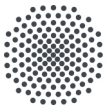
Robust Weighting and Matching Techniques for Causal Inference in Observational Studies with Continuous Treatment

Universität Stuttgart



Universität Stuttgart

Ioan Scheffel

December 1, 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 [7], 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 [5]. This issue becomes more pressing when moving from binary to continuous treatment [4]. Therefore methods have been developed to directly target imbalances in the data [2] [3] [12]. We take a closer look at [11] and extend the analysis to settings with continuous treatment [10] [9].

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

We want to estimate the populition mean of the outcome under treatment, i.e. $\mathbf{E}[Y^1]$.

Since Y_i^1 is only observed for the treated units, i.e. if $T_i = 1$ we will consider a weighted mean of the observed outcomes as an estimator, i.e. $\hat{Y}_w^1 = \sum_{i=1}^n w_i T_i Y_i$ where we use convex optimization to compute the weights.

We consider the following decomposition

$$\hat{Y}_w^1 - \mathbf{E}[Y^1] = \frac{1}{n}\sum_{i=1}^{n} S_i + R_0 + R_1 + R_2, \tag{2.3}$$

where

$$S_i := \frac{T_i}{\pi_i}\left(Y_i - \mathbf{E}[Y_i^1|X_i]\right) + \left(\mathbf{E}[Y_i^1|X_i] - \mathbf{E}[Y^1]\right) \quad \text{for } i \in \{1,\dots,n,\},$$

$$R_0 := \sum_{i=1}^{n} T_i\left(w_i - \frac{1}{n\pi_i}\right)\left(Y_i - \mathbf{E}[Y_i^1|X_i]\right),$$

$$R_1 := \sum_{i=1}^{n}\left(T_i w_i - \frac{1}{n}\right)\left(\mathbf{E}[Y_i^1|X_i] - B(X_i)^\top\lambda\right) \quad \text{and} \quad R_2 := \sum_{i=1}^{n}\left(T_i w_i - \frac{1}{n}\right)B(X_i)^\top\lambda \quad \text{for } \lambda \in \mathbb{R}^K.$$

We want to prove asyptotic normality

Theorem 2.1. *Suppose that conditions hold. Then*

$$\sqrt{n}\left(\hat{Y}^1_{w^*} - \mathbf{E}[Y^1]\right) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,\sigma_*^2).$$

To accomplish this we need

Theorem 2.2. *If $T_i = 1$ then $w^*(X_i)$ is a consistent estimator of $\frac{1}{n\pi(X_i)}$.*

We study the following problem:

Problem 2.1.

$$\underset{w_1,\dots,w_n\in\mathbb{R}}{\text{minimize}} \qquad \sum_{i=1}^{n} T_i f(w_i)$$

subject to the constraints

$$w_i T_i \geq 0, \qquad i = 1,\dots,n,$$

$$\sum_{i=1}^{n} w_i T_i = 1$$

$$\left|\sum_{i=1}^{n} w_i T_i B_k(X_i) - \frac{1}{n}\sum_{i=1}^{n} B_k(X_i)\right| \leq \delta_k, \qquad k = 1,\dots,K$$

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

Theorem 2.3. *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*

$$\begin{aligned} \mathbf{P}\left(\left\|w_i^* - \frac{1}{n\pi(X_i)}\right\|_\infty \leq c_1 c_\tau \varepsilon_n^1\right) &\geq 1-\tau, \\ \left\|w_i^* - \frac{1}{n\pi(X_i)}\right\|_{\mathbf{P},2} &\leq c_2\varepsilon_n^2, \end{aligned} \tag{2.4}$$

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

Assumption 2.1. *Assume, the following conditions hold:*

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

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

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

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

2.1.5. *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* $x \in \mathcal{X}$

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

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

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

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

Plan of Proof

It is easier to study the dual of Problem (**??**). Thus we employ results from convex analysis [6] to establish

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

$$\underset{\lambda\in\mathbb{R}^K}{minimize} \quad \frac{1}{n}\sum_{j=1}^{n}\left[-T_j n\rho\left(B(X_j)^T\lambda\right) + B(X_j)^T\lambda\right] + |\lambda|^T\delta, \tag{2.5}$$

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*

$$w_j^* = \rho^{'}\left(B(X_j)^T\lambda^\dagger\right) \tag{2.6}$$

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 2.1.4, 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 [8] 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*

$$\mathbf{P}\left(\left\|\lambda^\dagger - \lambda_1^*\right\|_2 \leq c^3\tilde{c_\tau}(\varepsilon_n^3)\right) \geq 1 - \tau. \tag{2.7}$$

It is then straightforward to prove a more general result than Theorem 2.3.

Theorem 2.4. *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*

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

where $w^*(X)$ is as in (2.6) without the index.

Proof of theorem 2.2

Proof. Motivated by Proposition 3.1 we consider

$$G(\lambda) := \frac{1}{n}\sum_{j=1}^n \left[-T_j n\rho\left(B(X_j)^T\lambda\right) + B(X_j)^T\lambda\right] + |\lambda|^T\delta. \tag{2.8}$$

Since $\rho \in C^2(\mathbb{R})$ we can employ (2.8), Corollary 6.1.1 and Proposition 6.1 to get

$$\begin{aligned}
&G(\lambda_1^* + \Delta) - G(\lambda_1^*) \\
&\geq \frac{1}{n}\sum_{j=1}^n \left[-T_j n\rho'\left(B(X_j)^T\lambda_1^*\right) + 1\right]\Delta^T B(X_j) \\
&\quad + \frac{1}{2}\sum_{j=1}^n -T_j\rho''\left(B(X_j)^T(\lambda_1^* + \xi\Delta)\right)\Delta^T\left(B(X_j)B(X_j)^T\right)\Delta \\
&\quad - |\Delta|^T\delta \\
&\geq -\|\Delta\|_2\left(\left\|\frac{1}{n}\sum_{j=1}^n\left[-T_j n\rho'\left(B(X_j)^T\lambda_1^*\right) + 1\right]B(X_j)\right\|_2 + \|\delta\|_2\right) \\
&\quad + n\|\Delta\|_2^2\,\varphi_{\rho''}\underline{\varphi_{BB^T}} \\
&:= -\|\Delta\|_2\left(I_1 + \|\delta\|_2\right) + \|\Delta\|_2^2 I_2.
\end{aligned} \tag{2.9}$$

The second inequality is due to the Cauchy-Schwarz-Inequality and Assumptions 2.1.6 and 2.1.7 . 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 (2.9) yields $G(\lambda_1^* + \Delta) - G(\lambda_1^*) > 0$ with arbitrarily large probabitity for n large enough. With Proposition 3.1 it follows then immediately Proposition 2.2 .

Analysis of I_1

We want to use Assumption 2.1.3. Thus we perform the following split:

$$\begin{aligned}
I_1 &\leq \left\|\sum_{j=1}^n T_j\left[\rho'\left(B(X_j)^T\lambda_1^*\right) - \frac{1}{n\pi(X_j)}\right]B(X_j)\right\|_2 \\
&\quad + \left\|\frac{1}{n}\sum_{j=1}^n\left[\frac{T_j}{\pi(X_j)} - 1\right]B(X_j)\right\|_2 \\
&=: J_1 + J_2
\end{aligned} \tag{2.10}$$

Analysis of J_1

By the Lipschitz-continuity of ρ', Assumption 2.1.8 and Assumption 2.1.4, $T \in \{0,1\}$ and the triangle inequality we have

$$J_1 \leq nL_{\rho'}\varphi_{\|B(x)\|}\varphi_{m^*} \tag{2.11}$$

Analysis of J_2

We want to employ Theorem 4.2. To this end we define the independent random matrices

$$\begin{aligned} A_j &:= \frac{1}{n}\left[\frac{T_j}{\pi(X_j)} - 1\right] B(X_j)\,, \quad j = 1,\dots,n, \\ S &:= \sum_{j=1}^{n} A_j \end{aligned} \tag{2.12}$$

and check conditions (4.12) and (4.3). Note that $\|S\|_2 = J_2$. By the properties of conditional expectation it holds

$$\mathbf{E}\left[\frac{T_j}{\pi(X_j)}B(X_j)\right] = \mathbf{E}\left[\mathbf{E}\left[T_j \mid X_j\right]\frac{1}{\pi(X_j)}B(X_j)\right] = \mathbf{E}[B(X_j)]. \tag{2.13}$$

Taking the expectation in (2.12) and using (2.13) we get $\mathbf{E}[A_j] = 0$ for all $j = 1,\dots,n$. Since

$$\left|\frac{T_j}{\pi(X_j)} - 1\right| \leq 1 + \frac{1-\varphi_\pi}{\varphi_\pi} = \frac{1}{\varphi_\pi} \tag{2.14}$$

by Assumption 2.1.5, we can employ Assumption 2.1.8 together with (2.14) and (2.12) to get

$$\|A_j\|_2 \leq \frac{\varphi_{\|B\|}}{n\varphi_\pi} =: L. \tag{2.15}$$

Thus, condition (4.12) is satisfied. Next we turn to the matrix variance statistic $v(S)$ (4.3). By (2.12) and (2.14) we have

$$\mathbf{E}\left[A_j A_j^T\right] \leq \left(\frac{1}{n\varphi_\pi}\right)^2 \mathbf{E}\left[B(X)B(X)^T\right] \tag{2.16}$$

and by (2.15)

$$\mathbf{E}\left[A_j^T A_j\right] \leq L^2. \tag{2.17}$$

Since $\max\{a,b\} \leq |a| + |b|$ we can use (2.16) and (2.17) to get

$$v(S) \leq \frac{1}{n}\frac{\lambda_{\max}}{\varphi_\pi^2} + nL^2, \tag{2.18}$$

where $\lambda_{\max}$ is the maximal eigenvalue of the symmetric (non-random) matrix $\mathbf{E}\left[B(X)B(X)^T\right]$. Having dealt with (4.12) and (4.3) we can establish the expectation bound (4.4) of Theorem 4.2. Together with (2.15) and (2.18) we get

$$\begin{aligned} &\mathbf{E}[J_2] \\ &\leq \sqrt{\frac{2\log(K+1)\left(\lambda_{\max} + \varphi_{\|B\|}^2\right)}{n\varphi_\pi^2}} + \frac{\log(K+1)\varphi_{\|B\|}}{3n\varphi_\pi} \\ &\leq \frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(\sqrt{2} + \frac{1}{3}\sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{2\lambda_{\max}}\right]. \end{aligned} \tag{2.19}$$

Since $K = o(n)$ by Assumption 2.1.9 we can discuss the other influences on the quality of the bound (2.19). On a high-level it is readily clear that appropriate bounds on $\varphi_\pi, \varphi_{\|B\|}$ and $\lambda_{\max}$ will shrink $\mathbf{E}[J_2]$ to 0 and will assist in establishing learning rates.

We could also have invoked the probabitity bound (4.5) of Theorem 4.2. But for the sake of simplicity we prefer the combination of the expectation bound (2.19) and the Markov inequality. With the latter we get

$$J_2 \leq \frac{1}{\tau}\frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(\sqrt{2}+\frac{1}{3}\sqrt{\frac{\log(K+1)}{n}}\right)+\sqrt{2\lambda_{\max}}\right] \tag{2.20}$$

with probabitity $\geq 1-\tau$.

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

$$\begin{aligned}\left(\sqrt{2}\frac{1}{\tau}\frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n^3}}\left[\varphi_{\|B\|}\left(1+\sqrt{\frac{\log(K+1)}{n}}\right)+\sqrt{\lambda_{\max}}\right]\right.\\ \left.+\,L_{\rho'}\varphi_{\|B\|}\varphi_{m^*}+\frac{\|\delta\|_2}{n}\right)\frac{1}{\varphi_{\rho''}\underline{\varphi_{BB^T}}}\end{aligned} \tag{2.21}$$

we get by (2.9), (2.10), (2.11), (2.20) and Proposition 3.1

$$\begin{aligned}\mathbf{P}\left(\left\|\lambda^\dagger-\lambda_1^*\right\|_2\leq C\right)&=\mathbf{P}\left(\inf_{\|\Delta\|_2=C}G(\lambda_1^*+\Delta)-G(\lambda_1^*)>0\right)\\ &\geq 1-\tau,\end{aligned} \tag{2.22}$$

where C is as in (2.21). With appropriate Assumptions (as discussed before) we can then establish Proposition 2.2.

We can invoke (2.22) to derive bounds as in Theorem 2.4:

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

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

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

Remark 2.1. *By Corollary 4.2.1 we can get rid of the* $\log(K)$ *term in* (2.21). ♢

Remark 2.2. *By the matrix Rosenthal-Pinelis Inequality [1][Thm.A.1] we can weaken Assumption 2.1.5 to a lower bound on the expectition of* $\pi(X)$ ♢

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

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 [6] 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*

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

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

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

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.*

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*

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{3.4}$$

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 3.4. (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*

$$(f_1 \square \ldots \square f_m)(x) := \inf_{\substack{x_i \in \mathbb{R}^n \\ \sum_{i=1}^m x_i = x}} \sum_{i=1}^m f_i(x_i) \tag{3.5}$$

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*

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*) \tag{3.6}$$

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

$$(f + g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{3.7}$$

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*

$$g(x) = \delta_\Omega(x) := \tag{3.8}$$

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

The Fenchel dual problem is then

$$\text{maximize} \quad -f^*\left(A^T y\right) - g^*(-y) \quad \text{subject to} \quad y \in \mathbb{R}^n. \tag{3.9}$$

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 (3.9) are equal, i.e.*

$$\inf_{x \in \mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y \in \mathbb{R}^n} \left\{-f^*\left(A^T y\right) - g^*(-y)\right\}. \tag{3.10}$$

Lemma 3.1. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be convex. Then for all $y \in \mathbb{R}^n$ and $C > 0$*

$$\inf_{\|\Delta\|=C} f(y+\Delta) - f(y) \geq 0 \quad \Longrightarrow \quad \exists y^* \in \mathbb{R}^n \colon y^* \text{ is global minimum of } f \text{ and } \|y^* - y\| \leq C. \tag{3.11}$$

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 (3.11) is true. Then it holds

$$f(x) < f(y^*) \quad \text{for some } x \in \mathbb{R}^n \setminus y + \mathcal{C}. \tag{3.12}$$

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.

$$\theta x + (1-\theta)y^* = y + \Delta_x \quad \text{for some } \theta \in (0,1) \text{ and } \Delta_x \text{ with } \|\Delta_x\| = C. \tag{3.13}$$

It follows

$$\begin{aligned} f(y^*) \leq f(y) \leq f(y + \Delta_x) &= f(\theta x + (1-\theta)y^*) \\ &\leq \theta f(x) + (1-\theta) f(y^*) < f(y^*), \end{aligned} \tag{3.14}$$

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 (3.11) being true, the equality is due to (3.13), the third inequality is due to the convexity of f and the strict inequality is due to (3.12). □

4 Matrix Concentration Inequalities

Definition 4.1. (Hermitian Dilation) *The Hermitian dilation*

$$\mathfrak{H} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{H}_{d_1 \times d_2}$$

is a map from a general matrix to an Hermitian matrix defined by

$$\mathfrak{H}(B) := \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} \tag{4.1}$$

Theorem 4.1. (Matrix Bernstein Inequality) *Let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be independent, random matrices with dimension $d_1 \times d_2$. Assume that*

$$\mathbf{E}(\mathbf{A}_k) = 0 \quad \textit{and} \quad \|\mathbf{A}_k\| \leq L \quad \textit{for each} \quad k \in \{1, \ldots, n\}. \tag{4.2}$$

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:

$$v(\mathbf{S}) := \left\|\mathbf{E}[\mathbf{S}\mathbf{S}^\top]\right\| \vee \left\|\mathbf{E}[\mathbf{S}^\top\mathbf{S}]\right\| = \left\|\sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k\mathbf{A}_k^\top]\right\| \vee \left\|\sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k^T\mathbf{A}_k]\right\|. \tag{4.3}$$

Then

$$\mathbf{E}\left[\|\mathbf{S}\|\right] \leq \sqrt{2v(\mathbf{S})\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2). \tag{4.4}$$

Furthermore,

$$\mathbf{P}[\|\mathbf{S}\| \geq t] \leq (d_1+d_2)\exp\left(\frac{-t^2/2}{v(\mathbf{S}) + Lt/3}\right) \quad \textit{for all } t \geq 0. \tag{4.5}$$

Corollary 4.1.1. (Scalar Bernstein Inequality) *Let $X_1, \ldots, X_n$ be centered, independent real random variables with $\|X_i\|_\infty \leq L$ and $\mathbf{E}(X_1^2) \leq \sigma^2$ for all $i = 1, \ldots, n$. Then it holds for all $t \geq 0$*

$$\mathbf{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{1}{2}\frac{t^2}{n\sigma^2 + Lt/3}\right) \tag{4.6}$$

Proof. We verify the scalar versions of (4.12) and (4.3). Since the 2- and ∞- norm coincide in one dimension and the X_i are centered and independent we get

$$v(\mathbf{S}) = \sum_{i=1}^{n} \mathbf{E}(X_i^2) \leq n\sigma^2 \tag{4.7}$$

We get (4.6) by applying Theorem 4.2. □

Definition 4.2. Intrinsic Dimension *For a positive-semidefinite matrix* $\mathbf{A}$*, the intrinsic dimension is the quantity*

$$\operatorname{intdim}(\mathbf{A}) := \frac{\operatorname{tr}\mathbf{A}}{\|\mathbf{A}\|_2},$$

where tr is the trace of a matrix.

Theorem 4.2. (Intrinsic Matrix Bernstein) *Let* $(\mathbf{A}_k)_{1\le k\le n}$ *be a finite sequence of independent, random matrices with the same size. Assume that*

$$\mathbf{E}(\mathbf{A}_k) = 0 \quad \textit{and} \quad \|\mathbf{A}_k\| \le L \quad \textit{for each} \quad k \in \{1,\dots,n\}. \tag{4.8}$$

Introduce the random matrix

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

Let $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *be semidefinite upper bounds for the matrix-valued variances* $\mathbf{Var}_1(\mathbf{S})$ *and* $\mathbf{Var}_2(\mathbf{S})$*:*

$$\mathbf{V}_1 \succcurlyeq \mathbf{Var}_1(\mathbf{S}) := \mathbf{E}(\mathbf{S}) = \sum_{k=1}^{n} \mathbf{E}(\mathbf{A}_k\mathbf{A}_k^\top),$$
$$\mathbf{V}_2 \succcurlyeq Var_2(\mathbf{S}) := \mathbf{E}(\mathbf{S}) = \sum_{k=1}^{n} \mathbf{E}(\mathbf{A}_k^\top\mathbf{A}_k).$$

Define an intrinsic dimension bound and a variance bound

$$d := \operatorname{intdim}\begin{bmatrix}\mathbf{V}_1 & 0\\ 0 & \mathbf{V}_2\end{bmatrix} \qquad \textit{and} \qquad v := \max\left\{\|\mathbf{V}_1\|_2, \|\mathbf{V}_2\|_2\right\}. \tag{4.9}$$

Then it holds

$$\mathbf{E}\,\|\mathbf{S}\| \le \textit{Const.}\left(\sqrt{v\log(d+1)} + L\log(d+1)\right). \tag{4.10}$$

Furthermore, for all $t \ge \sqrt{v} + \frac{L}{3}$*,*

$$\mathbf{P}(\|\mathbf{S}\| \ge t) \le 4d\exp\left(\frac{-t^2/2}{v + Lt/3}\right). \tag{4.11}$$

It turns out that for random vectors $d \le 2$. This is remarkable, because this property of the intrinsic dimension is invarient under the dimension of the vector. This fact motivates the following result:

Corollary 4.2.1. *Let* $(\mathbf{A}_k)_{1\le k\le n} \subseteq \mathbb{R}^K$ *be a finite sequence of independent, random vectors. Assume that*

$$\mathbf{E}(\mathbf{A}_k) = 0 \quad \textit{and} \quad \|\mathbf{A}_k\| \le L \quad \textit{for each} \quad k \in \{1,\dots,n\}. \tag{4.12}$$

Let $v(\mathbf{S})$ *be the matrix variance statistic of the sum as defined in* (4.3) *. Then it holds*

$$\mathbf{E}\,\|\mathbf{S}\| \le \textit{Const.}\left(\sqrt{v(\mathbf{S})\log(3)} + L\log(3)\right). \tag{4.13}$$

Furthermore, for all $t \ge \sqrt{v(\mathbf{S})} + \frac{L}{3}$*,*

$$\mathbf{P}(\|\mathbf{S}\| \ge t) \le 8\exp\left(\frac{-t^2/2}{v(\mathbf{S}) + Lt/3}\right). \tag{4.14}$$

Proof. First, let's verify

$$d := \operatorname{intdim} \begin{bmatrix} \operatorname{Var}_1(\mathbf{S}) & 0 \\ 0 & \operatorname{Var}_2(\mathbf{S}) \end{bmatrix} \leq 2. \tag{4.15}$$

Since

$$\operatorname{tr} BB^\top = B^\top B \quad \text{and} \quad \left\| \begin{bmatrix} BB^\top & 0 \\ 0 & B^\top B \end{bmatrix} \right\|_2 = \max \left\{ \left\| BB^\top \right\|, B^\top B \right\} \tag{4.16}$$

for all $B \in \mathbb{R}^K$ it follows

$$\operatorname{intdim} \begin{bmatrix} BB^\top & 0 \\ 0 & B^\top B \end{bmatrix} = \frac{\operatorname{tr} BB^\top + B^\top B}{\max \left\{ \left\| BB^\top \right\|, B^\top B \right\}} \leq 2 \tag{4.17}$$

for all $B \in \mathbb{R}^K$. By the linearity of the expectation it holds for random vectors $X = (X_1, \ldots, X_K)$ with $\|X\|_\infty < \infty$

$$\mathbf{E}(\operatorname{tr} XX^\top) = \operatorname{tr} \mathbf{E}(XX^\top). \tag{4.18}$$

Indeed

$$\begin{aligned} \mathbf{E}(\operatorname{tr} XX^\top) &= \mathbf{E} \left(\sum_{i=1}^n X_i^2 \right) = \sum_{i=1}^n \mathbf{E}(X_i^2) = \operatorname{tr} \left(E(X_i X_j) \right)_{1 \leq i,j \leq K} \\ &= \operatorname{tr} \mathbf{E}(XX^\top) \end{aligned}$$

Note that by the linearity of the trace and the scalar product, the above also holds for finite sums of bounded random vectors. We have thus established (4.15). Applying Theorem 4.2 finishes the proof. □

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] Richard Y. Chen, Alex Gittens, and Joel A. Tropp. The Masked Sample Covariance Estimator: An Analysis via Matrix Concentration Inequalities, June 2012.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

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

[12] 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.