tag{5.4}$$

□

Lemma 5.2. *For any finite class $\mathcal{F}$ of bounded, measurable, square-integrable functions, with $|\mathcal{F}|$ elements, it holds*

$$\mathbf{E} \|G_n f\|_{\mathcal{F}} \lesssim \max_{f \in \mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}} \log\left(1 + |\mathcal{F}|\right) + \max_{f \in \mathcal{F}} \|f\|_{\mathbf{P},2} \sqrt{\log\left(1 + |\mathcal{F}|\right)}. \tag{5.5}$$

6 Simple yet useful Calculations

Theorem 6.1. (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*

$$\begin{aligned} f(x+\Delta) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\Delta_i + \sum_{\substack{i,j=1\\ i\neq j}} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i \partial x_j}\Delta_i\Delta_j \\ + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i^2}\Delta_i^2 \end{aligned} \tag{6.1}$$

Corollary 6.1.1. *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*

$$f(a^T(x+\Delta)) - f(a^T x) = f^{'}(a^T x)\,\Delta^T a + \frac{1}{2} f^{''}(a^T(x+\xi\Delta))\,\Delta^T A\;\Delta, \tag{6.2}$$

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 $\xi \in [0,1]$

$$\frac{\partial^2 f(a^T(x+\xi\Delta))}{\partial x_i \partial x_j} = f^{''}(a^T(x+\xi\Delta))\,a_i a_j. \tag{6.3}$$

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

$$\Delta^T A\;\Delta = 2\sum_{\substack{i,j=1\\ i\neq j}}^{n} a_i a_j \Delta_i \Delta_j + \sum_{i=1}^{n} a_i^2 \Delta_i^2. \tag{6.4}$$

Plugging (6.3) and (6.4) into (6.1) yields (6.2). □

Proposition 6.1. *For all $x, y \in \mathbb{R}$ it holds*

$$|x+y| - |x| \geq -|y| \tag{6.5}$$

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

Notation Index

$\#A$ cardinality of the set A

$\mathbf{E}[X|Y]$ conditional expectation of the random variable X with respect to $\sigma(Y)$

$\mathbf{E}[X]$ expectation of the random variable X

$\mathbf{Var}[X]$ variance of the random variable X

$\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ extension of the real numbers

$\xrightarrow{\mathcal{D}}$ convergence of distributions

$\mathbf{P}$ generic probability measure

$\mathbf{P}_X = \mathbf{P} \circ X^{-1}$ distribution of the random variable X

$\mathbb{R}$ set of real numbers

$x \vee y, x \wedge y, x^+, x^-$ maximum, minimum, positive part, negative part of real numbers

$X \sim \mu$ the random variable has distribution μ

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

[8] 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.

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

[10] 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.