Todo list

Solving missing survival times with entropy balancing weights

Universität Stuttgart

Ioan Scheffel

February 16, 2023

Contents

1 Introduction

We consider a study population in which we want to test the effect of a treatment. We introduce the **indicator of treatment** $T \in \{0,1\}$. For each treatment level there exist the **marginal potential outcomes** $(Y(0), Y(1))$. We would like to estimate $\mathbf{E}[Y(1)]$. If we succeed the same technique shall yield an estimate of $\mathbf{E}[Y(0)]$. We shall compare $\mathbf{E}[Y(1)]$ and $\mathbf{E}[Y(0)]$ and find out something about the effect of the treatment in the population.

The data we acquire is independent and identically distributed. But usually

$$Y(1)|T = 1 \nsim Y(1)\,, \tag{1.1}$$

that is, $T = 1$ carries more information than observing the outcome under treatment. We say that $Y(1)|T = 1$ is **confounded**. To extract that plus of information from $T = 1$ and put it where it belongs by collecting more data. We gather it in $X \in \mathbb{R}^d$ and assume

$$(Y(0), Y(1)) \perp T \mid X\,, \tag{1.2}$$

that is, **conditional unconfoundedness**. Thus, we end up collecting $N \in \mathbb{N}$ independent and identically distributed copies of $(T, X, Y(T))$. For convenience, we assume that the first $n \in \mathbb{N}$ copies have $T = 1$.

A natural estimator for $\mathbf{E}[Y(1)]$ is the weighted mean

$$\frac{1}{n}\sum_{i=1}^{n} w_i Y_i\,. \tag{1.3}$$

The weights should satisfy (in a broader sense)

$$w_i \cdot Y_i \to Y(1) \qquad \text{for } N \ \to \ \infty\,. \tag{1.4}$$

One class of such weights has been recently analyzed in [WZ19]. We take ideas and extend.

The algorithm

Problem 1.1.

$$
\begin{aligned}
&\underset{w_1,\dots,w_n\in\mathbb{R}}{\text{minimize}} && \sum_{i=1}^{n} f(w_i) && \\
&\text{subject to} && w_i \geq 0 && \text{for all } i \in \{1,\dots,n\}\ , \\
& && \frac{1}{N}\sum_{i=1}^{n} w_i = 1 && \\
& && \left|\frac{1}{N}\left(\sum_{i=1}^{n} w_i B_k(X_i) - \sum_{i=1}^{N} B_k(X_i)\right)\right| \ \leq\ \delta_k && \text{for all } k \in \{1,\dots,N\}\ .
\end{aligned}
$$

This is a (convex) optimization problem. We will talk about the **objective function** f and the **equality** and **inequality constraints**, especially about the **regression basis** B .

Objective Function

Strictly speaking, we consider the sum

$$[w_1,\dots,w_n]^\top \ \mapsto\ \sum_{i=1}^{n} f(w_i) \tag{1.5}$$

as the objective function. It is natural to consider the dual formulation of the optimization problem. This involves the **convex conjugate**(cf.Definition ?) of the original objective function. We show in Example that for the sum this is

$$[\lambda_1,\dots,\lambda_n]^\top \ \mapsto\ \sum_{i=1}^{n} f^*(\lambda_i) \tag{1.6}$$

where f^* is the Legendre transformation of f .

In the sequel we need f to be strictly convex and its convex conjugate (or Legendre transformation) to be continuously differentiable and strictly non-decreasing. Two popular choices of f are the **negative entropy** and the **sample variance**.

Negative Entropy

We define the negative entropy to be

$$f\colon [0,\infty) \to \mathbb{R}, \quad w \mapsto \begin{cases} 0 & \text{if } w = 0, \\ w\log w & \text{else.} \end{cases} \tag{1.7}$$

It is strictly convex. To compute its Legendre transformation we note, that

$$(f')^{-1} = \lambda \mapsto e^{\lambda - 1} \tag{1.8}$$

Thus

$$\begin{aligned} f^*(\lambda) &= \lambda \cdot (f')^{-1}(\lambda) \; - \; f\left((f')^{-1}(\lambda)\right) \\ &= \lambda \cdot e^{\lambda-1} \; - \; e^{\lambda-1} \log\left(e^{\lambda-1}\right) \\ &= e^{\lambda-1} \, . \end{aligned}$$

Thus f^* is smooth and strictly non-decreasing.

Sample Variance

We define the sample variance to be

$$f \colon \mathbb{R} \to \mathbb{R}, \quad w \mapsto (w - 1/n)^2 \tag{1.9}$$

It is strictly convex. To compute its Legendre transformation we note, that

$$(f')^{-1} = \lambda \mapsto \frac{\lambda}{2} + \frac{1}{n} \tag{1.10}$$

Thus

$$\begin{aligned} f^*(\lambda) &= \lambda \cdot \left(\frac{\lambda}{2} + \frac{1}{n}\right) \; - \; \left(\left(\frac{\lambda}{2} + \frac{1}{n}\right) - \frac{1}{n}\right)^2 \\ &= \frac{\lambda^2}{4} + \frac{\lambda}{n} \, . \end{aligned}$$

Thus f^* is smooth. To eliminate some variables in the optimization problem, we need f^* also to be strictly non-decreasing. But the sample variance violates this assumption.

Constraints

Let's turn our attention to the constraints. The first constraint makes sure we do not extrapolate from the poputation. The second constraint norms the weights. The third constraint controls the bias of the resulting estimator.

Regression Basis

We adopt ideas from [GKKW02]. Another angle would be sieve estimates [New97] where the number of basis functions can grow slower than N. Their notion of (weak) consistency [GKKW02, Definitien 1.1] for noiseless estimands is

$$\mathbf{E}\left[\int_{\mathcal{X}}\left|\sum_{k=1}^{N} B_k(x)\cdot m(X_k) - m(x)\right|^2 \mathbf{P}_X(dx)\right] \to 0 \qquad \text{as } n \to \infty\,. \tag{1.11}$$

Universal consistency in this sense holds, if this is true for all distributions with $\mathbf{E}[m(X)^2] < \infty$ (cf. [GKKW02, Definition 1.3]).

We adopt a slightly different notion of consistency. The next theorem dose the translation work.

Theorem 1.1. *Assume $\mathbf{E}[m(X)^2] < \infty$ and the basis function are (weak) universal consistency in the sense of [GKKW02, Definitien 1.3]. Then it holds for all $\varepsilon > 0$*

$$\mathbf{P}\left[\left|\sum_{k=1}^{N} B_k(X)\cdot m(X_k) - m(X)\right| \geq \varepsilon\right] \to 0 \qquad \textit{as } n \to \infty\,. \tag{1.12}$$

Proof. By Markov's inequality it holds

$$\begin{aligned}
&\mathbf{P}\left[\left|\sum_{k=1}^{N} B_k(X)\cdot m(X_k) - m(X)\right| \geq \varepsilon\right] \\
&\leq \frac{\mathbf{E}\left[\left|\sum_{k=1}^{N} B_k(X)\cdot m(X_k) - m(X)\right|^2\right]}{\varepsilon^2} \\
&= \frac{\mathbf{E}\left[\mathbf{E}\left[\left|\sum_{k=1}^{N} B_k(X)\cdot m(X_k) - m(X)\right|^2 |X_1,\ldots,X_N\right]\right]}{\varepsilon^2} \\
&= \frac{\mathbf{E}\left[\int_{\mathcal{X}}\left|\sum_{k=1}^{N} B_k(x)\cdot m(X_k) - m(x)\right|^2 \mathbf{P}_X(dx)\right]}{\varepsilon^2}\,.
\end{aligned}$$

The last equality is due to [GKKW02, (1.2)]. By the weak universal consistency of B the last expression goes to 0 as $N \to \infty$. □

Classical choices of the basis functions are **partitioning estimates** and **kernel estimates**(cf. [GKKW02, §4,§5]).

Partitioning Estimates

We consider a partition $\mathcal{P}_N = \{A_{N,1}, A_{N,2}, \ldots\}$ of $\mathbb{R}^d$ and define $A_N(x)$ to be the cell of $\mathcal{P}_N$ containing x. We define N basis functions B_k of the covariates by

$$B_k(x) := \frac{\mathbf{1}_{X_k \in A_N(x)}}{\sum_{j=1}^{N} \mathbf{1}_{X_j \in A_N(x)}}\,, \qquad k = 1,\ldots,N\,.$$

The euclidian norm of the basis functions is bounded above by 1.

$$\|B(x)\|^2 = \sum_{k=1}^{n} \left(\frac{\mathbf{1}_{X_k \in A_n(x)}}{\sum_{j=1}^{n} \mathbf{1}_{X_j \in A_n(x)}} \right)^2 \leq \sum_{k=1}^{n} \frac{\mathbf{1}_{X_k \in A_n(x)}}{\sum_{j=1}^{n} \mathbf{1}_{X_j \in A_n(x)}} = 1\,.$$

Under mild conditions, the basis functions are universally consistent.

Theorem 1.2. *If for each sphere S centered at the origin*

$$\max_{j\colon A_{N,j} \cap S \neq \emptyset} \operatorname{diam} A_{N,j} \;\to\; 0 \qquad \text{for } N \to \infty \tag{1.13}$$

and

$$\frac{\#\{j\colon A_{N,j} \cap S \neq \emptyset\}}{N} \;\to\; 0 \qquad \text{for } N \to \infty \tag{1.14}$$

then the partitioning regression function estimate (definition) is universally consistent (definition).

Proof. [GKKW02, Theorem 4.2.] □

Corollary 1.2.1. *Assume $\mathbf{E}[m(X)^2] < \infty$ and the basis functions B belong to a partitioning estimate. Furthermore assume that the conditions of Theorem 1.2 are met. Then it holds for all $\varepsilon > 0$*

$$\mathbf{P}\left[\left| \sum_{k=1}^{N} B_k(X) \cdot m(X_k) - m(X) \right| \geq \varepsilon \right] \to 0 \qquad \text{as } n \to \infty\,. \tag{1.15}$$

Kernel Estimates

Let $K\colon \mathbb{R}^d \to [0,1]$ (bounded kernel) and $h_n > 0$ (bandwith). For examples see [GKKW02, §5.1.]. We define

$$B_k(x) := \frac{K\left(\frac{x - X_k}{h_n}\right)}{\sum_{i=1}^{N} K\left(\frac{x - X_i}{h_n}\right)}\,. \tag{1.16}$$

By the boundedness of the kernel it follows $\|B(x)\| \leq 1$.

Theorem 1.3. *Assume that there are balls $S_{0,r}$ of radius r and balls $S_{0,R}$ of radius R centered at the origin with $0 < r \leq R$, and a constant $b > 0$ such that*

$$\mathbf{1}_{\{x \in S_{0,R}\}} \geq K(x) \geq b \cdot \mathbf{1}_{\{x \in S_{0,r}\}} \tag{1.17}$$

(boxed kernel). Then for bandwiths with $h_n \to 0$ and $n \cdot h_n^d \to \infty$ as $n \to \infty$ the kernel estimate is weakly universally consistent.

Corollary 1.3.1. *Assume* $\mathbf{E}[m(X)^2] < \infty$ *and the basis functions* B *belong to a kernel estimate. Furthermore assume that the conditions of Theorem 1.3 are met. Then it holds for all* $\varepsilon > 0$

$$\mathbf{P}\left[\left|\sum_{k=1}^{N} B_k(X)\cdot m(X_k) - m(X)\right| \geq \varepsilon\right] \to 0 \qquad \textit{as } n \to \infty\,. \tag{1.18}$$

In the sequel we mainly work with the dual problem.

Dual Problem

Theorem. *The dual of Problem 1.1 is the unconstrained optimization problem*

$$\underset{\lambda_0,\dots,\lambda_N \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^{N}\left[\, T_i \cdot f^*(\lambda_0 + \langle B(X_i), \lambda\rangle) \;-\; (\lambda_0 + \langle B(X_i), \lambda\rangle)\,\right] \;+\; \langle \delta, |\lambda|\rangle\,.$$

where

$$f^* : \mathbb{R} \;\to\; \mathbb{R}\,, \qquad x^* \;\mapsto\; x^* \cdot (f')^{-1}(x^*) \;-\; f\left((f')^{-1}(x^*)\right)$$

is the Legendre transformation of f *, the vector* $B(X_i) = [B_1(X_i), \dots, B_n(X_i)]^\top$ *denotes the* N *basis functions of the covariates of unit* $i \in \{1, \dots, N\}$ *and* $|\lambda| = [|\lambda_1|, \dots, |\lambda_N|]^\top$ *, where* $|\cdot|$ *is the absolute value of a real-valued scalar. Moreover, if* $\lambda^\dagger$ *is an optimal solution of the above problem then the optimal solution to problem Problem 1.1 is given by*

$$w_i^\dagger \;=\; (f')^{-1}\left(\langle B(X_i), \lambda^\dagger\rangle + \lambda_0^\dagger\right) \qquad \textit{for } i \in \{1\dots, n\}\;.$$

Plan of proof

We want to apply Theorem 4.4. To this end, we find the suitable **matrix notation**. ([WZ19, p.20-22] fail to do so. The problem is, that they divide by 0 in the second display on p.21). Theorem 4.4 covers only parts of the constraints, so we apply the argument in [WZ19, p.19-20] to eliminate the remaining **non-negativity constraints**.

Proof. **Matrix notation**

We consider the vector of basis functions of the covariates of unit $i \in \{1, \dots, n\}$, that is,

$$B(X_i) \;:=\; [B_1(X_i), \dots, B_N(X_i)]^\top\;,$$

the constraints vectors

$$\begin{aligned} d &:= \begin{bmatrix} 0_n \\ -N\cdot\delta \pm \sum_{i=1}^{N} B_k(X_i) \end{bmatrix}, \\ a &:= N \end{aligned}$$

the matrix of the basis functions of the treated

$$\mathbf{B}(\mathbf{X}) := \Big[B(X_1),\ldots,B(X_n)\Big]$$

and the constraint matrices

$$\begin{aligned} \mathbf{U} &:= \begin{bmatrix} \mathbf{I}_n \\ \pm\,\mathbf{B}(\mathbf{X}) \end{bmatrix}. \\ \mathbf{A} &:= 1_n \end{aligned}$$

By Example 4.1 the convex conjugate of the objective function of Problem 1.1 is

$$[x_1^*,\ldots,x_n^*]^\top \;\mapsto\; \sum_{i=1}^{n} f^*(x_i^*)\,,$$

Before we apply Theorem 4.4 we eliminate the non-negativity constraints. To this end, we consider the objective function G of the dual problem and update it until we reach its final form. We write

$$\lambda_d =: \begin{bmatrix} \rho \\ \lambda^+ \\ \lambda^- \end{bmatrix} \tag{1.19}$$

$$\begin{aligned} G(\lambda_d,\lambda_0) &= G\left(\rho,\lambda^+,\lambda^-,\lambda_0\right) \\ &:= \sum_{i=1}^{N} -f^*\left(\rho_i+\lambda_0+\langle B(X_i),\lambda^+-\lambda^-\rangle\right) \;+\; \left(\lambda_0+\langle B(X_i),\lambda^+-\lambda^-\rangle\right) \\ &\quad -\; N\cdot\langle\delta,\lambda^+ + \lambda^-\rangle \end{aligned}$$

Since we maximize G and f^* is strictly non-decreasing, $\rho = 0$ is optimal. We update G.

$$\begin{aligned} G\left(\lambda^+,\lambda^-,\lambda_0\right) &= \sum_{i=1}^{N} -f^*\left(\lambda_0+\langle B(X_i),\lambda^+-\lambda^-\rangle\right) \;+\; \left(\lambda_0+\langle B(X_i),\lambda^+-\lambda^-\rangle\right) \\ &\quad -\; N\cdot\langle\delta,\lambda^+ + \lambda^-\rangle \end{aligned}$$

Non-negativity constraints

Next we want to remove the non-negativity constraints on $\lambda^{\pm}$. We show for all $i \in \{1, \dots, N\}$

$$\begin{aligned} \text{either} \qquad & \lambda_i^+ > 0 \\ \text{or} \qquad & \lambda_i^- > 0\,. \end{aligned}$$

Assume towards a contradiction that there exists $i \in \{1, \dots, N\}$ such that $\lambda_i^+ > 0$ and $\lambda_i^- > 0$ and that $\lambda^{\pm}$ is optimal. Consider

$$\tilde{\lambda} \;:=\; \Big[\, \lambda_1^+, \dots,\; \lambda_i^+ - (\lambda_i^+ \wedge \lambda_i^-)\,,\; \dots, \lambda_N^+,\; \lambda_1^-, \dots, \lambda_i^- - (\lambda_i^+ \wedge \lambda_i^-)\,,\; \dots, \lambda_N^-\,, \lambda_0 \Big]^\top . \tag{1.20}$$

Since $\lambda_i^{\pm} - (\lambda_i^+ \wedge \lambda_i^-) \geq 0$, the perturbed vector $\tilde{\lambda}$ is in the domain of the optimization problem. But

$$G(\tilde{\lambda}) - G(\lambda) \;=\; 2N \cdot \delta_i \cdot (\lambda_i^+ \wedge \lambda_i^-) \;>\; 0\,, \tag{1.21}$$

which contradicts the optimality of λ. But then $\lambda_i^{\pm} \geq 0$ collapses to $\lambda_i \in \mathbb{R}$ for all $i \in \{0, \dots, N\}$, that is, $\lambda_i = \lambda_i^+ - \lambda_i^-$. Note that $|\lambda_i| = \lambda_i^+ + \lambda_i^-$.

We update the objective function one more time. Multiplying with $-1/N$ and introducing T we get

$$\underset{\lambda_0, \dots, \lambda_N \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} \big[\, T_i \cdot f^*(\lambda_0 + \langle B(X_i), \lambda \rangle) \;-\; (\lambda_0 + \langle B(X_i), \lambda \rangle) \,\big] \;+\; \langle \delta, |\lambda| \rangle\,.$$

We apply Theorem 4.4 to finish the proof. □

We have gathered all the tools to tackle consistency of the weighted mean.

2 Consistency

Throughout this section assume the existence of an optimal solution $(\lambda^\dagger, \lambda_0^\dagger)$. We use a hint from the last display of [WZ19, p.22]. The high-level idea is, to connect the optimality of a dual solution to being in the neighborhood of an oracle parameter by looking at the objective function of the dual. We deliver the omitted technical details.

Neighbourhood of Oracle Parameter

Let λ^* denote the vector with coordinates

$$\lambda_i^* := f^{'}(1/\pi_i) - \lambda_0^\dagger\,, \tag{2.1}$$

where $\pi_i = \mathbf{P}[T_i = 1|X_i]$ is the **propensity score** of the i-th unit.

Theorem 2.1. *For all $\varepsilon > 0$ it holds*

$$\mathbf{P}\left[\left\|\lambda^\dagger - \lambda^*\right\| \geq \varepsilon\right] \;\to\; 0 \qquad \textit{for } N \to \infty\,. \tag{2.2}$$

We want to leverage the convexity of the objective function of the dual to get

$$\mathbf{P}\left[\left\|\lambda^\dagger - \lambda^*\right\| \leq \varepsilon\right] = \mathbf{P}\left[\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0\right]\,.$$

We learned about a similar idea from [WZ19, p.22]. The next Lemma makes this rigorous.

Lemma 2.1. *Let $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ be convex. Then for all $y \in \mathbb{R}^m$ and $\varepsilon > 0$*

$$\inf_{\|\Delta\|=\varepsilon} g(y + \Delta) - g(y) \geq 0 \tag{2.3}$$

implies the existence of a global minimum $y^* \in \mathbb{R}^m$ of g satisfying $\|y^* - y\| \leq \varepsilon$.

Proof. Since $y + \varepsilon B$ is convex, it contains a local minimum of g. Suppose towards a contradiction that $y^* \in y + \varepsilon B$ is a local minimum, but not a global one, and (2.3) is true. Then it holds

$$g(x) < g(y^*) \quad \text{for some } x \in \mathbb{R}^m \setminus (y + \varepsilon B)\,. \tag{2.4}$$

Furthermore, since $y + \varepsilon B$ is compact and contains y^*, the line segment connecting y^* and x intersects the boundary of $y+\mathcal{C}$, that is, there exist $\theta \in (0,1)$ and Δ_x with $\|\Delta_x\| = \varepsilon$ such that

$$\theta x + (1-\theta)y^* = y + \Delta_x \,. \tag{2.5}$$

It follows

$$\begin{aligned} g(y^*) \le g(y) \le g(y+\Delta_x) &= g(\theta x + (1-\theta)y^*) \\ &\le \theta g(x) + (1-\theta)g(y^*) < g(y^*), \end{aligned} \tag{2.6}$$

which is a contradiction. The first inequality is due to y^* being a local minimum of g in $y + \varepsilon B$, the second inequality is due to (2.3) being true, the equality is due to (2.5), the third inequality is due to the convexity of g and the strict inequality is due to (2.4). Thus every local minimum of g in $y + \varepsilon B$ is also a global minimum. □

Proof. The objective function G of the dual satisfies

$$G(\lambda,\lambda_0) \;:=\; \frac{1}{N}\sum_{i=1}^{N}\left[\, T_i \cdot f^*(\lambda_0 + \langle B(X_i),\lambda\rangle) \;-\; (\lambda_0 + \langle B(X_i),\lambda\rangle)\,\right] \;+\; \langle \delta, |\lambda|\rangle \,.$$

Without the last term, this is a differentiable convex function.

It is well know that a differentiable convex functions g satisfies

$$g(x) - g(y) \ge \nabla g(y)^\top (x-y) \qquad \text{for all } x, y \,. \tag{2.7}$$

The gradient of

$$g := (\lambda,\lambda_0) \mapsto \frac{1}{N}\sum_{i=1}^{N}\left[\, T_i \cdot f^*(\lambda_0 + \langle B(X_i),\lambda\rangle) \;-\; (\lambda_0 + \langle B(X_i),\lambda\rangle)\,\right] \tag{2.8}$$

is

$$\nabla g = (\lambda,\lambda_0) \mapsto \frac{1}{N}\sum_{i=1}^{N}\left[T_i \cdot (f')^{-1}(\lambda_0 + \langle B(X_i),\lambda\rangle) \;-\; 1\right][B(X_i)^\top, 1]^\top \tag{2.9}$$

Thus

$$\begin{aligned} &G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) \;-\; G(\lambda^*, \lambda_0^\dagger) \\ &\quad\ge -\frac{1}{N}\sum_{i=1}^{N}\left[B(X_i)^\top, 1\right] \cdot \begin{bmatrix} \Delta \\ \Delta_0 \end{bmatrix} \left(1 \;-\; T_i \cdot (f')^{-1}\left(\langle B(X_i),\lambda^*\rangle + \lambda_0^\dagger\right)\right) \\ &\qquad + \;\langle \delta, |\lambda^* + \Delta| - |\lambda^*|\rangle \,. \end{aligned} \tag{2.10}$$

We fix $\tilde\varepsilon > 0$ and establish the lower bound $-\tilde\varepsilon$ with probability going to 1 as $N \to \infty$. We control the **first term** by (what?) and the **second term** by $\|\delta\|$.

First Term

We note, that by $\|B(x)\| \leq 1$ and the Cauchy-Schwarz inequality it holds

$$\Big[B(X_i)^\top, 1\Big] \cdot \begin{bmatrix} \Delta \\ \Delta_0 \end{bmatrix} \lesssim \|(\Delta, \Delta_0)\| = \varepsilon \,. \tag{2.11}$$

Next, we see that

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N}\Big(1 \;-\; T_i \cdot (f')^{-1}\Big(\langle B(X_i), \lambda^*\rangle + \lambda_0^\dagger\Big)\Big) \\ &\lesssim \frac{1}{N}\sum_{i=1}^{N}\left|1 - \frac{T_i}{\pi_i}\right| \;+\; \frac{1}{N}\sum_{i=1}^{N}\left|\langle B(X_i), \lambda^*\rangle + \lambda_0^\dagger \;-\; f'\left(\frac{1}{\pi_i}\right)\right| \\ &=: \; S_N \;+\; M_N \,. \end{aligned} \tag{2.12}$$

With $\tilde{\varepsilon} > 0$ fixed previously, we want to establish the upper bound $\tilde{\varepsilon}/(2\varepsilon)$ with probability going to 1 as $N \to \infty$.

First, we bound S_N. By the properties of conditional expectation it holds

$$\mathbf{E}\left[\frac{T}{\pi(X)}\right] = \mathbf{E}\left[\frac{\mathbf{E}[T|X]}{\pi(X)}\right] = 1 \,.$$

By the weak law of large numbers (L1 version ? some assumption on 1/pi?)

$$\mathbf{P}\left[S_N \geq \tilde{\varepsilon}/(4\varepsilon)\right] \;\to\; 0 \qquad \text{for } N \to \infty \,. \tag{2.13}$$

Next, we bound M_N. Recall that $\sum_{k=1}^{N} B_k(x) = 1$. Thus

$$\langle B(X), \lambda^*\rangle \;+\; \lambda_0^\dagger \;=\; \sum_{k=1}^{N} B_k(X)\left(f'\left(\frac{1}{\pi_k}\right) - \lambda_0^\dagger\right) \;+\; \lambda_0^\dagger \;=\; \sum_{k=1}^{N} B_k(X) \cdot f'\left(\frac{1}{\pi_k}\right) \,.$$

By Markov's inequality it holds

$$\begin{aligned} &\mathbf{P}\left[M_N \geq \tilde{\varepsilon}/(4\varepsilon)\right] \\ &\leq \; \frac{4\varepsilon}{\tilde{\varepsilon}}\,\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\left[\left|\sum_{k=1}^{N} B_k(X_i) \cdot f'\left(\frac{1}{\pi_k}\right) - f'\left(\frac{1}{\pi_i}\right)\right|\right] \\ &\leq \; \frac{4\varepsilon}{\tilde{\varepsilon}}\,\mathbf{E}\left[\left|\sum_{k=1}^{N} B_k(X) \cdot f'\left(\frac{1}{\pi_k}\right) - f'\left(\frac{1}{\pi(X)}\right)\right|\right] \\ &\leq \; \frac{4\varepsilon}{\tilde{\varepsilon}}\,\mathbf{E}\left[\left|\sum_{k=1}^{N} B_k(X) \cdot f'\left(\frac{1}{\pi_k}\right) - f'\left(\frac{1}{\pi(X)}\right)\right|^2\right]^{1/2} \quad \to \; 0 \qquad \text{for } N \to \infty \,. \end{aligned}$$

The convergence is due to the universal consistency of B. This establishes the desired bound of $\tilde{\varepsilon}/(2\varepsilon)$ in (2.12). Together with (2.11) we conclude that the **first term** in (2.10) is bounded below by $-\tilde{\varepsilon}/2$ with probability going to 1 as $N \to \infty$.

Second Term

It holds

$$|x+y|-|x| \geq -|y| \qquad \text{for all } x, y\,.$$

Since $\delta \geq 0$ we get

$$\begin{aligned}&\langle \delta, |\lambda^* + \Delta| - |\lambda^*|\rangle \\ &\quad \geq\ -\langle \delta, |\Delta|\rangle\ \geq\ -\|\delta\|\,\|\Delta\|\ \geq\ -\|\delta\|\,\|(\Delta, \Delta_0)\|\ \geq\ -\|\delta\|\,\varepsilon\ \geq\ -\tilde{\varepsilon}/2\,,\end{aligned}$$

with probability going to 1 as $N \to \infty$. The convergence is due to $\|\delta\|$ converging to 0 in probability.

Conclusion

With the analysis of the **first** and **second term** in (2.10) we conclude

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0)\ -\ G(\lambda^*, \lambda_0^\dagger) \geq -\tilde{\varepsilon} \tag{2.14}$$

with probability going to 1 as $N \to \infty$. Since this holds true for all $\tilde{\varepsilon} > 0$ we get

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0)\ -\ G(\lambda^*, \lambda_0^\dagger) \geq 0 \tag{2.15}$$

with probability going to 1 as $N \to \infty$. But this holds for all (Δ, Δ_0) with $\|(\Delta, \Delta_0)\| = \varepsilon$. Thus

$$\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0 \tag{2.16}$$

with probability going to 1 as $N \to \infty$. Thus by Lemma 2.1

$$\mathbf{P}\left[\left\|\lambda^\dagger - \lambda^*\right\| \geq \varepsilon\right]\ \to\ 0 \qquad \text{for } N \to \infty\,. \tag{2.17}$$

Finally, note that this holds for all $\varepsilon > 0$. This finishes the proof. □

Consistency for Inverse Propensitiy Score

Theorem 2.2. *For all $\varepsilon > 0$ it holds*

$$\mathbf{P}\left[|w(X) - 1/\pi(X)| \geq \varepsilon\right]\ \to\ 0 \qquad \textit{for } N \to \infty\,.$$

Furthermore, it holds

$$\mathbf{E}\left[|w(X) - 1/\pi(X)|^2\right]^{1/2}\ \to\ 0 \qquad \textit{for } N \to \infty\,.$$

Proof. We employ the consistency of the dual variable, the universal consistency and boundedness of the regression basis and the constraint on the arithmetic mean of the weights.

$$\begin{aligned}
\left|w(X) - \frac{1}{\pi(X)}\right| &= \left|(f')^{-1}\left(\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger\right) - \frac{1}{\pi(X)}\right| \\
&\lesssim \left|\langle B(X), \lambda^\dagger - \lambda^*\rangle\right| + \left|\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger - f'\left(\frac{1}{\pi(X)}\right)\right| \\
&\lesssim \left\|\lambda^\dagger - \lambda^*\right\| + \left|\sum_{i=1}^{N} B_k(X) \cdot f'\left(\frac{1}{\pi_k}\right) - f'\left(\frac{1}{\pi(X)}\right)\right| \\
&\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \le \varepsilon\,,
\end{aligned} \tag{2.18}$$

with probability going to 1 as $N \to \infty$.

If we prove boundedness, L_2-convergence follows readily.

$$\begin{aligned}
\left|w(X) - \frac{1}{\pi(X)}\right|^2 &\le \left|w(X) - \frac{1}{C_\pi}\right|^2 \\
&\lesssim \left|\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger - f'\left(\frac{1}{C_\pi}\right)\right|^2 \\
&\lesssim \left(\operatorname{diam}\Theta + f'\left(\frac{1}{C_\pi}\right)\right)^2
\end{aligned}$$

□

Consistency of the Weighted Mean

Theorem 2.3. *For all $\varepsilon > 0$ it holds*

$$\mathbf{P}\left[\left|\frac{1}{N}\sum_{i=1}^{n} w_i Y_i \;-\; \mathbf{E}[Y(1)]\right| \;\ge\; \varepsilon\right] \;\to\; 0 \qquad \text{for } N \to \infty\,,$$

that is, the weighted mean is a consistent estimator. Furthermore, it holds for all $p \in [1, \infty)$

$$\mathbf{E}\left|\frac{1}{N}\sum_{i=1}^{n} w_i Y_i \;-\; \mathbf{E}[Y(1)]\right|^p \;\to\; 0 \qquad \text{for } N \to \infty\,.$$

Proof. Let $\mathbf{Y}(1)$ be the vector with i-th coordinate $Y_i(1)$, that is, the vector of marginal potential outcomes under treatment. Note, that $\mathbf{Y}(1)$ is usually unknown. Nevertheless, we can leverage its existence in the following error decomposition. Also note, that for $i > n$ we may choose $w_i = 1/\pi_i$.

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \;\le\; & \left|\frac{1}{N}\left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i)\right)^{\top} \mathbf{Y}(1)\right| \\ & + \left|\frac{1}{N}\sum_{i=1}^{N} (T_i \cdot w_i - 1)\left(\mathbf{E}[Y(1)|X_i] - \langle B(X_i), \mathbf{Y}(1)\rangle\right)\right| \\ & + \left|\frac{1}{N}\sum_{i=1}^{N} T_i(w_i - 1/\pi_i)\left(Y_i - \mathbf{E}[Y(1)|X_i]\right)\right| \\ & + \left|\frac{1}{N}\sum_{i=1}^{N} T_i/\pi_i\left(Y_i - \mathbf{E}[Y(1)|X_i]\right) + \left(\mathbf{E}[Y(1)|X_i] - \mathbf{E}[Y(1)]\right)\right| \\ =: \; & R_1 + R_2 + R_3 + R_4\end{aligned}$$

Analysis of R_1

By the Cauchy-Schwarz inequality it holds

$$\left|\frac{1}{N}\left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i)\right)^{\top} \mathbf{Y}(1)\right| \;\le\; \|\delta\|\,\|\mathbf{Y}(1)\| \;\lesssim\; \|\delta\|\, N \to 0 \qquad \text{for } N \to \infty\,.$$

Analysis of R_2

This calculation will be central to the asymptotic normality.

$$\begin{aligned}&\mathbf{P}[R_2 \ge \varepsilon] \\ &\;\le\; \varepsilon^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\left[\left|(T_i \cdot w_i - 1)\left(\mathbf{E}[Y(1)|X_i] - \langle B(X_i), \mathbf{Y}(1)\rangle\right)\right|\right] \\ &\;\le\; \varepsilon^{-1}\mathbf{E}\left[|w(X) - 1/\pi(X)|^2\right]^{1/2}\mathbf{E}\left[|\mathbf{E}[Y(1)|X_i] - \langle B(X_i), \mathbf{Y}(1)\rangle|^2\right]^{1/2} \;\to\; 0\end{aligned}$$

for $N \to \infty$. Notice that the rates multiply. This is important for later.

Analysis of R_3

By Theorem? it holds

$$\mathbf{P}[R_3 \ge \varepsilon] \le \frac{\mathbf{E}[|w(X) - 1/\pi(X)|^2]^{1/2}}{\varepsilon} \to 0 \qquad \text{for } N \to \infty\,.$$

Analysis of R_4

$$\begin{aligned}
\mathbf{E}\left[Y(T) \cdot (T \,/\, \pi(X))\right] &= \mathbf{E}\left[Y(1) \,/\, \pi(X) \,|\, T = 1\right] \cdot \mathbf{P}[T = 1] \\
&= \int_{\mathcal{X}} \mathbf{E}\left[Y(1) \,|\, X = x, T = 1\right] \cdot (\mathbf{P}[T = 1] \,/\, \pi(x))\, \mathbf{P}_{X|T}(dx \,|\, 1) \\
&= \int_{\mathcal{X}} [Y(1)|X = x]\, \mathbf{P}_X(dx) \\
&= \mathbf{E}[Y(1)]\,.
\end{aligned} \tag{2.19}$$

The first equality holds because of the definition of the weights. The second, third and last equality stem from $T \in \{0, 1\}$, and the law of total expectation, applied with T and X. The fourth equality is justified by the assumption of conditional unconfoundedness. The density transformation is due to Bayes's Theorem. Thus the expectation of the summands in R_4 is 0. Convergence follows from the WLLN.

Conclusion

We conclude that

$$\left| \frac{1}{N} \sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)] \right| \leq \varepsilon \tag{2.20}$$

with probability going to 1 as $N \to \infty$. To show the second statement we calculate

$$\left| \frac{1}{N} \sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)] \right|^p = \left| \frac{1}{N} \sum_{i=1}^{n} w_i \left(Y_i - \mathbf{E}[Y(1)]\right) \right|^p \tag{2.21}$$

$$\leq \frac{1}{N} \sum_{i=1}^{n} w_i \left|(Y_i - \mathbf{E}[Y(1)])\right|^p \leq (2M)^p \tag{2.22}$$

□

3 Asymptotic Normality and Convergence to Gaussian Bridge

Theorem 3.1. *Under conditions the partition estimate has*

$$\mathbf{E}\left\|m_N - m\right\|^2 \leq C_{\mathbf{P}} N^{-\frac{2}{d+2}} \tag{3.1}$$

Theorem 3.2. *Under conditions the kernel estimate has*

$$\mathbf{E}\left\|m_N - m\right\|^2 \leq C_{\mathbf{P}} N^{-\frac{2}{d+2}} \tag{3.2}$$

Learning Rates for the Dual

Theorem 3.3. *Under conditions*

$$\mathbf{P}\left[\left\|\lambda^\dagger - \lambda^*\right\| \leq C_{\mathbf{P}} C_\tau \varepsilon_n\right] \geq 1 - \tau\,, \tag{3.3}$$

where ε_n *is the square root of the basis function Learning rate and* C_τ *depends on the Concentration Inequality. We need bernstein confidence* $\sqrt{\log(1/\tau)}$ *to preserve minimal Learning rate for* $d = 1$ *.*

Learning Rates for the Primal

Theorem 3.4. *Under conditions the weights satisfy*

$$\mathbf{E}[|w^\dagger(X) - 1/\pi(X)|^2]^{1/2} \leq C_{\mathbf{P}} \sqrt{\log(n)} n^{-1/(2+d)} \tag{3.4}$$

where $varepsilon_n$ *depends on the Learning rate of the basis functions and the confidence of the dual.* $C_{\mathbf{P}}$ *depends on the size of the parameter space.*

Proof.

$$\begin{aligned}
\mathbf{E}[|w^\dagger(X) - 1/\pi(X)|^2]^{1/2} &= \mathbf{E}\left[\left|(f')^{-1}\left(\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger\right) - 1/\pi(X)\right|^2\right]^{1/2} && (3.5)\\
&\leq \left|(f')^{-1}\right|_L (I_1 + I_2) && (3.6)
\end{aligned}$$

where

$$I_1 := \left(\mathbf{E}\left\|\lambda^\dagger - \lambda^*\right\|^2\right)^{1/2} \tag{3.7}$$

$$I_2 := \mathbf{E}\left[\left|\sum_{k=1}^{N} B_k(X)\cdot f^{'}(X_k) - f^{'}(X)\right|^2\right]^{1/2} \tag{3.8}$$

It holds $I_2 \leq n^{-1/(d+2)}$ by the lr of the basis. To analyse I_1 we use the lr of the dual.

$$I_1 \leq C_\tau n^{-1/(d+2)} + \sqrt{\tau}\cdot \operatorname{diam}\Theta \tag{3.9}$$

Note that the Markov confidence $1/\sqrt{\tau}$ is insufficient. We need the Bernstein confidence $\sqrt{\log(1/\tau)}$. With Bernstein confidence, bounded diameter and $\tau = n^{-2/(d+2)}$ we get

$$I_1 \leq C\cdot\sqrt{\log(n)}n^{-1/(2+d)} \tag{3.10}$$

Thus

$$\mathbf{E}[|w^\dagger(X) - 1/\pi(X)|^2]^{1/2} \leq C\cdot\sqrt{\log(n)}n^{-1/(2+d)} \tag{3.11}$$

□

Asymptotic Normality of the Weighted Mean

Theorem 3.5. *The estimate*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{n} w_i\mathbf{1}_{\{Y_i\leq t\}} - \mathbf{P}[Y(1)\leq t]\right) \tag{3.12}$$

is Asymptoticaly normal.

Proof.

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{n} w_i\mathbf{1}_{\{Y_i\leq t\}} \ - \ \mathbf{P}[Y(1)\leq t]\\
&= \frac{1}{N}\left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i)\right)^\top \mathbf{Y}(1)\\
&\quad + \frac{1}{N}\sum_{i=1}^{N}(T_i\cdot w_i - 1)\left(\mathbf{P}[Y(1)\leq t|X_i] - \langle B(X_i),\mathbf{Y}(1)\rangle\right)\\
&\quad + \frac{1}{N}\sum_{i=1}^{N} T_i(w_i - 1/\pi_i)\left(\mathbf{1}_{\{Y_i\leq t\}} - \mathbf{P}[Y(1)\leq t|X_i]\right)\\
&\quad + \frac{1}{N}\sum_{i=1}^{N} T_i/\pi_i\left(\mathbf{1}_{\{Y_i\leq t\}} - \mathbf{E}[Y(1)\leq t|X_i]\right) + \left(\mathbf{P}[Y(1)\leq t|X_i] - \mathbf{P}[Y(1)\leq t]\right)\\
&=: \ R_1 + R_2 + R_3 + R_4
\end{aligned}$$

The term

$$\left| \frac{1}{N} \sum_{i=1}^{N} (T_i w_i - 1) \left(\mathbf{P}[Y(1) \leq t | X_i] - \langle B(X_i)), \lambda^* \rangle \right) \right| \tag{3.13}$$

gives back the expectation. We control its rate with the factor of basis rates and primal weights. Choose d to be faster than $\sqrt{n}$. The rest follows with standard empirical theory. Expectations are 0. $\|\delta\|$ has to converge fast enough.

$$\begin{aligned}
&\sqrt{N} \cdot \mathbf{E}\left[(T \cdot w(X) - 1) \left(\mathbf{P}[Y(1) \leq t | X] - \langle B(X), \lambda^* \rangle \right) \right] \\
&\quad = \sqrt{N} \cdot \mathbf{E}\left[\pi(X) \cdot (w(X) - 1/\pi(X)) \left(\mathbf{P}[Y(1) \leq t | X] - \langle B(X), \lambda^* \rangle \right) \right] \\
&\quad \leq \sqrt{N} \cdot \mathbf{E}\left[|w(X) - 1/\pi(X)|^2 \right]^{1/2} \mathbf{E}\left[|\mathbf{P}[Y(1) \leq t | X] - \langle B(X), \lambda^* \rangle|^2 \right]^{1/2} \\
&\quad \leq C\sqrt{N}\sqrt{\log(N)} N^{-1/(d+2)} \cdot N^{-1/(d+2)} \\
&\quad \leq C\sqrt{\log(N)} N^{-1/6} \to 0 \qquad \text{for } N \to \infty
\end{aligned}$$

Note, that we get no convergence for $d > 1$. Also note that

$$\mathbf{E}\left[T \cdot w(X) - 1 | X, X_1, \ldots, X_N \right] = \mathbf{E}\left[T | X \right] w(X) - 1 = \pi(X) \left(w(X) - 1/\pi(X) \right)$$

□

Lemma 3.1. *Under conditions it holds for all* $\varepsilon > 0$

$$\mathbf{P}\left[\|G_N\|^*_{\mathcal{F}_N} \geq \varepsilon \right] \;\to\; 0 \qquad \textit{for } N \to \infty\,. \tag{3.14}$$

Proof. By maximal inequalities it holds

$$\mathbf{E}^*\left[\|G_N\|_{\mathcal{F}_N} \right] \;\lesssim\; J_{[]}\left(\varepsilon_N, \mathcal{F}_N, \mathrm{L}_2(\mathbf{P}) \right) \tag{3.15}$$

$$= \int_0^{\varepsilon_N} \sqrt{\log N_{[]}\left(\varepsilon, \mathcal{F}_N, \mathrm{L}_2(\mathbf{P}) \right)}\, d\varepsilon = \int_0^{\varepsilon_N} \sqrt{\log N_{[]}\left(\varepsilon/\varepsilon_N, B_{\mathcal{F}}, \mathrm{L}_2(\mathbf{P}) \right)}\, d\varepsilon \tag{3.16}$$

$$\leq \int_0^{\varepsilon_N} \left(\frac{\varepsilon_N}{\varepsilon} \right)^{k/2} d\varepsilon \tag{3.17}$$

$$\lesssim \varepsilon_N^{k/2} \frac{1}{1 - k/2} \varepsilon_N^{1-k/2} \lesssim \varepsilon_N \to 0 \qquad \text{for } N \to \infty\,. \tag{3.18}$$

Note, that $k < 2$. By the boundedness of $\mathbf{E}^*$ there is no measurability problem. By Markov's Inequality it holds

$$\mathbf{P}\left[\|G_N\|^*_{\mathcal{F}_N} \geq \varepsilon \right] \leq \mathbf{E}^*\left[\|G_N\|_{\mathcal{F}_N} \right] \tag{3.19}$$

□

Gaussian Bridge

We can even view $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i$ as an empirical process $\mathbb{G}_n f$ indexed over

$$f_\Phi(T,X,Y) = \frac{T}{\pi(X)}\left(\Phi(Y) - \mathbf{E}[\Phi(Y)|X]\right) + \mathbf{E}[\Phi(Y)|X]. \tag{3.20}$$

If $\mathcal{F} = \{f_\Phi \colon \Phi \in \text{ some set}\}$ is $\mathbf{P}$-Donsker, the empirical process converges to a tight gaussian process. Then the functional delta Method is applicable.

3.1 Application to Plug In Estimators

A plethora of applications of the delta method to estimates of the distribution function are to be found in [vdV00] and [vdvW13]. This includes Quantile estimation [vdV00, §21] [vdvW13, §3.9.21/24], survival analysis via Nelson-Aalen and Kaplan-Meier estimator [vdvW13, §3.9.19/31], Wilcoxon Test [vdvW13, §3.9.4.1], and much more. Maybe Boostrapping from the weighted distribution is also sensible .

3.2 Double Robustness

By double robustness we mean the property of an estimator that it is consistent if either one of treatment or outcome model is well specified. The augmented weighting estimator was designed to have exactly this property. Surprisingly, the weighted mean estimator in the balancing weights approach of [WZ19] retains this feature despite its simplicity [ZP17]. In the following we explore double robustness in the weighted mean estimator from the perspective of learning rates. We will adopt a similar notion of learning rates as in [SC08]. To the best of our knowledge this approach is new. We investigate how learning rates change in different scenarios.

3.2.1 Learning Rates of the weighted mean

What is the speed of convergence in the weak law of large numbers? A clear-cut answer is: The arithmetic mean of independent, identically distributed, square-integrable random variables learns with rate $n^{-1/2}$. Furthermore, using Bienaymé's formula and Chebyshev's inequality, the statement is easy to prove (cf. [Kle20, Theorem 5.14]).

Theorem. *Let $X_1, X_2, \dots$ be i.i.d, square-integrable random variables with $V := \mathbf{Var}[X_1] < \infty$. Then, for any $\tau \in (0,1]$ and all $n \in \mathbb{N}$, we have*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}[X_i])\right| \le \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \ge 1-\tau\,. \tag{3.21}$$

Reflection. Bernsteins's inequality yields better confidence. ♠

Theorem 3.6. (Bernstein's inequality) *Let* $(\Omega, \mathcal{A}, \mathbf{P})$ *be a probability space,* $B > 0$ *and* $\sigma > 0$ *be real numbers, and* $n \geq 1$ *be an integer. Furthermore, let* $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ *be independent random variables satisfying* $\mathbf{E}[X_i] = 0$ *,* $\|X_i\|_\infty \leq B$ *and* $\mathbf{E}[X_i^2] \leq \sigma^2$ *for all* $i = 1, \ldots, n$ *. Then we have*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \;\leq\; \sqrt{\frac{2\,\sigma^2\log(e/\tau)}{n}} \;+\; \frac{2B\log(e/\tau)}{3n}\right] \;\geq\; 1-\tau\,, \qquad \tau > 0\,.$$

Proof. Confer [SC08, Theorem 6.12] for the one-sided version. The two-sided version, as stated in the above theorem, is an easy consequence. We omit the details. □

$\sqrt{2}(\sigma \vee B)\frac{1}{\sqrt{n}}\log(e/\tau)$

We conflate *observed outcomes* with *marginal potential outcomes*. But few study populations allow for such disregard. If, by chance, we compare the health of systematically different persons, such as that of a thriving smoker with that of an asthmatic non-smoker, the conclusion leads us astray. Smoking is not healthy. **Confounders**, as the literature calls unbalanced external influences on both treatment and outcome, are lurking in the data. The term *Simpson's paradox* is a collection of painful stories telling about surprising changes of effect in subclasses of the data (cf. [Wag82] for a brief discussion of Simpson's paradox and some real world examples). The probabilistic tools of the community were not sharp enough, so non-probabilistic frameworks, like causal framework, found widespread acceptance. Their proper use in empirical data analysis is, however, still subject to debate [Pea09, §6].

In experimental studies we usually specify treatment assignment as opposed to merely observing a unit receiving treatment.

The next statement makes use of the propensity score.

Theorem. *Consider the weighted mean estimator with weights*

$$w_i \;=\; \frac{1}{n}\frac{T_i}{\pi(X_i)}\,. \tag{3.22}$$

Denote $V := \mathbf{E}[(Y(1))^2\,/\,\pi(X)] - \mathbf{E}[Y(1)]^2$ *. Assume that weak unconfoundedness holds. Then, for any* $\tau \in (0,1]$ *and all* $n \in \mathbb{N}$ *, we have*

$$\mathbf{P}\left[\left|\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \;\leq\; \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \;\geq\; 1-\tau\,. \tag{3.23}$$

Proof. We want to reinforce coherent use of the weak law of large numbers. To this end, we verify

$$\begin{aligned} n\,\mathbf{E}[w(T,X)\,Y(T)] &= \mathbf{E}[Y(1)]\,, \\ n^2\,\mathbf{Var}[w(T,X)\,Y(T)] &= \mathbf{E}[(Y(1))^2\,/\,\pi(X)] - \mathbf{E}[Y(1)]^2\,. \end{aligned}$$

Essentially, the random weight $w(T,X)$ acts on $Y(T)$ through $T\,/\,\pi(X)$. It does so by inducing independence of observed outcome $Y(T)$ and treatment T. This requires that weak unconfoundedness holds, i.e.,

$$(Y(0),Y(1)) \perp\!\!\!\perp T\,|\,X\,. \tag{3.24}$$

To showcase the details we added an n and n^2 factor in the above display. The calculations go as follows.

$$\begin{aligned} n\,\mathbf{E}[w(T,X)\,Y(T)] &= \mathbf{E}\,[Y(T)\cdot(T\,/\,\pi(X))] \\ &= \mathbf{E}\,[Y(1)\,/\,\pi(X)\,|\,T=1]\cdot\mathbf{P}[T=1] \\ &= \int_{\mathcal{X}} \mathbf{E}\,[Y(1)\,|\,X=x,T=1]\cdot(\mathbf{P}[T=1]\,/\,\pi(x))\,\mathbf{P}_{X|T}(dx\,|\,1) \\ &= \int_{\mathcal{X}} [Y(1)|X=x]\,\mathbf{P}_X(dx) = \mathbf{E}[Y(1)]. \end{aligned} \tag{3.25}$$

The first equality holds because of the definition of the weights. The second, third and last equality stem from $T \in \{0,1\}$, and the law of total expectation, applied with T and X. The fourth equality is justified by the assumption of weak unconfoundedness. The density transformation is due to Bayes's Theorem. With slight modifications in the above argument, it follows

$$\mathbf{E}\left[\Big(Y(T)\cdot(T\,/\,\pi(X))\Big)^2\right] = \mathbf{E}\Big[(Y(1))^2\,/\,\pi(X)\Big]\,. \tag{3.26}$$

We omit the details. Invoking the weak law of large numbers finishes the proof. □

Formulate improved version with Bernstein.

Formulate better transition to weights with estimated propensity score.

We started by asking an easy question, so it is time for a more challenging one: How do we proceed in deriving learning rates if the propensity score is unknown. How do we generally procede? [what has been done in the past. Why are some methods obsolete] A naive answer would be: We hope to select a proper model and try to estimate the propensity score. Stunningly, a lot of practicioners still settle for obsolete methods when it comes to propensity score analysis.

Next, we consider the event that we have a consistent estimator of the propensity score and the distribution of the covariate vector X, along with that of the outcome Y, has compact support. The next assumptions reduce the technical task to a minimum.

Assumptions. *Let the following hold.*

(i) *There exists* $C_Y \geq 1$ *such that* $|Y(1)| \leq C_Y$ *almost surely.*

(ii) *There exists* $C_\pi > 0$ *such that* $C_\pi < \pi(X)$ *almost surely.*

(iii) *There exists a function class* $\mathcal{F}$ *with unit ball* $B_\mathcal{F} := \{f \in \mathcal{F} \colon \|f\|_\infty \leq 1\}$ *such that* $\log N_{[]}(\varepsilon, B_\mathcal{F}, L_2(\mathbf{P})) \leq C_\mathcal{F}(1/\varepsilon)^{1/k}$ *for some* $k > 1/2$ *and some constant* $C_\mathcal{F} \geq 1$.

(iv) *The random function* f_w *defined by* $f_w(T, X, Y) := \left(n\, w(X) - \frac{1}{\pi(X)}\right) T\, Y$ *satisfies* $f_w \in \mathcal{F}$ *almost surely .*

(v) *There exist a learning rate* (r_n), *confidence constants* (γ_τ) *and uniform constant* $C_w \geq 1$ *such that for all* $\tau \in (0,1]$ *and all* $n \in \mathbb{N}$ *it holds* $\mathbf{P}\left[\left\|w(X) - \frac{1}{\pi(X)}\right\|_\infty \leq C_w c_\tau \varepsilon_n\right] \geq 1 - \tau$

(vi) *There exists* $\alpha > 1$ *such that* $\varepsilon_n \cdot c_{n^{-\alpha}} \to 0$ *as* $n \to \infty$ *and* $\varepsilon_n \cdot c_{n^{-\alpha}} \leq 1$ *for all* $n \in \mathbb{N}$.

Elaborate on assumptions. For example (iii) [vdvW13, §2.7.1] [vdV00, §19.9]

Theorem. *Let the assumptions. Then the weighted mean learns with rate* (ε_n) *defined by*

$$\varepsilon_n \;:=\; \inf\big\{\, t \,\in\, (0,1] \,:\, r_n \cdot \gamma_{t/n} \;\leq\; t \,\big\} \;\wedge\; 1 \qquad \textit{for all } n \in \mathbb{N}\,.$$

Furthermore, it has confidence constants (c_τ) *given by*

$$c_\tau = \gamma_\tau / \tau \tag{3.27}$$

and uniform constant

$$C_\mathbf{P} = \max\left\{ \sqrt{C_\mathcal{F}} C_w, C_Y C_w, \frac{C_Y}{C_\pi} \right\} \tag{3.28}$$

Proof. We consider the following error decomposition.

$$\begin{aligned}
&\sum_{i=1}^{n} w_i T_i Y_i \; - \; \mathbf{E}[Y(1)] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)} \left(Y_i - \mathbf{E}[Y(1)]\,\right) \; + \; \sum_{i=1}^{n} T_i \left(w_i - \frac{1}{n\,\pi(X_i)} \right) Y_i \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)} \left(Y_i - \mathbf{E}[Y(1)]\,\right) \\
&\qquad\qquad\qquad + \; \frac{1}{\sqrt{n}}\,\mathbb{G}_n f_w \\
&\qquad\qquad\qquad\qquad - \; \mathbf{E}[f_w(T,X,Y)\,]\,.
\end{aligned}$$

We already bounded the first term. To bound the remaining terms we will use the learning rates of w. To this end, we employ maximal inequalities for empirical processes to bound the second term. We bound the third term by the law of total expectation and balancing learning rates and confidence.

2nd term

Denote $\mathcal{F}_{n,\tau} := (C_Y C_w \gamma_\tau r_n)\cdot B_{\mathcal{F}} =: \delta_{n,\tau}\cdot B_{\mathcal{F}}$. It holds by maximal inequalities

$$\begin{aligned}
\mathbf{E}^*\left[\|\mathbb{G}_n\|_{\mathcal{F}_{n,\tau}}\right] \; &\leq \; \int_0^{\delta n,\tau} \sqrt{\log N_{[\,]}\left(\varepsilon/\delta_{n,\tau}, B_{\mathcal{F}}, L_2(\mathbf{P})\right)}\, d\varepsilon \\
&\leq \; \int_0^{\delta_{n,\tau}} \left(\frac{\delta_{n,\tau}}{\varepsilon}\right)^{1/(2k)} d\varepsilon \; = \; \delta_{n,\tau}\,.
\end{aligned}$$

For $t > 0$, Markov's inequality gives

$$\mathbf{P}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}} \geq t\right] \; \leq \; \frac{1}{t}\,\mathbf{E}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}}\right] \; \leq \; \frac{1}{t}\,\mathbf{E}^*\left[\|\mathbb{G}_n\|_{\mathcal{F}_{n,\tau}}\right] \leq \frac{\delta_{n,\tau}}{t}\,, \tag{3.29}$$

and consequently

$$\mathbf{P}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}} \leq \frac{1}{\tau} C_Y C_w \gamma_\tau r_n\right] \; \geq \; 1-\tau \tag{3.30}$$

Next, note that

$$\|f_w\|_\infty \leq C_Y \left\| nw - \frac{1}{\pi(X)} \right\| \leq C_Y C_w \gamma_\tau r_n \tag{3.31}$$

with probability greater than $1-\tau$. Thus $f_w \in \mathcal{F}_{n,\tau}$ with probability greater than $1-\tau$. It follows

$$\mathbf{P}\left[\mathbb{G}_n f_w \leq \frac{1}{\tau} C_Y C_w \gamma_\tau r_n\right] \geq 1 - 2\tau\,. \tag{3.32}$$

Streamline analysis of second term.

3rd term

We localize with regards to $f_w \in \mathcal{F}_{n,\tau}$. We require the weights to be smalle that 1, such that we always have $\left\|nw - \frac{1}{\pi}\right\| \leq n + \frac{1}{C_\pi}$.

$$\begin{aligned} \mathbf{E}[f_w] &\leq C_Y C_w \gamma_\tau r_n (1-\tau) + C_Y(n + \frac{1}{C_\pi})\tau && (3.33)\\ &\leq C_Y \left(C_w \gamma_\tau r_n + \left(n + \frac{1}{C_\pi}\right)\tau \right) . && (3.34) \end{aligned}$$

If we choose $\tau = \varepsilon_n$ we get

$$\mathbf{E}[f_w] \leq C_Y \left(C_w + 1 + \frac{1}{C_\pi} \right) \varepsilon_n \, . \tag{3.35}$$

Selecting the worst instance of learning rate, confidence and uniform constant concludes the proof.

Streamline analysis of third term.

□

Reflection. Why did we use empirical process theory instead of conventional concentration inequalities? The weights w are random, so f_w is random as well.

Best is $\gamma_\tau = 1$, when we recover the learning rate of the estimator, that is, (r_n). The confidence γ_τ/τ is substandard. Improvements may involve Bernstein like concentration for empirical processes (cf. [vdvW13, Section 2.14.2]) ♠

Continue section with rates for right outcome model and then both models right. Do the rates improve?

Next we need to assume that the weights and outcome are independent given treatment and covariates, that is,

$$Y(1) \perp\!\!\!\perp w \mid X, T \, . \tag{3.36}$$

Also the bias of the outcome regression has to be bounded by the weights. We get the following error decomposition

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} n w_i T_i (Y_i - \mathbf{E}[Y(1)|X_i]) + (\mathbf{E}[Y(1)|X_i] - \mathbf{E}[Y(1)]) && (3.37)\\ + \sum_{i=1}^{n} (T_i w_i - 1/n)(\mathbf{E}[Y(1)|X_i] - B(X_i)\lambda) && (3.38)\\ + \sum_{i=1}^{n} (w_i - 1/n) B(X_i)\lambda && (3.39) \end{aligned}$$

The expectation of summand in the first term is zero. Hence the convergence by wlln or more refined methods. The second term goes to 0 by the consistency of the outcome regression. The third term is bounded by the weights.

$$\mathbf{E}[Tw \cdot (Y(T) - \mathbf{E}[Y(1) \,|\, X\,])] \;=\; \mathbf{E}[T \cdot \mathbf{E}[w \,|\, T, X] \cdot \mathbf{E}[Y(T) - Y(1) \,|\, T, X]\,] \;=\; 0$$

The first equality stems from the conditional independence assumptions. The resulting term vanishes because the difference $Y(T) - Y(1)$ does so after conditioning on $T = 1$.

$$\left|\sum_{i=1}^{n}(T_i w_i - 1/n)(\mathbf{E}[Y(1)|X_i] - B(X_i)\lambda)\right| \leq 2 \max_{i=1,\dots,n} |\mathbf{E}[Y(1)|X_i] - B(X_i)\lambda| \tag{3.40}$$

$$\leq 2\,\text{lerning rate triple of regression} \tag{3.41}$$

with probability greater than $1 - \tau$. Note, that the weights sum to 1. Assume $|\sum_{i=1}^{n}(T_i w_i - 1/n)B_k(X_i)| \leq \delta_k$ for all $k = 1, \dots, K$ and the deltas go to zero with some rate. The by Cauchy-Schwarz it follows

$$\left|\sum_{i=1}^{n}(T_i w_i - 1/n)B(X_i) \cdot \lambda\right| \leq |\langle(\delta_k), \lambda\rangle| \leq \text{rate of deltas} \cdot \|\lambda\| \tag{3.42}$$

So we assume $\lambda \in \Theta$, where the parameter space Θ is compact or grows moderately with K.

Introduce concept of semiparametric efficiency. For the semiparametric efficiency bound of propensity score weighting, [Hah98] is a good reference. General introduction to semiparametric models see [vdV00, §25].

Takeaways Each error decomposition furnishes information about the asymptotic properties of the weighted mean. We always get consistency, but learning rates may differ. Obtaining good confidence depends on employing more refined concentration inequalities like Bernstein's inequality. If both treatment and outcome model are well specified we reach the semiparametric efficiency bound of weighting with the true inverse propensity score.

4 Convex Analysis

In our application we want to analyse a convex optimization problem by its dual problem. In particular we want to obtain primal optimal solutions from dual solutions. To accomplish the task we need technical tools from convex analysis, mainly conjugate calculus and some KKT related results.

Our starting point is the support function intersection rule [MMN22, Theorem 4.23]. We give the details in the case of finite dimensions and refer for the rest of the proof to the book. The support function intersection rule is applied to give first conjugate sum and then chain rule, which are vital to calculating convex conjugates. The proofs are omited, since the book is thorough enough. The material we present is very well known. As an introduction, we recommend the recent book [MMN22] and classical reference [Roc70]. We finish the chapter with ideas from [TB91]. They provide the high-level ideas to obtain for strictly convex functions a dual relationship between optimal solutions. We will deliver the details that are omited in the paper.

4.1 A Convex Analysis Primer

Excursively, we present some well known definitions and facts from convex analysis. For details, see, e.g., [MMN22].

A subset $C \subseteq \mathbb{R}^n$ is called **convex set**, if for all $x, y \in C$ and all $\lambda \in [0, 1]$, we have $\lambda x + (1 - \lambda)y \in C$. The Cartesian product of convex sets is convex. The intersection of a collection of convex sets is also convex.

A set $A \subseteq \mathbb{R}^n$ is called **affine set**, if $\alpha x + (1 - \alpha)y \in A$ for all $x, y \in A$ and $\alpha \in \mathbb{R}$. The **affine hull** $\operatorname{aff}(\Omega)$ of a set $\Omega \subseteq \mathbb{R}^n$ is the smallest affine set that includes Ω. A mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ is called **affine mapping** if there exist a linear mapping $L : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $b \in \mathbb{R}^m$ such that $A(x) = L(x) + b$ for all $x \in \mathbb{R}^n$. The image and inverse image/preimage of convex sets under affine mappings are also convex.

Because the notion of interior is not precise enough for our purposes we define the relative interior which is the interior relative to the affine hull.

Definition. *Let $\Omega \subseteq \mathbb{R}^n$. We define the **relative interior** of Ω by*

$$\operatorname{ri}(\Omega) := \{x \in \Omega \colon \textit{there exists } \varepsilon > 0 \textit{ such that } (x + \varepsilon B) \cap \operatorname{aff}(\Omega) \subset \Omega\}\,. \tag{4.1}$$

Next we collect some useful properties of relative interiors.

Proposition 4.1. *Let C be a non-empty convex set in $\mathbb{R}^n$. The following holds:*

(i) $\operatorname{ri}(C) \neq \emptyset$ *if and only if* $C \neq \emptyset$

(ii) $\operatorname{cl}(\operatorname{ri} C) = \operatorname{cl} C$ *and* $\operatorname{ri}(\operatorname{cl} C) = \operatorname{ri}(C)$

(iii) $\operatorname{ri}(C) = \{z \in C \colon \textit{for all } x \in C \textit{ there exists } t > 0 \textit{ such that } z + t(z - x) \in C\}$

(iv) Suppose $\bigcap_{i \in I} C_i \neq \emptyset$ *for a finite index set* I *. Then* $\operatorname{ri}\left(\bigcap_{i\in I} C_i\right) = \bigcap_{i \in I} \operatorname{ri}(C_i)$ *.*

(v) Let $L : \mathbb{R}^n \to \mathbb{R}^m$ *be a linear function. Then* $\operatorname{ri} L(C) = L(\operatorname{ri} C)$ *. If it also holds* $L^{-1}(\operatorname{ri} C) \neq \emptyset$ *, we have* $\operatorname{ri} L^{-1}(C) = L^{-1}(\operatorname{ri} C)$ *.*

(vi) $\operatorname{ri}(C_1 \times C_2) = \operatorname{ri} C_1 \times \operatorname{ri} C_2$

Proof. For a proof of (i)-(v) we refer to [Roc70, Theorem 6.2 - 6.7].

To prove (vi) we use (iii). Let $(z_1, z_2) \in \operatorname{ri}(C_1 \times C_2)$. Then for all $(x_1, x_2) \in C_1 \times C_2$ there exists $t > 0$ such that

$$z_i + t(z_i - x_i) \in C_i \qquad \text{for all } i \in \{1, 2\}\,. \tag{4.2}$$

Using (iii) again, we get $\operatorname{ri}(C_1 \times C_2) \subseteq \operatorname{ri} C_1 \times \operatorname{ri} C_2$. Suppose $(z_1, z_2) \in \operatorname{ri} C_1 \times \operatorname{ri} C_2$. By (iii), for all $(x_1, x_2) \in C_1 \times C_2$ there exist $(t_1, t_2) > 0$ such that

$$z_i + t_i(z_i - x_i) \in C_i \qquad \text{for all } i \in \{1, 2\}\,. \tag{4.3}$$

If $t_1 = t_2$ we recover (4.2) from (4.3). By (iii) it holds $(z_1, z_2) \in \operatorname{ri}(C_1 \times C_2)$. If $t_1 < t_2$ we define $\theta := \frac{t_1}{t_2} \in (0, 1)$. Consider (4.3) with $i = 2$, together with $z_2 \in C_2$ and the convexity of C_2 . It follows

$$z_2 + t_1(z_2 - x_2) = \theta \cdot (z_2 + t_2(z_2 - x_2)) + (1 - \theta) \cdot z_2 \in C_2\,. \tag{4.4}$$

Now we consider (4.4) and (4.3) with $i = 1$. This gives (4.2) with $t = t_1$. As before, it follows $(z_1, z_2) \in \operatorname{ri}(C_1 \times C_2)$. If $t_1 > t_2$ similar arguments lead to the same result. We have proven $\operatorname{ri}(C_1 \times C_2) \supseteq \operatorname{ri} C_1 \times \operatorname{ri} C_2$ and equality. □

We procede with convex separation results which are vital to the subsequent developments.

Definition. *Let C_1 and C_2 be two non-empty convex sets in $\mathbb{R}^n$. A hyperplane H is said to* ***separate*** *C_1 and C_2 if C_1 is contained in one of the closed half-spaces associated with H and C_2 lies in the opposite closed half-space. It is said to separate C_1 and C_2* ***properly*** *if C_1 and C_2 are not both actually contained in H itselef.*

Theorem 4.1. (Convex separation in finite dimension) *Let C_1 and C_2 be two non-empty convex sets in $\mathbb{R}^n$. Then C_1 and C_2 can be properly separated if and only if* $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$.

Proof. [Roc70, Theorem 11.3] □

Definition. *Given a nonempty subset $\Omega \subseteq \mathbb{R}^n$, we define the* ***support function*** *of Ω to be*

$$\sigma_\Omega \,:\, \mathbb{R}^n \to \overline{\mathbb{R}}, \qquad x^* \mapsto \sup_{x\in\Omega} \langle x^*, x\rangle \,.$$

Definition 4.1. *Given functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ for $i = 1, \ldots, m$, we define the* ***infimal convolution*** *of these functions to be*

$$f_1 \Box \cdots \Box f_m \,:\, \mathbb{R}^n \to \overline{\mathbb{R}}, \quad x \mapsto \inf \left\{ \sum_{i=1}^{m} f_i(x_i) \,:\, x_i \in \mathbb{R}^n \text{ and } \sum_{i=1}^{m} x_i = x \right\} .$$

The next result establishes a connection between the support function of the intersection of two convex sets and the infimal convolution of the support functions of the sets taken by themselfes. The proof translates the geometric concept of convex separation to the world of convex functions.

Lemma 4.1. *Let C_1 and C_2 be two non-empty convex sets in $\mathbb{R}^n$. For any $x^* \in$* $\operatorname{dom} \sigma_{C_1 \cap C_2}$ *the sets*

$$\begin{aligned} \Theta_1 &:= C_1 \times [0, \infty)\,, \\ \Theta_2(x^*) &:= \{(x, \lambda) \in \mathbb{R}^n \,:\, x \in C_2 \textit{ and } \lambda \leq \langle x^*, x\rangle - \sigma_{C_1\cap C_2}(x^*)\} \end{aligned}$$

can by properly separated.

Proof. We fix $x^* \in \operatorname{dom} \sigma_{C_1\cap C_2}$ and write $\alpha := \sigma_{C_1\cap C_2}(x^*)$. In order to apply convex separation in finite dimension (Theorem 4.1) to the sets Θ_1 and $\Theta_2(x^*)$, it suffics to show their convexity and $\operatorname{ri}\Theta_1 \cap \operatorname{ri}\Theta_2(x^*) = \emptyset$.

Convexity of Θ_1 and $\Theta_2(x^*)$

Clearly, Θ_1 is convex by the convexity of C_1 and $[0,\infty)$. To see that $\Theta_2(x^*)$ is convex consider the linear function

$$L \,:\, \mathbb{R}^n \times \mathbb{R} \;\to\; \mathbb{R}\,, \qquad (x,\lambda) \;\mapsto\; \langle x^*, x\rangle - \lambda\,.$$

From the definitions of L and $\Theta_2(x^*)$ we get

$$\Theta_2(x^*) \;=\; (C_2 \times \mathbb{R}) \;\cap\; L^{-1}[\,\alpha,\infty)\,.$$

Thus, by Proposition 4.1 (v) and the convexity of C_2 we get the convexity of $L^{-1}[\,\alpha,\infty)$ and with it that of $\Theta_2(x^*)$.

Relative interiors of Θ_1 and $\Theta_2(x^*)$ are disjoint

We start by calculating the relative interiors. It holds

$$\begin{aligned} \operatorname{ri}\Theta_1 &\;=\; \operatorname{ri}(C_1 \times [0,\infty)) \;=\; \operatorname{ri} C_1 \times \operatorname{ri}[0,\infty) \;=\; \operatorname{ri} C_1 \times (0,\infty)\,, \\ \operatorname{ri}\Theta_2(x^*) &\;=\; \operatorname{ri}(L^{-1}[\,\alpha,\infty)) \;=\; L^{-1}(\operatorname{ri}[\,\alpha,\infty)) \;=\; L^{-1}(\alpha,\infty)\,. \end{aligned}$$

Suppose there exists $(\lambda, x) \in \operatorname{ri}\Theta_1 \cap \operatorname{ri}\Theta_2(x^*)$. Then it holds $x \in C_1 \times C_2$ and $\lambda > 0$. We also note, that

$$\alpha \;=\; \sigma_{C_1 \cap C_2}(x^*) \;=\; \sup_{z \in C_1 \cap C_2} \langle x^*, z\rangle \;\geq\; \langle x^*, x\rangle\,.$$

Then it follows

$$\alpha \;<\; \langle x^*, x\rangle - \lambda \;\leq\; \alpha\,,$$

a contradiction. Thus, the relative interiors of Θ_1 and $\Theta_2(x^*)$ are disjoint.

Applying Theorem 4.1 finishes the proof. □

Theorem. *Let C_1 and C_2 be two non-empty convex sets in $\mathbb{R}^n$ with $\operatorname{ri} C_1 \cap \operatorname{ri} C_2 \neq \emptyset$. Then the support function of the intersection $C_1 \cap C_2$ is represented as*

$$(\sigma_{C_1 \cap C_2})(x^*) = (\sigma_{C_1} \square\, \sigma_{C_2})(x^*) \qquad \textit{for all } x^* \in \mathbb{R}^n. \tag{4.5}$$

Furthermore, for any $x^* \in \operatorname{dom}(\sigma_{C_1 \cap C_2})$ there exist dual elements $x_1^*, x_2^* \in \mathbb{R}^n$ such that $x^* = x_1^* + x_2^*$. and

$$(\sigma_{C_1 \cap C_2})(x^*) = \sigma_{C_1}(x_1^*) + \sigma_{C_2}(x_2^*). \tag{4.6}$$

Proof. Using Lemma 4.1 the rest of the proof is as that of *[MMN22, Theorem 4.23(b)]*. □

Takeaways The support function intersection rule connects the geometric property of convex separation to an identity of support functions This result is central to the analysis of convex conjugates.

4.2 Conjugate Calculus

The goal of this section is to establish the tools to calculate convex conjugates. We cite the conjugate sum and chain rule without proof. After some examples, we cite the Fenchel-Rockafellar Theorem.

Definition 4.2. (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*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \tag{4.7}$$

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

Proposition 4.2. *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.*

Proof. [MMN22, Proposition 4.2] □

Theorem 4.2. *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***

$$(f+g)^*(x^*) = (f^* \square g^*)(x^*) \tag{4.8}$$

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*

$$(f+g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{4.9}$$

Proof. [MMN22, Theorem 4.27(c)] □

Theorem 4.3. *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 the* ***conjugate chain rule***

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{4.10}$$

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

Proof. [MMN22, Theorem 4.28(c)] □

Example 4.1. Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be a proper convex function, that is, $\operatorname{dom} f \neq \emptyset$ and f is convex. In steps we apply the conjugate chain and sum rule, together with mathematical induction, to prove the conjugate relationship

$$\begin{aligned} S_{f,n} : \mathbb{R}^n \to \overline{\mathbb{R}}, \qquad & (x_1, \dots, x_n) \mapsto \sum_{i=1}^{n} f(x_i), \\ S^*_{f,n} : \mathbb{R}^n \to \overline{\mathbb{R}}, \qquad & (x^*_1, \dots, x^*_n) \mapsto \sum_{i=1}^{n} f^*(x^*_i) . \end{aligned}$$

This relationship is very natural and the ensuing calculations serve to confirm our intuition.

First, we work in the projections on the coordinates. For the i-th coordinate, where $i = 1, \dots, n$, this is

$$p_i : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto x_i . \tag{4.11}$$

All projections p_i are linear function with matrix representation $e_i^\top$, where e_i is i-the coordinate vector. The adjoint of p_i is therefore

$$p^*_i : \mathbb{R} \to \mathbb{R}^n, \quad x \mapsto e_i \cdot x . \tag{4.12}$$

For the inverse image of the adjoint of p_i it holds

$$(p^*_i)^{-1} \{(x^*_1, \dots, x^*_n)\} \;=\; \begin{cases} \{x^*_i\} , & \text{if } x^*_j = 0 \text{ for all } j \neq i , \\ \emptyset & \text{else.} \end{cases} \tag{4.13}$$

Throughout this example we use the asterisk character * somewhat inconsistently. Note that f^* is the convex conjugate of the function f and p^*_i is the adjoint linear function of the projection on the i-th coordinate. Likewise, we denote dual variables, that is, the arguments of convex conjugates, as x^* .

Next, we employ the conjugate chain rule to establish the conjugate relationship

$$\begin{aligned} f_i : \mathbb{R}^n \to \overline{\mathbb{R}}, \quad & (x_1, \dots, x_n) \mapsto x_i \mapsto f(x_i) , \\ f^*_i : \mathbb{R}^n \to \overline{\mathbb{R}}, \quad & (x^*_1, \dots, x^*_n) \mapsto \begin{cases} f^*(x^*_i), & \text{if } x^*_j = 0 \text{ for all } j \neq i , \\ \infty & \text{else.} \end{cases} \end{aligned}$$

Note, that $f_i = (f \circ p_i)$ and $f_i^* = (f \circ p_i)^*$. Since $\operatorname{Im} p_i = \mathbb{R}$ and $\operatorname{dom} f \neq \emptyset$, it holds $\operatorname{Im} p_i \cap \operatorname{ri}(\operatorname{dom} f) \neq \emptyset$. Then f and p_i conform with the demands of the conjugate chain rule. It follows

$$\begin{aligned} f_i^*(x_1^*, \dots, x_n^*) &= (f \circ p_i)^*(x_1^*, \dots, x_n^*) = \inf\left\{f^*(y) \mid y \in (p_i^*)^{-1}\left\{(x_1^*, \dots, x_n^*)\right\}\right\} \\ &= \begin{cases} f^*(x_i^*), & \text{if } x_j^* = 0 \text{ for all } j \neq i\,, \\ \infty & \text{else,} \end{cases} \end{aligned}$$

where we keep to the convention $\inf \emptyset = \infty$. In the same way it follows

$$\left(S_{f,n} \circ p_{\{1,\dots,n\}}\right)^* (x_1^*, \dots, x_{n+1}^*) = \begin{cases} S_{f,n}^*(x_1^*, \dots, x_n^*) & \text{if } x_{n+1}^* = 0\,, \\ \infty & \text{else,} \end{cases} \tag{4.14}$$

Next, note that for $n = 1$ we arrive at the result. Thus, for some $n \in \mathbb{N}$ it holds $(S_{f,n})^* = S_{f,n}^*$. In order to apply the conjugate sum rule to $S_{f,n}$ and f_{n+1} we note that

$$\begin{aligned} \operatorname{dom} f_i &= \left\{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} : x_i \in \operatorname{dom} f\right\} \neq \emptyset \qquad \text{for all } i = 1, \dots, n+1\,, \\ \bigcap_{i=1}^{n+1} \operatorname{dom} f_i &= \left\{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} : x_i \in \operatorname{dom} f \text{ for all } i = 1, \dots, n+1\right\} \neq \emptyset\,, \end{aligned}$$

and

$$\begin{aligned} &\operatorname{ri}\left(\operatorname{dom}\left(S_{f,n} \circ p_{\{1,\dots,n\}}\right)\right) \cap \operatorname{ri}\left(\operatorname{dom} f_{n+1}\right) \\ &\qquad = \operatorname{ri}\left(\operatorname{dom}\left(S_{f,n} \circ p_{\{1,\dots,n\}}\right) \cap \operatorname{dom} f_{n+1}\right) = \operatorname{ri}\left(\bigcap_{i=1}^{n+1} \operatorname{dom} f_i\right) \neq \emptyset\,. \end{aligned}$$

By the conjugate sum rule it follows

$$\begin{aligned} (S_{f,n+1})^* = (S_{f,n} \circ p_{\{1,\dots,n\}} + f_{n+1})^* &= (S_{f,n} \circ p_{\{1,\dots,n\}})^* \square f_{n+1}^* \\ &= S_{f,n}^* \circ p_{\{1,\dots,n\}} + f_{n+1}^* = S_{f,n+1}^*\,. \end{aligned}$$

◊

Takeaways Conjugate sum and chain rule are direct consequences of the support function intersection rule. They are powerful tools, that allow us to compute convex conjugates of difficult expressions as well as proving the Fenchel-Rockafellar Duality theorem.

4.3 Duality of Optimal Solutions

We consider a general convex optimization problem with matrix equality and inequality constraints. For this problem there exists a related problem, which we call its dual. With ideas from [TB91] we establish a functional relationship between the optimal solution of the original problem and optimal solutions of the dual. The main assumption is that in the original problem we have a strictly convex objective function with continuously differentiable convex conjugate(cf. Definition 4.2).

Theorem 4.4. *Consider the optimization problem*

$$\begin{aligned} &\underset{w\in\mathbb{R}^n}{\text{minimize}} && f(w) && (4.15)\\ &\text{subject to} && \mathbf{U}w \;\geq\; d\,. \\ & && \mathbf{A}w \;=\; a\,, \end{aligned}$$

and its dual problem

$$\begin{aligned} &\underset{\lambda_d\in\mathbb{R}^r,\lambda_a\in\mathbb{R}^s}{\text{maximize}} && \langle\lambda_d, d\rangle \;+\; \langle\lambda_a, a\rangle \;-\; f^*\Big(\mathbf{U}^\top\lambda_d + \mathbf{A}^\top\lambda_a\Big) && (4.16)\\ &\text{subject to} && \lambda_d \;\geq\; 0\,. \end{aligned}$$

Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ *be an optimal solution to* (4.16)*. If the objective function* f *of* (4.15) *is strictly convex and its convex conjugate* f^* *is continuously differentiable, then the unique optimal solution to* (4.15) *is given by*

$$w^\dagger = \nabla f^*\Big(\mathbf{U}^\top\lambda_d^\dagger + \mathbf{A}^\top\lambda_a^\dagger\Big)\;. \tag{4.17}$$

Plan of Proof

We show that $w^\dagger$ and $(\lambda_d^\dagger, \lambda_a^\dagger)$ meet the Karush-Kuhn-Tucker conditions for 4.15, that is, **complementary slackness**

$$\langle\lambda_d^\dagger, d - \mathbf{U}w^\dagger\rangle \;=\; 0\,, \tag{4.18}$$

primal and **dual feasibility**

$$\begin{aligned} \mathbf{U}w^\dagger &\;\geq\; d\,, && (4.19)\\ \mathbf{A}w^\dagger &\;=\; a\,, \\ \lambda_d^\dagger &\;\geq\; 0\,, && (4.20) \end{aligned}$$

and **stationarity**

$$0_n \ \in \ [\partial f(w^\dagger) \ + \ \partial\,(w \mapsto d - \mathbf{U}w)\,(w^\dagger) \cdot \lambda_d^\dagger \ + \ \partial\,(w \mapsto a - \mathbf{A}w)\,(w^\dagger) \cdot \lambda_a^\dagger\,]\,. \tag{4.21}$$

Applying the well know result [Roc70, Theorem 28.3] finishes the proof. Apart from elementary calculations, our main tools are the strict convexity of f, the smoothness of f^* and

Proposition 4.3. [Roc70, Theorem 23.5(a)-(b)]. *For any proper convex function g and any vector w, it holds $t \in \partial f(w)$ if and only if $x \mapsto \langle x, t\rangle - f(x)$ achieves its supremum at w.*

Proof. Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ be an optimal solution to (4.16).

Complementary Slackness

We fix $\lambda_a^\dagger$ and work with the objective function G of the dual problem, that is,

$$G(\lambda_d) \ := \ \langle \lambda_d, d\rangle \ + \ \langle \lambda_a^\dagger, a\rangle \ - \ f^*\left(\mathbf{U}^\top \lambda_d + \mathbf{A}^\top \lambda_a^\dagger\right)\,.$$

Since f^* is continuously differentiable, so is G. Thus

$$\nabla G(\lambda_d^\dagger) \ := \ d \ - \ \mathbf{U} \cdot \nabla f^*\left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger\right) \ = \ d \ - \ \mathbf{U}w^\dagger\,.$$

Let $\lambda_{d,i}^\dagger$ be the i-th coordinate of $\lambda_d^\dagger$ and $\nabla G_i(\lambda_d^\dagger)$ be the i-th coordinate of $\nabla G(\lambda_d^\dagger)$. To establish (4.18) we will show for all coordinates

$$\begin{array}{rlcl} \text{either} & \lambda_{d,i}^\dagger = 0 & \text{and} & \nabla G_i(\lambda_d^\dagger) \le 0 \\ \text{or} & \lambda_{d,i}^\dagger > 0 & \text{and} & \nabla G_i(\lambda_d^\dagger) = 0\,. \end{array}$$

It is well know that a concave functions g satisfies

$$g(x) - g(y) \ge \nabla g(x)^\top (x - y) \qquad \text{for all } x, y\,. \tag{4.22}$$

But G is concave by the convexity of f^* (cf. Proposition 4.2).

First, we show

$$\nabla G_i(\lambda_d^\dagger) \ \le \ 0 \qquad \text{for all } i \in \{1, \ldots, s\}\,. \tag{4.23}$$

Assume towards a contradiction that $\nabla G_i(\lambda_d^\dagger) > 0$ for some $i \in \{1, \ldots, s\}$. By the continuity of ∇G there exists $\varepsilon > 0$ such that $\nabla G_i(\lambda_d^\dagger + e_i \cdot \varepsilon) > 0$. It follows from (4.22)

$$G(\lambda_d^\dagger + e_i \cdot \varepsilon) \ - \ G(\lambda_d^\dagger) \ \ge \ \nabla G_i(\lambda_d^\dagger + e_i \cdot \varepsilon) \cdot \varepsilon \ > \ 0\,,$$

which contradicts the optimality of $\lambda_d^\dagger$ for (4.16). It follows (4.23).

Next, we assume that $\lambda_{d,i}^\dagger > 0$ and $\nabla G_i(\lambda_d^\dagger) < 0$ for some $i \in \{1, \dots, s\}$. Again, by the continuity of ∇G there exists $\varepsilon > 0$ such that $\nabla G_i(\lambda_d^\dagger - e_i \cdot \varepsilon) < 0$ and $\varepsilon - \lambda_{d,i}^\dagger < 0$. Thus

$$G(\lambda_d^\dagger - e_i \cdot \varepsilon) - G(\lambda_d^\dagger) \geq \nabla G_i(\lambda_d^\dagger - e_i \cdot \varepsilon) \cdot (-\varepsilon) > 0\,,$$

which contradicts the optimality of $\lambda_d^\dagger$. It follows (4.18), that is, we proved complementary slackness.

Primal Feasibility

Since f^* is continuously differentiable it holds

$$\nabla G(\lambda_d^\dagger) \;=\; d \;-\; \mathbf{U} \cdot \nabla f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \;=\; d - \mathbf{U} w^\dagger\,.$$

Thus, by (4.23), $w^\dagger$ satisfies the inequality constraints in (4.15). To prove this for the equality constraints, we view G from a different angel. Let for fixed $\lambda_d^\dagger$

$$G(\lambda_a) \;:=\; \langle \lambda_a, a \rangle \;-\; \left(f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a \right) - \langle \lambda_d^\dagger, d \rangle \right) \;=:\; \langle \lambda_a, a \rangle \;-\; g(\lambda_a)\,.$$

The function g inherits convexity and differentiability from f^*. From the optimality of $\lambda_a^\dagger$ we know that G takes its maximum there. But then by Proposition 4.3 and the differentiability of g it holds

$$a \;\in\; \partial g(\lambda_a^\dagger) \;=\; \left\{ \mathbf{A} \cdot \nabla f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \right\} \;=\; \left\{ \mathbf{A} w^\dagger \right\}. \tag{4.24}$$

Thus $a = \mathbf{A} w^\dagger$. But then $w^\dagger$ satisfies also the equality constraints. We proved (4.19).

Stationarity

First we show

$$\mathbf{U}^\top \lambda_d^\dagger \;+\; \mathbf{A}^\top \lambda_a^\dagger \;\in\; \partial f(w^\dagger)\,. \tag{4.25}$$

By Proposition 4.3 it suffices to show that

$$w \;\mapsto\; \langle w, \mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \rangle \;-\; f(w)$$

achieves its supremum at $w^\dagger$. Since f is strictly convex there exists a unique vector $x^\dagger$ where the above expression achieves its maximum. Since f^* is differentiable it holds

$$w^\dagger \;=\; \nabla f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \;=\; \nabla \left(\lambda \mapsto \langle x^\dagger, \lambda \rangle \;-\; f(x^\dagger) \right) \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \;=\; x^\dagger\,.$$

It follows (4.25). Next we show

$$-\mathbf{U}^\top \in \ \partial\left(w \mapsto d - \mathbf{U}w\right)(w^\dagger) \qquad \text{and} \qquad -\mathbf{A}^\top \in \ \partial\left(w \mapsto d - \mathbf{A}w\right)(w^\dagger)\,. \tag{4.26}$$

To this end, note that

$$\langle -\mathbf{U}^\top e_i, w - w^\dagger\rangle \ = \ (d - \mathbf{U}w)_i \ - \ (d - \mathbf{U}w^\dagger)_i \qquad \text{for all } i \in \{1, \ldots, r\}\ .$$

Thus $-\mathbf{U}^\top \in \partial\left(w \mapsto d - \mathbf{U}w\right)(w^\dagger)$. In the same way it follows $-\mathbf{A}^\top \in \partial\left(w \mapsto d - \mathbf{A}w\right)(w^\dagger)$. From (4.25) and (4.26) we conclude

$$\begin{aligned} 0_n \ &= \ \left(\mathbf{U}^\top \lambda_d^\dagger \ + \ \mathbf{A}^\top \lambda_a^\dagger\right) - \mathbf{U}^\top \lambda_d^\dagger \ - \ \mathbf{A}^\top \lambda_a^\dagger \\ &\in \ [\partial f(w^\dagger) \ + \ \partial\left(w \mapsto d - \mathbf{U}w\right)(w^\dagger) \cdot \lambda_d^\dagger \ + \ \partial\left(w \mapsto a - \mathbf{A}w\right)(w^\dagger) \cdot \lambda_a^\dagger\,]\,. \end{aligned}$$

We have proved (4.21), that is, stationarity.

Dual Feasibility and Conclusion

Dual feasibility (4.20) follows immediately from the optimality of $\lambda_d^\dagger$ for (4.16). Thus, $(\lambda_d^\dagger, \lambda_a^\dagger)$ and $w^\dagger$ satisfy the Karush-Kuhn-Tucker conditions for (4.15). Applying [Roc70, Theorem 28.3] finishes the proof. □

Takeaways For strictly convexity objective functions with continuously differentiable convex conjugate we get a functional relationship of primal and dual solutions via the Karush-Kuhn-Tucker conditions.

Bibliography

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer, New York, NY, 2002.

[Hah98] Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315, March 1998.

[Kle20] Achim Klenke. *Probability Theory: A Comprehensive Course.* Universitext. Springer International Publishing, Cham, 2020.

[MMN22] Boris S. Mordukhovich and Nguyen Mau Nam. *Convex Analysis and Beyond: Volume I: Basic Theory.* Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, 2022.

[New97] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, July 1997.

[Pea09] J. Pearl. *Causality.* Cambridge University Press, 2009.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Springer Science & Business Media, September 2008.

[TB91] Paul Tseng and Dimitri P. Bertsekas. Relaxation Methods for Problems with Strictly Convex Costs and Linear Constraints. *Mathematics of Operations Research*, 16(3):462–481, 1991.

[vdV00] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, June 2000.

[vdvW13] Aad van der vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Science & Business Media, March 2013.

[Wag82] Clifford H. Wagner. Simpson's Paradox in Real Life. *The American Statistician*, 36(1):46–48, 1982.

[WZ19] Yixin Wang and José R. Zubizarreta. Minimal Dispersion Approximately Balancing Weights: Asymptotic Properties and Practical Considerations. *Biometrika*, page asz050, October 2019.

[ZP17] Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010, September 2017.