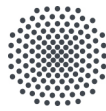


## Todo list



# Title?

Universität Stuttgart



Universität Stuttgart

Ioan Scheffel

February 24, 2023



# Contents

<b>1</b>	<b>Intro for all</b>	<b>5</b>
<b>2</b>	<b>True Introduction</b>	<b>7</b>
<b>3</b>	<b>Introduction</b>	<b>9</b>
<b>4</b>	<b>Consistency</b>	<b>17</b>
<b>5</b>	<b>Asymptotic Normality and Convergence to Gaussian Bridge</b>	<b>25</b>
5.1	Application to Plug In Estimators . . . . .	30
<b>6</b>	<b>Convex Analysis</b>	<b>31</b>
6.1	A Convex Analysis Primer . . . . .	31
6.2	Conjugate Calculus . . . . .	34
6.3	Duality of Optimal Solutions . . . . .	38



# 1 Intro for all

Is study design more important than statistical analysis?

I think, they are at least equal.

But a bad analysis can be undone, whereas a bad design can not.

You have to stick with the data.

If you are not familiar with study design the distinction between randomized and observational study is helpful.

If you read the literature and are unsure about the design of a study, ask for this terms.

You are likely to find an answer.

It is all about how we collect the data.

Say, we want to test the effect of a drug in a study population.

There usually are differences among the units of the study population.

Some are more healthy than others.

We form a treatment and control group, that is, one group takes the drug and the other doesn't.

Then we compare the groups by their health.

Then a critical review comes in. What do you mean by healthy.

We mean this and that.

It seems you did not consider this factor.

Maybe the drug is not effective, but the effect we see in your analysis comes from something else.

What do we answer to this?

A good method to avoid this awkward situation is to randomize.

For every unit of the population we toss a fair coin that decides if they get the drug.

Now comes the critic.

From the tables it seems there is an effect. But what about unknown influence?

We answer: Does the coin now of them?

It is not ideal, but this way you can prevent systematic damage to your analysis.

What if we can't decide who gets treatment?

Don't think treatment has to be something good, it should not carry any label of good or bad.

But what about smoking?

Would you smoke if a coin tells you to?

So this is unethical.

But it is also unethical not to investigate the effects of smoking on the health.

Let's accept, that we sometimes (often?) can not control who gets treatment.

Some smoke, some don't, and we merely observe.

This is typical example for an observational study.

Honestly, this is an oversimplification, but I hope you get the point.

Who still is insulted by the tone will maybe like [Rub07].

In [Rub07] you will find the propensity score.

The propensity score is the individual probability to receive treatment, that is,

$$\mathbf{P}[T = 1|X] \tag{1.1}$$

if  $T$  is the random variable that decides about treatment and  $X$  is the vector that carries your individual information.

This concept goes back to [RR83].

It is maybe worth to stop here and think about this definition and its connection to the two study designs.

Discover it for yourself.

**Reflection.** What is the propensity score in the above example. How does the propensity score behave in rs and os? ♠



## 2 True Introduction

Starting point: Propensity score analysis [RR83]. Two major branches: PS-weighting and matching. We study a weighting method. Procedure: Estimate PS, invert, weight. Problem: Extreme weights when PS close to 0. Bias when estimation model is misspecified. Solution: Balance some measure of dependence simultaneously, e.g. Covariate balancing PS [IR14]. Other solution: Doubly robust estimators [HJ05]. They incorporate treatment and outcome model. Problem: bad results if either are (slightly) misspecified [KS07]. A third option is obtaining weights (seemingly) unrelated to PS. Entropy Balancing [Hai12], ?balancing [Zub15]. Problem: constraint  $\delta = 0$  too strict. Bad convergence. Solution: relax to  $\delta \rightarrow 0$  for  $N \rightarrow \infty$ . Paper with mathematical analysis [WZ19]. Surprising connection to PS. Also doubly robust [ZP17]. This attracted my attention.

I choose different basis as in [GKKW02]. Does analysis work? Consistency? Asymptotic Normality? Beyond that?

[WZ19]: Proofs are substandard. Many mistakes. Missing assumptions. Theorems have to be differently.

This thesis is no erratum of [WZ19], but can be consulted for writing one.

We thank Wang and Zubi for discussions.



### 3 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

$$Y(1)|T = 1 \approx Y(1), \quad (3.1)$$

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

$$(Y(0), Y(1)) \perp T \mid X, \quad (3.2)$$

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

$$\frac{1}{n} \sum_{i=1}^n w_i Y_i. \quad (3.3)$$

The weights should satisfy (in a broader sense)

$$w_i \cdot Y_i \rightarrow Y(1) \quad \text{for } N \rightarrow \infty. \quad (3.4)$$

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

#### The algorithm

**Problem 3.1.**

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

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

$$[w_1, \dots, w_n]^\top \mapsto \sum_{i=1}^n f(w_i) \quad (3.5)$$

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

$$[\lambda_1, \dots, \lambda_n]^\top \mapsto \sum_{i=1}^n f^*(\lambda_i) \quad (3.6)$$

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

$$f: [0, \infty) \rightarrow \mathbb{R}, \quad w \mapsto \begin{cases} 0 & \text{if } w = 0, \\ w \log w & \text{else.} \end{cases} \quad (3.7)$$

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

$$(f')^{-1} = \lambda \mapsto e^{\lambda-1} \quad (3.8)$$

Thus

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

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

### Sample Variance

We define the sample variance to be

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad w \mapsto (w - 1/n)^2 \quad (3.9)$$

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

$$(f')^{-1} = \lambda \mapsto \frac{\lambda}{2} + \frac{1}{n} \quad (3.10)$$

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 population. 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)

### 3 Introduction

consistency [GKKW02, Definition 1.1] for noiseless estimands is

$$\mathbf{E} \left[ \int_{\mathcal{X}} \left| \sum_{k=1}^N B_k(x) \cdot m(X_k) - m(x) \right|^2 \mathbf{P}_X(dx) \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3.11)$$

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

$$\mathbf{P} \left[ \left| \sum_{k=1}^N B_k(X) \cdot m(X_k) - m(X) \right| \geq \varepsilon \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3.12)$$

**Proof.** By Markov's inequality it holds

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

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

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

#### Partitioning Estimates

We consider a partition  $\mathcal{P}_N = \{A_{N,1}, A_{N,2}, \dots\}$  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

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

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 3.2.** *If for each sphere  $S$  centered at the origin*

$$\max_{j: A_{N,j} \cap S \neq \emptyset} \text{diam } A_{N,j} \rightarrow 0 \quad \text{for } N \rightarrow \infty \quad (3.13)$$

and

$$\frac{\#\{j: A_{N,j} \cap S \neq \emptyset\}}{N} \rightarrow 0 \quad \text{for } N \rightarrow \infty \quad (3.14)$$

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

**Proof.** [GKKW02, Theorem 4.2.] □

**Corollary 3.2.1.** *Assume  $\mathbf{E}[m(X)^2] < \infty$  and the basis functions  $B$  belong to a partitioning estimate. Furthermore assume that the conditions of Theorem 3.2 are met. Then it holds for all  $\varepsilon > 0$*

$$\mathbf{P} \left[ \left| \sum_{k=1}^N B_k(X) \cdot m(X_k) - m(X) \right| \geq \varepsilon \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty . \quad (3.15)$$

## Kernel Estimates

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

$$B_k(x) := \frac{K\left(\frac{x-X_k}{h_n}\right)}{\sum_{i=1}^N K\left(\frac{x-X_i}{h_n}\right)} . \quad (3.16)$$

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

**Theorem 3.3.** *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*

$$\mathbf{1}_{\{x \in S_{0,R}\}} \geq K(x) \geq b \cdot \mathbf{1}_{\{x \in S_{0,r}\}} \quad (3.17)$$

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

### 3 Introduction

**Corollary 3.3.1.** *Assume  $\mathbf{E}[m(X)^2] < \infty$  and the basis functions  $B$  belong to a kernel estimate. Furthermore assume that the conditions of Theorem 3.3 are met. Then it holds for all  $\varepsilon > 0$*

$$\mathbf{P} \left[ \left| \sum_{k=1}^N B_k(X) \cdot m(X_k) - m(X) \right| \geq \varepsilon \right] \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3.18)$$

In the sequel we mainly work with the dual problem.

## Dual Problem

**Theorem.** *The dual of Problem 3.1 is the unconstrained optimization problem*

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

where

$$f^* : \mathbb{R} \rightarrow \mathbb{R}, \quad x^* \mapsto x^* \cdot (f')^{-1}(x^*) - f((f')^{-1}(x^*))$$

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 3.1 is given by

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

## Plan of proof

We want to apply Theorem 6.4. 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 6.4 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

$$\mathbf{B}(\mathbf{X}) := \begin{bmatrix} B(X_1), \dots, B(X_n) \end{bmatrix}$$

and the constraint matrices

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

By Example 6.1 the convex conjugate of the objective function of Problem 3.1 is

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

Before we apply Theorem 6.4 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

$$\lambda_d =: \begin{bmatrix} \rho \\ \lambda^+ \\ \lambda^- \end{bmatrix} \tag{3.19}$$

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

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

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

### 3 Introduction

#### Non-negativity constraints

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

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

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

$$\tilde{\lambda} := \left[ \lambda_1^+, \dots, \lambda_i^+ - (\lambda_i^+ \wedge \lambda_i^-), \dots, \lambda_N^+, \lambda_1^-, \dots, \lambda_i^- - (\lambda_i^+ \wedge \lambda_i^-), \dots, \lambda_N^-, \lambda_0 \right]^\top. \quad (3.20)$$

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

$$G(\tilde{\lambda}) - G(\lambda) = 2N \cdot \delta_i \cdot (\lambda_i^+ \wedge \lambda_i^-) > 0, \quad (3.21)$$

which contradicts the optimality of  $\lambda$ . 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

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

We apply Theorem 6.4 to finish the proof. □

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

## 4 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  $\lambda^*$  denote the vector with coordinates

$$\lambda_i^* := f'(1/\pi_i) - \lambda_0^\dagger, \quad (4.1)$$

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

**Theorem 4.1.** *For all  $\varepsilon > 0$  it holds*

$$\mathbf{P} \left[ \left\| \lambda^\dagger - \lambda^* \right\| \geq \varepsilon \right] \rightarrow 0 \quad \text{for } N \rightarrow \infty. \quad (4.2)$$

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 4.1.** *Let  $m \in \mathbb{N}$  and  $g : \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$  be convex. Then for all  $y \in \mathbb{R}^m$  and  $\varepsilon > 0$*

$$\inf_{\|\Delta\|=\varepsilon} g(y + \Delta) - g(y) \geq 0 \quad (4.3)$$

*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 (4.3) is true. Then it holds

$$g(x) < g(y^*) \quad \text{for some } x \in \mathbb{R}^m \setminus (y + \varepsilon B). \quad (4.4)$$

#### 4 Consistency

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  $\Delta_x$  with  $\|\Delta_x\| = \varepsilon$  such that

$$\theta x + (1 - \theta)y^* = y + \Delta_x. \quad (4.5)$$

It follows

$$\begin{aligned} g(y^*) &\leq g(y) \leq g(y + \Delta_x) = g(\theta x + (1 - \theta)y^*) \\ &\leq \theta g(x) + (1 - \theta)g(y^*) < g(y^*), \end{aligned} \quad (4.6)$$

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 (4.3) being true, the equality is due to (4.5), the third inequality is due to the convexity of  $g$  and the strict inequality is due to (4.4). Thus every local minimum of  $g$  in  $y + \varepsilon B$  is also a global minimum.  $\square$

**Proof.** The objective function  $G$  of the dual satisfies

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

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

It is well known that a differentiable convex function  $g$  satisfies

$$g(x) - g(y) \geq \nabla g(y)^\top (x - y) \quad \text{for all } x, y. \quad (4.7)$$

The gradient of

$$g := (\lambda, \lambda_0) \mapsto \frac{1}{N} \sum_{i=1}^N [T_i \cdot f^*(\lambda_0 + \langle B(X_i), \lambda \rangle) - (\lambda_0 + \langle B(X_i), \lambda \rangle)] \quad (4.8)$$

is

$$\nabla g = (\lambda, \lambda_0) \mapsto \frac{1}{N} \sum_{i=1}^N \left[ T_i \cdot (f')^{-1}(\lambda_0 + \langle B(X_i), \lambda \rangle) - 1 \right] [B(X_i)^\top, 1]^\top \quad (4.9)$$

Thus

$$\begin{aligned} &G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \\ &\geq -\frac{1}{N} \sum_{i=1}^N \left[ B(X_i)^\top, 1 \right] \cdot \begin{bmatrix} \Delta \\ \Delta_0 \end{bmatrix} \left( 1 - T_i \cdot (f')^{-1} \left( \langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger \right) \right) \\ &\quad + \langle \delta, |\lambda^* + \Delta| - |\lambda^*| \rangle. \end{aligned} \quad (4.10)$$

We fix  $\tilde{\varepsilon} > 0$  and establish the lower bound  $-\tilde{\varepsilon}$  with probability going to 1 as  $N \rightarrow \infty$ .

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

$$\left[ B(X_i)^\top, 1 \right] \cdot \begin{bmatrix} \Delta \\ \Delta_0 \end{bmatrix} \lesssim \|(\Delta, \Delta_0)\| = \varepsilon. \quad (4.11)$$

Next, we see that

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \left( 1 - T_i \cdot (f')^{-1} \left( \langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger \right) \right) \\ & \lesssim \frac{1}{N} \sum_{i=1}^N \left| 1 - \frac{T_i}{\pi_i} \right| + \frac{1}{N} \sum_{i=1}^N \left| \langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger - f' \left( \frac{1}{\pi_i} \right) \right| \\ & =: S_N + M_N. \end{aligned} \quad (4.12)$$

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

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  $1/\pi$ ?)

$$\mathbf{P}[S_N \geq \tilde{\varepsilon}/(4\varepsilon)] \rightarrow 0 \quad \text{for } N \rightarrow \infty. \quad (4.13)$$

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

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

The convergence is due to the universal consistency of  $B$ . This establishes the desired bound of  $\tilde{\varepsilon}/(2\varepsilon)$  in (4.12). Together with (4.11) we conclude that the **first term** in (4.10) is bounded below by  $-\tilde{\varepsilon}/2$  with probability going to 1 as  $N \rightarrow \infty$ .

## Second Term

It holds

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

Since  $\delta \geq 0$  we get

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

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

## Conclusion

With the analysis of the **first** and **second term** in (4.10) we conclude

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq -\tilde{\varepsilon} \quad (4.14)$$

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

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0 \quad (4.15)$$

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

$$\inf_{\|(\Delta, \Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0 \quad (4.16)$$

with probability going to 1 as  $N \rightarrow \infty$ . Thus by Lemma 4.1

$$\mathbf{P} \left[ \left\| \lambda^\dagger - \lambda^* \right\| \geq \varepsilon \right] \rightarrow 0 \quad \text{for } N \rightarrow \infty. \quad (4.17)$$

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

## Consistency for Inverse Propensity Score

**Theorem 4.2.** *For all  $\varepsilon > 0$  it holds*

$$\mathbf{P} [|w(X) - 1/\pi(X)| \geq \varepsilon] \rightarrow 0 \quad \text{for } N \rightarrow \infty.$$

*Furthermore, it holds*

$$\mathbf{E} \left[ |w(X) - 1/\pi(X)|^2 \right]^{1/2} \rightarrow 0 \quad \text{for } N \rightarrow \infty.$$

**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.

$$\begin{aligned}
\left| w(X) - \frac{1}{\pi(X)} \right| &= \left| (f')^{-1} \left( \langle B(X), \lambda^\dagger \rangle + \lambda_0^\dagger \right) - \frac{1}{\pi(X)} \right| \\
&\lesssim \left| \langle B(X), \lambda^\dagger - \lambda^* \rangle \right| + \left| \langle B(X), \lambda^\dagger \rangle + \lambda_0^\dagger - f' \left( \frac{1}{\pi(X)} \right) \right| \\
&\lesssim \left\| \lambda^\dagger - \lambda^* \right\| + \left| \sum_{i=1}^N B_k(X) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi(X)} \right) \right| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \leq \varepsilon,
\end{aligned} \tag{4.18}$$

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

If we prove boundedness,  $L_2$ -convergence follows readily.

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

□

## Consistency of the Weighted Mean

**Theorem 4.3.** *For all  $\varepsilon > 0$  it holds*

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

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

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

**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

#### 4 Consistency

note, that for  $i > n$  we may choose  $w_i = 1/\pi_i$ .

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

##### Analysis of $R_1$

By the Cauchy-Schwarz inequality it holds

$$\left| \frac{1}{N} \left( \sum_{i=1}^n w_i B(X_i) - \sum_{i=1}^N B(X_i) \right)^\top \mathbf{Y}(1) \right| \leq \|\delta\| \|\mathbf{Y}(1)\| \lesssim \|\delta\| N \rightarrow 0 \quad \text{for } N \rightarrow \infty.$$

##### Analysis of $R_2$

This calculation will be central to the asymptotic normality.

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

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

##### Analysis of $R_3$

By Theorem? it holds

$$\mathbf{P}[R_3 \geq \varepsilon] \leq \frac{\mathbf{E}[|w(X) - 1/\pi(X)|^2]^{1/2}}{\varepsilon} \rightarrow 0 \quad \text{for } N \rightarrow \infty.$$



## Analysis of $R_4$

$$\begin{aligned}
\mathbf{E}[Y(T) \cdot (T / \pi(X))] &= \mathbf{E}[Y(1) / \pi(X) | T = 1] \cdot \mathbf{P}[T = 1] \\
&= \int_{\mathcal{X}} \mathbf{E}[Y(1) | X = x, T = 1] \cdot (\mathbf{P}[T = 1] / \pi(x)) \mathbf{P}_{X|T}(dx | 1) \\
&= \int_{\mathcal{X}} [Y(1) | X = x] \mathbf{P}_X(dx) \\
&= \mathbf{E}[Y(1)].
\end{aligned} \tag{4.19}$$

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

$$\left| \frac{1}{N} \sum_{i=1}^n w_i Y_i - \mathbf{E}[Y(1)] \right| \leq \varepsilon \tag{4.20}$$

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

$$\left| \frac{1}{N} \sum_{i=1}^n w_i Y_i - \mathbf{E}[Y(1)] \right|^p = \left| \frac{1}{N} \sum_{i=1}^n w_i (Y_i - \mathbf{E}[Y(1)]) \right|^p \tag{4.21}$$

$$\leq \frac{1}{N} \sum_{i=1}^n w_i |Y_i - \mathbf{E}[Y(1)]|^p \leq (2M)^p \tag{4.22}$$

□



## 5 Asymptotic Normality and Convergence to Gaussian Bridge

**Theorem 5.1.** *Under conditions the partition estimate has*

$$\mathbf{E} \|m_N - m\|^2 \leq C_{\mathbf{P}} N^{-\frac{2}{d+2}} \quad (5.1)$$

**Theorem 5.2.** *Under conditions the kernel estimate has*

$$\mathbf{E} \|m_N - m\|^2 \leq C_{\mathbf{P}} N^{-\frac{2}{d+2}} \quad (5.2)$$

### Learning Rates for the Dual

**Theorem 5.3.** *Under conditions*

$$\mathbf{P} \left[ \left\| \lambda^\dagger - \lambda^* \right\| \leq C_{\mathbf{P}} C_\tau \varepsilon_n \right] \geq 1 - \tau, \quad (5.3)$$

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

### Plan of Proof

The path is similar to that of the previous chapter. We will use strong convexity to obtain a quadratic term. This gives us the flexibility to obtain learning rates.

**Theorem 5.4.** (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  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  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  $i = 1, \dots, n$ . Then we have*

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

**Proof.** See [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.  $\square$

### Strong Convexity

For simpler notation we define  $B_0(x) := 1$ . The Hessian matrix of  $g$  as in ?? is

$$\nabla^2 g(\lambda_0^\dagger, \lambda^\dagger) = \left[ \frac{1}{N} \sum_{i=1}^n \left( (f')^{-1} \right)' \left( \lambda_0^\dagger + \langle B(X_i), \lambda^\dagger \rangle \right) B_k(X_i) \cdot B_l(X_i) \right]_{0 \leq k, l \leq N}. \quad (5.4)$$

Since  $\left( (f')^{-1} \right)' > 0$  by the strict convexity of  $f$ ,  $(\lambda_0^\dagger, \lambda^\dagger) \in \Theta$ , where  $\Theta$  is compact parameter space, and the support of  $X$  is compact it holds

$$\left( (f')^{-1} \right)' \left( \lambda_0^\dagger + \langle B(X_i), \lambda^\dagger \rangle \right) > \frac{1}{C}. \quad (5.5)$$

Consider

$$\mathbf{B}(X) := \begin{bmatrix} 1 & B_1(X_1) & \cdots & B_N(X_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & B_1(X_n) & \cdots & B_N(X_n) \end{bmatrix} \in \mathbb{R}^{n \times (N+1)} \quad (5.6)$$

Since  $B_k(x) \geq 0$  for all  $k$  it holds

$$\nabla^2 g(\lambda_0^\dagger, \lambda^\dagger) \succcurlyeq \frac{1}{C} N^{-1} \mathbf{B}(X)^\top \mathbf{B}(X) \succcurlyeq \frac{\lambda_{\min}(N^{-1} \mathbf{B}(X)^\top \mathbf{B}(X))}{C} \cdot \mathbf{I} \succcurlyeq \frac{1}{C \sqrt{\log(1/\tau)}} \cdot \mathbf{I} \quad (5.7)$$

with probability greater than  $1 - \tau$ . Thus  $g$  is strongly convex with parameter  $\left( C \sqrt{\log(1/\tau)} \right)^{-1}$  with probability greater than  $1 - \tau$ . We analyze  $1 - T_i/\pi_i$ . It holds  $\mathbf{E}[1 - T_i/\pi_i] = 0$ . Furthermore,

$$\left| 1 - \frac{T_i}{\pi_i} \right| \leq 1 \wedge \frac{1 - \pi_i}{\pi_i} \leq 1 + \frac{1 - \pi_i}{\pi_i} = \frac{1}{\pi_i} \leq \frac{1}{C_\pi} \quad \text{almost surely.} \quad (5.8)$$

Also

$$\mathbf{E} \left[ \left| 1 - \frac{T_i}{\pi_i} \right|^2 \right] = 1 - 2\mathbf{E} \left[ \frac{T_i}{\pi_i} \right] + \mathbf{E} \left[ \frac{T_i^2}{\pi_i^2} \right] = \mathbf{E} \left[ \frac{1}{\pi_i} \right] - 1 \leq \frac{1}{C_\pi}. \quad (5.9)$$

Thus, by Bernstein's inequality, it holds

$$\mathbf{P} \left[ \left| \frac{1}{n} \sum_{i=1}^n 1 - \frac{T_i}{\pi_i} \right| \lesssim \sqrt{\frac{\log(e/\tau)}{N}} \right] \geq 1 - \tau \quad \text{for all } \tau > 0. \quad (5.10)$$

Next, we analyze

$$\sum_{k=1}^N B_k(X_i) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) - \mathbf{E}[\cdots]. \quad (5.11)$$

To this end,

$$\begin{aligned}
& \left| \sum_{k=1}^N B_k(X_i) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) \right| \\
& \leq \left| \sum_{k=1}^N B_k(X_i) \cdot \left( f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) \right) \right| \leq \sum_{k=1}^N B_k(X_i) \cdot \left| f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) \right| \\
& \leq 2f' \left( \frac{1}{C_\pi} \right).
\end{aligned}$$

Also

$$\begin{aligned}
& \mathbf{E} \left[ \left| \sum_{k=1}^N B_k(X_i) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) - \mathbf{E}[\dots] \right|^2 \right]^{1/2} \\
& \leq \mathbf{E} \left[ \left| \sum_{k=1}^N B_k(X_i) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) \right|^2 \right]^{1/2} + \left| \mathbf{E} \left[ \sum_{k=1}^N B_k(X_i) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) \right] \right| \\
& \lesssim N^{-1/(d+2)} + 2f' \left( \frac{1}{C_\pi} \right).
\end{aligned}$$

Thus, by Bernstein's inequality, it holds

$$\mathbf{P} \left[ \left| \frac{1}{N} \sum_{i,k=1}^N B_k(X_i) \cdot f' \left( \frac{1}{\pi_k} \right) - f' \left( \frac{1}{\pi_i} \right) - \mathbf{E}[\dots] \right| \lesssim \sqrt{\frac{\log(e/\tau)}{N}} \right] \geq 1 - \tau \quad \text{for all } \tau > 0. \quad (5.12)$$

Similar to the analysis of the first and second term we get

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq \|\Delta\| \left( \frac{\|\Delta\|}{C\sqrt{\log(1/\tau)}} - \sqrt{\frac{\log(1/\tau)}{N}} - N^{-1/(d+2)} \right) \quad (5.13)$$

with probability greater than  $1 - \tau$ . Thus for

$$\|\Delta\| = C \log(1/\tau) N^{-1/(d+2)} \quad (5.14)$$

it holds

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0 \quad (5.15)$$

with probability greater than  $1 - \tau$ . We conclude

$$\mathbf{P} \left[ \left\| \lambda^\dagger - \lambda^* \right\| \lesssim \log(1/\tau) N^{-1/(d+2)} \right] \geq 1 - \tau. \quad (5.16)$$

### Learning Rates for the Primal

**Theorem 5.5.** *Under conditions the weights satisfy*

$$\mathbf{E}[|w^\dagger(X) - 1/\pi(X)|^2]^{1/2} \leq C_{\mathbf{P}} \sqrt{\log(n)} n^{-1/(2+d)} \quad (5.17)$$

where  $\text{varepsilon}_n$  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.**

$$\mathbf{E}[|w^\dagger(X) - 1/\pi(X)|^2]^{1/2} = \mathbf{E} \left[ \left| (f')^{-1} \left( \langle B(X), \lambda^\dagger \rangle + \lambda_0^\dagger \right) - 1/\pi(X) \right|^2 \right]^{1/2} \quad (5.18)$$

$$\leq \left| (f')^{-1} \right|_L (I_1 + I_2) \quad (5.19)$$

where

$$I_1 := \left( \mathbf{E} \left\| \lambda^\dagger - \lambda^* \right\|^2 \right)^{1/2} \quad (5.20)$$

$$I_2 := \mathbf{E} \left[ \left| \sum_{k=1}^N B_k(X) \cdot f'(X_k) - f'(X) \right|^2 \right]^{1/2} \quad (5.21)$$

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.

$$I_1 \leq C_\tau n^{-1/(d+2)} + \sqrt{\tau} \cdot \text{diam } \Theta \quad (5.22)$$

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

$$I_1 \lesssim \log(N) \cdot N^{-1/(2+d)} \quad (5.23)$$

Thus

$$\mathbf{E}[|w^\dagger(X) - 1/\pi(X)|^2]^{1/2} \lesssim \log(N) \cdot N^{-1/(2+d)} \quad (5.24)$$

□

### Asymptotic Normality of the Weighted Mean

**Theorem 5.6.** (Slutzky's theorem) *Let  $(E, d)$  be a metric space and let  $X, X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  be random variables with values in  $E$ . Assume  $X_n \rightarrow X$  in distribution and  $d(X_n, Y_n) \rightarrow 0$  in probability. Then  $Y_n \rightarrow X$  in distribution.*

**Proof.** [Kle20, Theorem 13.8]

□

**Theorem 5.7.** *Under conditions the stochastic process*

$$\sqrt{N} \left( \frac{1}{N} \sum_{i=1}^n w_i \mathbf{1}_{\{Y_i \leq z\}} - \mathbf{P}[Y(1) \leq z] \right)_{z \in \mathbb{R}}. \quad (5.25)$$

*converges in  $l^\infty(\mathbb{R})$  to a Gaussian process with mean 0 and covariance ??.*

**Proof.** For fixed  $z \in \mathbb{R}$  we use the following error decomposition. Recall  $\pi(x) := \mathbf{P}[T = 1 | X = x]$  and  $w(x) := (f')^{-1} \left( \langle B(x), \lambda^\dagger \rangle + \lambda_0^\dagger \right)$ , where  $(\lambda^\dagger, \lambda_0^\dagger)$  is the optimal dual solution. We also write  $F_{Y(1)}(z|x) = \mathbf{P}[Y(1) \leq z | X = x]$  and  $F_{Y(1)}(z) = \mathbf{P}[Y(1) \leq z]$ .

$$\begin{aligned} & \sqrt{N} \left( \frac{1}{N} \sum_{i=1}^n w(X_i) \mathbf{1}_{\{Y_i \leq z\}} - \mathbf{P}[Y(1) \leq z] \right) \\ &= \sqrt{N} \sum_{k=1}^N \left[ \frac{1}{N} \left( \sum_{i=1}^n w(X_i) B_k(X_i) - \sum_{i=1}^N B_k(X_i) \right) F_{Y(1)}(z|X_k) \right] \\ & \quad + \sqrt{N} \sum_{i=1}^N \left[ \frac{T_i \cdot w(X_i) - 1}{N} \left( F_{Y(1)}(z|X_i) - \sum_{k=1}^N B_k(X_i) \cdot F_{Y(1)}(z|X_k) \right) \right] \\ & \quad + \sqrt{N} \left( \frac{1}{N} \sum_{i=1}^N \left[ T_i \left( w(X_i) - \frac{1}{\pi(X_i)} \right) (\mathbf{1}_{\{Y_i \leq z\}} - F_{Y(1)}(z|X_i)) \right] \right) \\ & \quad + \sqrt{N} \left( \frac{1}{N} \sum_{i=1}^N \frac{T_i}{\pi(X_i)} (\mathbf{1}_{\{Y_i \leq z\}} - F_{Y(1)}(z|X_i)) + (F_{Y(1)}(z|X_i) - F_{Y(1)}(z)) \right) \\ &=: R_1(z) + R_2(z) + R_3(z) + R_4(z) \end{aligned}$$

We show that  $\sup_{z \in \mathbb{R}} |R_i(z)| \rightarrow 0$  in probability for  $i = 1, 2, 3$ . The term  $(R_4)_{z \in \mathbb{R}}$  is  $\mathbf{P}$ -Donsker and determines the covariance.  $\square$

## 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

$$f_\Phi(T, X, Y) = \frac{T}{\pi(X)} (\Phi(Y) - \mathbf{E}[\Phi(Y)|X]) + \mathbf{E}[\Phi(Y)|X]. \quad (5.26)$$

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

## 5.1 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 Bootstrapping from the weighted distribution is also sensible .



## 6 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 omitted, 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 omitted in the paper.

### 6.1 A Convex Analysis Primer

Throughout this section let  $n \in \mathbb{N}$ .

#### Sets

A subset  $C \subseteq \mathbb{R}^n$  is called **convex set**, if for all  $x, y \in C$  and all  $\theta \in [0, 1]$ , we have  $\theta x + (1 - \theta)y \in C$ . Many set operations preserve convexity. Among them forming the **Cartesian product** of two convex sets, **intersection** of a collection of convex sets and taking the **inverse image under linear functions**.

The classical theory evolves around the question if convex sets can be separated.

**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 contained in  $H$ .

We need a refined concept of interiors, since some convex sets have empty interior. To this end, we call a set  $A \subseteq \mathbb{R}^n$  **affine set**, if  $\alpha x + (1 - \alpha)y \in A$  for all  $x, y \in A$  and  $\alpha \in \mathbb{R}$ . The **affine hull**  $\text{aff}(\Omega)$  of a set  $\Omega \subseteq \mathbb{R}^n$  is the smallest affine set that includes  $\Omega$ . We define the **relative interior**  $\text{ri}\Omega$  of a set  $\Omega \subseteq \mathbb{R}^n$  to be the interior relative to the affine hull, that is,

$$\text{ri}(\Omega) := \{x \in \Omega \mid \exists \varepsilon > 0 : (x + \varepsilon B_{\mathbb{R}^n}) \cap \text{aff}(\Omega) \subset \Omega\}. \quad (6.1)$$

**Theorem 6.1.** (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  $\text{ri}(C_1) \cap \text{ri}(C_2) = \emptyset$ .*

**Proof.** [Roc70, Theorem 11.3] □

We collect some useful properties of relative interiors before we get on to convex functions.

**Proposition 6.1.** *Let  $C$  be a non-empty convex set in  $\mathbb{R}^n$ . The following holds:*

- (i)  $\text{ri}(C) \neq \emptyset$  if and only if  $C \neq \emptyset$
- (ii)  $\text{cl}(\text{ri } C) = \text{cl } C$  and  $\text{ri}(\text{cl } C) = \text{ri}(C)$
- (iii)  $\text{ri}(C) = \{z \in C : \text{for all } x \in C \text{ there exists } t > 0 \text{ such that } z + t(z - x) \in C\}$
- (iv) Suppose  $\bigcap_{i \in I} C_i \neq \emptyset$  for a finite index set  $I$ . Then  $\text{ri}(\bigcap_{i \in I} C_i) = \bigcap_{i \in I} \text{ri}(C_i)$ .
- (v) Let  $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a linear function. Then  $\text{ri } L(C) = L(\text{ri } C)$ . If it also holds  $L^{-1}(\text{ri } C) \neq \emptyset$ , we have  $\text{ri } L^{-1}(C) = L^{-1}(\text{ri } C)$ .
- (vi)  $\text{ri}(C_1 \times C_2) = \text{ri } C_1 \times \text{ri } C_2$

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

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

$$z_i + t(z_i - x_i) \in C_i \quad \text{for all } i \in \{1, 2\}. \quad (6.2)$$

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

$$z_i + t_i(z_i - x_i) \in C_i \quad \text{for all } i \in \{1, 2\}. \quad (6.3)$$

If  $t_1 = t_2$  we recover (6.2) from (6.3). By (iii) it holds  $(z_1, z_2) \in \text{ri}(C_1 \times C_2)$ . If  $t_1 < t_2$  we define  $\theta := \frac{t_1}{t_2} \in (0, 1)$ . Consider (6.3) with  $i = 2$ , together with  $z_2 \in C_2$  and the convexity of  $C_2$ . It follows

$$z_2 + t_1(z_2 - x_2) = \theta \cdot (z_2 + t_2(z_2 - x_2)) + (1 - \theta) \cdot z_2 \in C_2. \quad (6.4)$$

Now we consider (6.4) and (6.3) with  $i = 1$ . This gives (6.2) with  $t = t_1$ . As before, it follows  $(z_1, z_2) \in \text{ri}(C_1 \times C_2)$ . If  $t_1 > t_2$  similar arguments lead to the same result. We have proven  $\text{ri}(C_1 \times C_2) \supseteq \text{ri } C_1 \times \text{ri } C_2$  and equality.  $\square$

## Functions

A function  $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is called **convex function**, if the area above its graph, that is, its epigraph(cf. [MMN22, §2.4.1]), is convex. We shall often use an equivalent definition. To this end, a function  $f$  is convex if and only if

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \text{for all } x, y \in \mathbb{R}^n \text{ and all } \theta \in [0, 1]. \quad (6.5)$$

This definition extends to convex combinations  $\theta_1, \dots, \theta_m \in [0, 1]$  with  $\sum_{i=1}^m \theta_i = 1$ , that is, a function  $f$  is convex if and only if

$$f\left(\sum_{i=1}^m \theta_i x_i\right) \leq \sum_{i=1}^m \theta_i f(x_i) \quad \text{for all } x_1, \dots, x_m \in \mathbb{R}^n. \quad (6.6)$$

We call a function **strictly convex** if the inequality in (6.5) is strict.

We define the **domain**  $\text{dom } f$  of a convex function  $f$  to be the set where  $f$  is finite, that is,

$$\text{dom } f := \{x \in \mathbb{R}^n : f(x) < \infty\}. \quad (6.7)$$

The domain of a convex function is convex. We say that  $f$  is a **proper function** if  $\text{dom } f \neq \emptyset$ .

For any  $\bar{x} \in \text{dom } f$  we call  $x^* \in \mathbb{R}^n$  a **subgradient** of  $f$  at  $\bar{x}$  if for all  $x \in \mathbb{R}^n$  it holds

$$\langle x^*, x - \bar{x} \rangle \leq f(x) - f(\bar{x}). \quad (6.8)$$

We denote the collection of all subgradients at  $\bar{x}$ , that is, the **subdifferential** of  $f$  at  $\bar{x}$ , as  $\partial f(\bar{x})$ . If  $f$  is differentiable at  $\bar{x}$  it holds  $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$  and thus

$$\langle \nabla f(\bar{x}), x - \bar{x} \rangle \leq f(x) - f(\bar{x}). \quad (6.9)$$

We call a differentiable function  $f$  **strongly convex** with parameter  $m > 0$  if for all  $x, y \in \text{dom } f$  it holds

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2. \quad (6.10)$$

If  $f$  is twice continuously differentiable, then it is strongly convex with parameter  $m > 0$  if and only if the matrix

$$\nabla^2 f(x) - m \cdot \mathbf{I} \quad \text{is positive semi-definite for all } x \in \text{dom } f, \quad (6.11)$$

where  $\nabla^2 f$  is the Hessian Matrix.

One important application of convex functions is in optimization. There we often analyse a dual problem instead, which relies on the notion of **convex conjugate**  $f^* : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  of  $f$  defined by

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} \langle x^*, x \rangle - f(x). \quad (6.12)$$

Even for arbitrary functions, the convex conjugate is convex(cf.). Like in differential calculus, there exist sum and chain rule for computing the convex conjugate.

## 6.2 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 6.1.** (Convex conjugate) *Given a function  $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ , the **convex conjugate**  $f^* : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  of  $f$  is defined as*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \quad (6.13)$$

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

**Proposition 6.2.** *Let  $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$  be a proper function. Then its convex conjugate  $f^* : \mathbb{R}^n \rightarrow (-\infty, \infty]$  is convex.*

**Proof.** [MMN22, Proposition 4.2] □

**Theorem 6.2.** *Let  $f, g : \mathbb{R}^n \rightarrow (-\infty, \infty]$  be proper convex functions and  $\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g)) \neq \emptyset$ . Then we have the **conjugate sum rule***

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*) \quad (6.14)$$

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

$$(f + g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \quad (6.15)$$

**Proof.** [MMN22, Theorem 4.27(c)] □

**Theorem 6.3.** Let  $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$  be a linear map (matrix) and  $g : \mathbb{R}^n \rightarrow (-\infty, \infty]$  a proper convex function. If  $\text{Im}(A) \cap \text{ri}(\text{dom}(g)) \neq \emptyset$  it follows the **conjugate chain rule**

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \quad (6.16)$$

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

**Proof.** [MMN22, Theorem 4.28(c)] □

**Example 6.1.** Let  $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$  be a proper convex function, that is,  $\text{dom } f \neq \emptyset$  and  $f$  is convex. In steps we apply the conjugate chain and sum rule, together with mathematical induction, to prove the conjugate relationship

$$\begin{aligned} S_{f,n} : \mathbb{R}^n &\rightarrow \overline{\mathbb{R}}, & (x_1, \dots, x_n) &\mapsto \sum_{i=1}^n f(x_i), \\ S_{f,n}^* : \mathbb{R}^n &\rightarrow \overline{\mathbb{R}}, & (x_1^*, \dots, x_n^*) &\mapsto \sum_{i=1}^n f^*(x_i^*). \end{aligned}$$

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

$$p_i : \mathbb{R}^n \rightarrow \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto x_i. \quad (6.17)$$

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

$$p_i^* : \mathbb{R} \rightarrow \mathbb{R}^n, \quad x \mapsto e_i \cdot x. \quad (6.18)$$

For the inverse image of the adjoint of  $p_i$  it holds

$$(p_i^*)^{-1} \{(x_1^*, \dots, x_n^*)\} = \begin{cases} \{x_i^*\}, & \text{if } x_j^* = 0 \text{ for all } j \neq i, \\ \emptyset & \text{else.} \end{cases} \quad (6.19)$$

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

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

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

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

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

$$(S_{f,n} \circ p_{\{1, \dots, n\}})^*(x_1^*, \dots, x_{n+1}^*) = \begin{cases} S_{f,n}^*(x_1^*, \dots, x_n^*) & \text{if } x_{n+1}^* = 0, \\ \infty & \text{else,} \end{cases} \quad (6.20)$$

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

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

and

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

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.

### 6.3 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 6.1).

**Theorem 6.4.** *Consider the optimization problem*

$$\begin{aligned} & \underset{w \in \mathbb{R}^n}{\text{minimize}} && f(w) \\ & \text{subject to} && \mathbf{U}w \geq d, \\ & && \mathbf{A}w = a, \end{aligned} \tag{6.21}$$

*and its dual problem*

$$\begin{aligned} & \underset{\lambda_d \in \mathbb{R}^r, \lambda_a \in \mathbb{R}^s}{\text{maximize}} && \langle \lambda_d, d \rangle + \langle \lambda_a, a \rangle - f^*(\mathbf{U}^\top \lambda_d + \mathbf{A}^\top \lambda_a) \\ & \text{subject to} && \lambda_d \geq 0. \end{aligned} \tag{6.22}$$

*Let  $(\lambda_d^\dagger, \lambda_a^\dagger)$  be an optimal solution to (6.22). If the objective function  $f$  of (6.21) is strictly convex and its convex conjugate  $f^*$  is continuously differentiable, then the unique optimal solution to (6.21) is given by*

$$w^\dagger = \nabla f^*(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger). \tag{6.23}$$

#### Plan of Proof

We show that  $w^\dagger$  and  $(\lambda_d^\dagger, \lambda_a^\dagger)$  meet the Karush-Kuhn-Tucker conditions for 6.21, that is, **complementary slackness**

$$\langle \lambda_d^\dagger, d - \mathbf{U}w^\dagger \rangle = 0, \tag{6.24}$$

#### primal and dual feasibility

$$\mathbf{U}w^\dagger \geq d, \tag{6.25}$$

$$\begin{aligned} \mathbf{A}w^\dagger &= a, \\ \lambda_d^\dagger &\geq 0, \end{aligned} \tag{6.26}$$



and **stationarity**

$$0_n \in [\partial f(w^\dagger) + \partial(w \mapsto d - \mathbf{U}w)(w^\dagger) \cdot \lambda_d^\dagger + \partial(w \mapsto a - \mathbf{A}w)(w^\dagger) \cdot \lambda_a^\dagger]. \quad (6.27)$$

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 6.3.** [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 (6.22).

### Complementary Slackness

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

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

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

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

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 (6.24) we will show for all coordinates

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

It is well know that a concave functions  $g$  satisfies

$$g(x) - g(y) \geq \nabla g(x)^\top (x - y) \quad \text{for all } x, y. \quad (6.28)$$

But  $G$  is concave by the convexity of  $f^*$  (cf. Proposition 6.2).

First, we show

$$\nabla G_i(\lambda_d^\dagger) \leq 0 \quad \text{for all } i \in \{1, \dots, s\}. \quad (6.29)$$

Assume towards a contradiction that  $\nabla G_i(\lambda_d^\dagger) > 0$  for some  $i \in \{1, \dots, s\}$ . By the continuity of  $\nabla G$  there exists  $\varepsilon > 0$  such that  $\nabla G_i(\lambda_d^\dagger + e_i \cdot \varepsilon) > 0$ . It follows from (6.28)

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

which contradicts the optimality of  $\lambda_d^\dagger$  for (6.22). It follows (6.29).

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  $\nabla G$  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 (6.24), 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 (6.29),  $w^\dagger$  satisfies the inequality constraints in (6.21). 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 6.3 and the differentiability of  $g$  it holds

$$a \in \partial g(\lambda_a^\dagger) = \left\{ \mathbf{A} \cdot \nabla f^* \left( \mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \right\} = \left\{ \mathbf{A} w^\dagger \right\}. \quad (6.30)$$

Thus  $a = \mathbf{A} w^\dagger$ . But then  $w^\dagger$  satisfies also the equality constraints. We proved (6.25).

### Stationarity

First we show

$$\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \in \partial f(w^\dagger). \quad (6.31)$$

By Proposition 6.3 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 (6.31). Next we show

$$-\mathbf{U}^\top \in \partial(w \mapsto d - \mathbf{U}w)(w^\dagger) \quad \text{and} \quad -\mathbf{A}^\top \in \partial(w \mapsto d - \mathbf{A}w)(w^\dagger). \quad (6.32)$$

To this end, note that

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

Thus  $-\mathbf{U}^\top \in \partial(w \mapsto d - \mathbf{U}w)(w^\dagger)$ . In the same way it follows  $-\mathbf{A}^\top \in \partial(w \mapsto d - \mathbf{A}w)(w^\dagger)$ .

From (6.31) and (6.32) we conclude

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

We have proved (6.27), that is, stationarity.

### Dual Feasibility and Conclusion

Dual feasibility (6.26) follows immediately from the optimality of  $\lambda_d^\dagger$  for (6.22). Thus,  $(\lambda_d^\dagger, \lambda_a^\dagger)$  and  $w^\dagger$  satisfy the Karush-Kuhn-Tucker conditions for (6.21). Applying [Roc70, Theorem 28.3] finishes the proof.  $\square$

**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.
- [Hai12] 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.
- [HJ05] Bang H and Robins Jm. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), December 2005.
- [IR14] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:243–263, 2014.
- [Kle20] Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer International Publishing, Cham, 2020.
- [KS07] 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.
- [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.
- [Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [RR83] 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.

## Bibliography

- [Rub07] Donald B. Rubin. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, January 2007.
- [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.
- [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.
- [Zub15] 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.