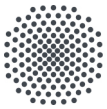
Title?

Universität Stuttgart



Universität Stuttgart

Ioan Scheffel

March 1, 2023

Contents

1 Introduction

How does action change an outcome? How should I guide my actions towards a better outcome? The first question is about causality, the second about ethics.

How do causality and ethics reflect on statistics? If you have not spent much time thinking about study design, this is a good way to start: As an analyst, ask yourself "Who acted? Who assigned treatments?" As researcher – plan your study accurately. You can ask yourself "How do we act? How do we assign treatment? Can we act?"

Let's say, you gather a sample from a study population, assign treatment (but forget how you did it). Some units get the drug, others don't. Then the statistical analysis shows a strong correlation of treatment and outcome. You hurry to your supervisor. "How was treatment assigned", asks she. "I forgot", says you. "How do you know your analysis is correct then?" You show her the data and together find out, that all units that received treatment were significantly taller than the rest of the sample. After all, is the drug or the height responsible for the change in outcome? You realise, that the data is worthless for answering this question. But you are lucky: It is just grass and fertiliser you were studying.

You get a second chance. A new medication needs testing before it enters the market. A company shall recruit participants, but the board requires you to write an outline for the study. You carefully explain steps to minimize risks for participants. You include plans to meet other requirements of human research. Then you have to decide how to assign treatment. No hand waving this time. You talk to your supervisor. "Last time, too many tall blades received fertiliser. The distribution of treatment was not really random..." You decide to determine treatment status by the flip a fair coin. You call the procedure 'randomization'.

Would you smoke if a coin tells you to? If you say yes - you are likely smoke anyway. The point is that forcing someone to smoke is unethical. But so is not studying the risks of smoking.

A professor is curious if the smoking habits of his students affect their grades. He observes the smoking area through his field glasses. His assistant gets to know his plans. He warns him. "Many students attend parties the night before exams. Maybe they are also more likely to smoke." "I shall see this for myself..." says the professor. He puts

away the field glasses. After a while, he visits the local club. He talks to a few of his students. Some smoke, some don't. The chats are enjoyable. He thinks: "Some of best students celebrate **before** the exam."

I hope, by now it's clear that sometimes it's all about how treatment was assigned. The probability of treatment given individual characteristics was introduced as the **propensity score** [RR83]. In the second example, where you flip a fair coin to decide treatment status, the propensity score is $1/2$. The coin ignores everything. What is the propensity score in the other examples? I admit, I don't know. It may vary. But we can see tendencies.

The propensity score is a simple concept that works well with potential outcomes. They are potential, because you only get the chance to observe on of them. If you treat, it's the potential outcome under treatment. On a high-level: If you act, you can't observe at the same time the effect of no action. Thus one of the potential outcomes always remains potential. Of course there are tricks. You can wait for the effect of an action to vanish and then observe the outcome again. This works well when the effect of an action is short term.

If treatment assignment is random we actually observe the potential outcome under treatment. This is because treatment assignment carries no more information. The coin ignores it. But we saw, that assignment often contains more information. Then it is not clear, if the effect on the outcome comes from the new information or the treatment. Then we don't even observe the potential outcome under treatment, but a confounded version. A simple idea to obtain information about the true potential outcome is to weight with the inverse probability of treatment, that is, 1 divided by the propensity score. Let's introduce some notation to be more precise.

Let $T \in \{0, 1\}$ be the **indicator of treatment**. This is a random variable. Let $X \in \mathcal{X}$ be a vector with individual characteristics. We call this the **covariate vector**. It is also a random variable. Last, let $(Y(0), Y(1))$ be the potential outcomes, that is, $Y(0)$ is the potential outcome without treatment and $Y(1)$ the potential outcome with treatment. They are random variables. We define the propensity score with individual characteristics x to be $\pi(x) := \mathbf{P}[T = 1|X = x]$. We observe

$$\text{either} \qquad Y(0)|T = 0 \qquad \text{or} \qquad Y(1)|T = 1\,. \tag{1.1}$$

We saw, that $Y(t)|T = t$ does not have the same distribution as $Y(t)$ for $t \in \{0, 1\}$. We can show, that if

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

holds we get

$$\mathbf{E}\left[\frac{T}{\pi(X)}Y(T)\right] = \mathbf{E}\left[Y(1)\right] . \tag{1.3}$$

That is, by weighting the observed outcome under treatment with the inverse propensity score we recover (in expectation) the potential outcome with treatment.

If the propensity score is unknown, one method is to use estimates of it. We hope to recover (1.3) from the estimate. But we have to be careful. After all, we want to extract informations on the potential outcomes and the propensity score is just a tool.

This is, why people started thinking about alternative ways to generate weights.

One way is to solve a constrained optimization problem.

2 Entropy Balancing Weights

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

$$Y(1)|T = 1 \nsim Y(1)\,, \tag{2.1}$$

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

$$(Y(0), Y(1)) \perp T \mid X\,, \tag{2.2}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} w_i Y_i\,. \tag{2.3}$$

The weights should satisfy (in a broader sense)

$$w_i \cdot Y_i \to Y(1) \qquad \text{for } N \to \infty\,. \tag{2.4}$$

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

My Contribution

I analyse the full optimization problem. In [WZ19] only the box constraints are considered. To eliminate the constraints on the dual variable of the first constraint in the primal optimization problem I need f^* to be strictly non-decreasing. This excludes the sample variance as an objective function, but the negative entropy still works. For technical reasons I change the box constraints. I discussed this change with the authors of [WZ19]. They approve it, because the method remains in tact. I consider a different regression basis. In [WZ19] they use sieve estimator [New97], whereas I chose the simpler partitioning estimate of [GKKW02]. The benefit of my method is, that I can work with a concrete oracle parameter. Also the basis of partitioning estimates forms a convex combination and is bounded. Thus I can avoid the use of matrix concentration inequalities as in [WZ19].

Problem 2.1.

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

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

$$[w_1,\dots,w_n]^\top \ \mapsto\ \sum_{i=1}^{n} f(w_i) \tag{2.5}$$

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

$$[\lambda_1, \ldots, \lambda_n]^\top \;\mapsto\; \sum_{i=1}^{n} f^*(\lambda_i) \tag{2.6}$$

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

$$f\colon [0,\infty) \to \mathbb{R}, \quad w \mapsto \begin{cases} 0 & \text{if } w = 0, \\ w \log w & \text{else.} \end{cases} \tag{2.7}$$

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

$$(f')^{-1} = \lambda \mapsto e^{\lambda - 1} \tag{2.8}$$

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

$$f\colon \mathbb{R} \to \mathbb{R}, \quad w \mapsto (w - 1/n)^2 \tag{2.9}$$

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

$$(f')^{-1} = \lambda \mapsto \frac{\lambda}{2} + \frac{1}{n} \tag{2.10}$$

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 partitioning estimates from [GKKW02]. Another angle would be sieve estimates [New97] where the number of basis functions can grow slower than N.

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.

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

In the sequel we mainly work with the dual problem.

My Contribution

I found important errors in the proof of a similar theorem in [WZ19]. After talking to the authors, I came up with a corrected proof. The Theorem has to be changed, but it becomes simpler. The key is to find the right matrix notation of the primal optimization problem. I adapted tools from convex analysis to make the proof work.

We introduce some more notation. Let $\mathbf{I}_n$ be the n-dimensional unit matrix, 0_n and 1_n the n-dimensional vectors containing only zeros or ones. Also we define the vector of basis functions of the covariates of unit $i \in \{1, \ldots, N\}$ to be

$$B(X_i) \;:=\; [B_1(X_i), \ldots, B_N(X_i)]^\top \in \mathbb{R}^N.$$

Let $\delta := [\delta_1, \ldots, \delta_N]^\top \in \mathbb{R}^N$ be the vector of upper bounds in the box constraints of Problem 2.1. Furthermore, we define the matrix of basis functions **for the treated** to be

$$\mathbf{B}(\mathbf{X}) \;:=\; \Big[B(X_1), \ldots, B(X_n)\Big] \in \mathbb{R}^{N \times n}.$$

Note, that these are random quantities and that the size of $\mathbf{B}(\mathbf{X})$ depends on the random size $n \in \mathbb{N}$ of the treatment group in the sample.

Theorem 2.1. *The dual of Problem 2.1 is the unconstrained optimization problem*

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

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 $B(X_i) = [B_1(X_i), \ldots, B_n(X_i)]^\top$ denotes the N basis functions of the covariates of unit $i \in \{1, \ldots, N\}$ and $|\lambda| = [|\lambda_1|, \ldots, |\lambda_N|]^\top$, 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 2.1 is given by

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

Lemma 2.1. *A matrix formulation of Problem 2.1 is*

$$\begin{aligned} &\underset{w \in \mathbb{R}^n}{\text{minimize}} && \varphi(w) \\ &\text{subject to} && \mathbf{U}w \; \geq \; d \, , \\ &&& \mathbf{A}w \; = \; a \, , \end{aligned} \tag{2.11}$$

with objective function

$$\varphi \; : \; \mathbb{R}^n \to \; \overline{\mathbb{R}} \, , \qquad [w_1, \ldots, w_n]^\top \; \mapsto \; \sum_{i=1}^{n} f(w_i) \, ,$$

inequality matrix and vector

$$\mathbf{U} \; := \; \begin{bmatrix} \mathbf{I}_n \\ \pm \, \mathbf{B}(\mathbf{X}) \end{bmatrix} \in \mathbb{R}^{(n+2N) \times n} \qquad d \; := \; \begin{bmatrix} 0_n \\ -N\delta \; \pm \; \sum_{i=1}^{N} B(X_i) \end{bmatrix} \in \mathbb{R}^{(n+2N)} \, ,$$

and equality matrix and vector

$$\mathbf{A} \; := \; 1_n^\top \in \mathbb{R}^{1 \times n} \qquad\qquad a \; := \; N \in \mathbb{N} \, .$$

Proof. Recall the box constraints of Problem 2.1.

$$\left| \frac{1}{N} \left(\sum_{i=1}^{n} w_i B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) \right| \; \leq \; \delta_k \qquad \text{for all } k \in \{1, \ldots, N\} \, .$$

A different notation is

$$\sum_{i=1}^{n} w_i B_k(X_i) \;\leq\; N\delta_k + \sum_{i=1}^{N} B_k(X_i)$$

and

$$\sum_{i=1}^{n} w_i B_k(X_i) \;\geq\; -N\delta_k + \sum_{i=1}^{N} B_k(X_i)$$

for all $k \in \{1, \ldots, N\}$. But this is $\pm\mathbf{B}(\mathbf{X})w \geq d$. Proving the rest of the statements is easy. We omit the details. □

Remark. The inequality constraints of Lemma 2.1 differ from its counterpart [WZ19, Proof of Lemma 1]. We don't transform the variable w, but shift to d what prevents us from keeping w. Note, that the choice of [WZ19, Proof of Lemma 1] leads to a mistake on page 21. The mistake is most obvious in the second display, where the first implication follows from dividing by 0. I discussed this with the authors and proposed a version of Lemma2.1 to solve the problem. I think it's best not to transform variables, because the mistake comes from (wrongly) calculating the convex conjugate of the (more complicated) transformed version of the objective function. The subsequent analysis even simplifies with my version. ♢

Lemma 2.2. *Let* $\rho \in \mathbb{R}^n$ *and* $\lambda^+, \lambda^- \in \mathbb{R}^N$ *. For*

$$\lambda_d \;:=\; \begin{bmatrix} \rho \\ \lambda^+ \\ \lambda^- \end{bmatrix} \qquad \textit{and} \qquad \lambda_a \;:=\; \lambda_0 \in \mathbb{R} \tag{2.12}$$

the dual optimization problem in the spirit of Theorem 4.4 of the matrix formulation in Lemma 2.1 is

$$\begin{aligned} &\underset{\rho,\lambda^+,\lambda^-,\lambda_0}{\text{maximize}} && G(\rho, \lambda^+, \lambda^-, \lambda_0) \\ &\text{subject to} && \rho, \lambda^+, \lambda^- \;\geq\; 0\,, \end{aligned}$$

where

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

Proof. We show in Example? the convex conjugate relationship

$$\varphi \;:\; \mathbb{R}^n \to\; \overline{\mathbb{R}}\,, \qquad [w_1,\ldots,w_n]^\top \;\mapsto\; \sum_{i=1}^{n} f(w_i)\,,$$
$$\varphi^* \;:\; \mathbb{R}^n \to\; \overline{\mathbb{R}}\,, \qquad [w_1^*,\ldots,w_n^*]^\top \;\mapsto\; \sum_{i=1}^{n} f^*(w_i^*)\,.$$

The rest of the proof consists of elementary computations. We omit the details. □

Proof. *(Theorem 2.1)* We eliminate the constraints in the dual problem of Lemma 2.2. Since we assume f^* to be strictly non-decreasing and $\rho \geq 0$, it follows that $\rho = 0_n$ is optimal. Thus, we consider the updated objective function G, that is,

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

To eliminate the remaining constraints, we paraphrase [WZ19, pages 19-20]. We show for all $i \in \{1,\ldots,N\}$

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

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

$$\tilde{\lambda} \;:=\; \left[\; \lambda_1^+,\ldots,\; \lambda_i^+-(\lambda_i^+\wedge\lambda_i^-)\,,\; \ldots,\lambda_N^+,\; \lambda_1^-,\ldots,\lambda_i^--(\lambda_i^+\wedge\lambda_i^-)\,,\; \ldots,\lambda_N^-\,,\lambda_0\right]^\top. \tag{2.13}$$

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

$$G(\tilde{\lambda},\lambda_0) - G(\lambda,\lambda_0) \;=\; 2N\cdot\delta_i\cdot(\lambda_i^+\wedge\lambda_i^-) \;>\; 0\,, \tag{2.14}$$

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

objective function one more time to get

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

Multiplying G with $-1/N$ and introducing the indicator of treatment T to fill up the entries for $i > n$, the final (unconstrained) optimization problem reads

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

Connecting Problem 2.1 via Lemma 2.1 to Lemma 2.2 and applying Theorem 4.4 we derive the last statement of Theorem 2.1. This finishes the proof. □

3 Asymptotic Analysis

3.1 Consistency of Optimal Solutions

3.1.1 Estimate of an Oracle Parameter by the Dual

My Contribution

I found out, that consistency for the dual variable is enough to prove later results. This simplifies the proof. In [WZ19] they use a quadratic Taylor expansion to obtain learning rates. I found out, that a simpler mean value result for differentiable convex functions is sufficient to proof consistency. Since I work with partitioning estimates, I found a suitable oracle parameter. I prove an (extended) lemma which is central but the details were omited.

Throughout this section we assume for all $N \in \mathbb{N}$ the existence of an optimal solution $(\lambda_0^\dagger, \lambda^\dagger)$ to Problem? We define the oracle parameter $\lambda^* \in \mathbb{R}^N$ to be the vector with coordinates

$$\lambda_k^* \;:=\; f'\left(\frac{1}{\pi(X_k)}\right) \;-\; \lambda_0^\dagger \qquad \text{for all } k \in \{1,\dots,N\}\;, \tag{3.1}$$

where $\pi(x) = \mathbf{P}[T=1|X=x]$ is the **propensity score** at $x \in \mathcal{X}$. Why this choice? First, the $\lambda_0^\dagger$ part is unimportant. We need it to eliminate the same factor in

$$w(x) := (f')^{-1}\left(\langle B(x), \lambda^\dagger\rangle + \lambda_0^\dagger\right) \tag{3.2}$$

that is

$$\langle B(x), \lambda^*\rangle + \lambda_0^\dagger = \sum_{k=1}^{N} B_k(x) f'\left(\frac{1}{\pi(X_k)}\right)\;. \tag{3.3}$$

The other part is the foundation of why everything works. We will show

$$\left|\sum_{k=1}^{N} B_k(X_i) f'\left(\frac{1}{\pi(X_k)}\right) - f'\left(\frac{1}{\pi(X_k)}\right)\right| \;\le\; \omega\left(f' \circ (x \mapsto 1/x) \circ \pi, h_N\right)\;. \tag{3.4}$$

Consequently, if π is continuous and positive (not 0) on $\mathcal{X}$ and the width of the partition h_N converges to 0, we get

$$\left|\langle B(X_i), \lambda^*\rangle + \lambda_0^\dagger - f'\left(\frac{1}{\pi(X_k)}\right)\right| \;\to\; 0 \qquad \text{almost surely.} \tag{3.5}$$

This helps proving

Theorem 3.1. *Let* $(\lambda_0^\dagger, \lambda^\dagger)$ *be an optimal solution to Problem? and define the oracle parameter* λ^* *as in* (3.1)*. Furthermore, assume that the propensity score function is continuous and positive on* $\mathcal{X}$*. Then* $\left\|\lambda^\dagger - \lambda^*\right\|_2 \xrightarrow{\mathbf{P}} 0$ *for* $N \to \infty$*.*

Proof. We use a hint from the last display of [WZ19, p.22]. The high-level idea is that the connection of optimality of $(\lambda_0^\dagger, \lambda^\dagger)$ to proximity to (any) oracle parameter λ^* is due to convexity and differentiability of (parts of) the the objective function of Problem. We deliver the omitted technical details. I proved the following lemma by myself.

Lemma 3.1. *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* $\varepsilon > 0$

$$\inf_{\|\Delta\|=\varepsilon} g(y+\Delta) - g(y) \geq 0 \tag{3.6}$$

implies the existence of a global minimum $y^* \in \mathbb{R}^m$ *of* g *satisfying* $\|y^* - y\|_2 \leq \varepsilon$*.*

Proof. Let B be the euclidian ball in $\mathbb{R}^m$. 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 (3.6) is true. Then it holds

$$g(x) < g(y^*) \quad \text{for some } x \in \mathbb{R}^m \setminus (y + \varepsilon B)\,. \tag{3.7}$$

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 $\|\Delta_x\|_2 = \varepsilon$ such that

$$\theta x + (1-\theta)y^* = y + \Delta_x\,. \tag{3.8}$$

It follows

$$\begin{aligned} g(y^*) \leq g(y) \leq g(y+\Delta_x) &= g(\theta x + (1-\theta)y^*) \\ &\leq \theta g(x) + (1-\theta)g(y^*) < g(y^*), \end{aligned} \tag{3.9}$$

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 (3.6) being true, the equality is due to (3.8), the third inequality is due to the convexity of g and the strict inequality is due to (3.7). Thus every local minimum of g in $y + \varepsilon B$ is also a global minimum. □

Since $(\lambda_0^\dagger, \lambda^\dagger)$ is a global minimum (in $\mathbb{R}^{N+1}$) of the objective function G of Problem?, that is,

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

an immediate consequence of Lemma 3.1 is

$$\mathbf{P}\left[\left\|\lambda^\dagger - \lambda^*\right\|_2 \le \varepsilon\right] \;\ge\; \mathbf{P}\left[\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \ge 0\right].$$

To prove Theorem 3.1 it thus suffices to prove

Lemma 3.2. *Under the conditions of Theorem 3.1 it holds for all* $\varepsilon > 0$

$$\mathbf{P}\left[\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \ge 0\right] \;\to\; 1 \qquad \textit{for } N \to \infty\,. \tag{3.10}$$

Proof. Recall the objective function G of Problem?

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

Since we assume the convex conjugate f^* to be differentiable (it always convex), without the last term, G would be a differentiable convex function.

It is well know that a differentiable convex functions g satisfies

$$g(x) - g(y) \ge \nabla g(y)^\top (x - y) \qquad \text{for all } x, y\,. \tag{3.11}$$

The gradient of

$$g := (\lambda, \lambda_0) \mapsto \frac{1}{N}\sum_{i=1}^{N} [\, T_i \cdot f^*(\lambda_0 + \langle B(X_i), \lambda\rangle) \;-\; (\lambda_0 + \langle B(X_i), \lambda\rangle)\,] \tag{3.12}$$

is

$$\nabla g = (\lambda, \lambda_0) \mapsto \frac{1}{N}\sum_{i=1}^{N} \Big[\, T_i \cdot (f')^{-1}(\lambda_0 + \langle B(X_i), \lambda\rangle) \;-\; 1\Big]\, [B(X_i)^\top, 1]^\top \tag{3.13}$$

Thus

$$\begin{aligned} &G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) \;-\; G(\lambda^*, \lambda_0^\dagger) \\ &\quad\ge -\frac{1}{N}\sum_{i=1}^{N} \Big[B(X_i)^\top, 1\Big] \cdot \begin{bmatrix}\Delta \\ \Delta_0\end{bmatrix} \Big(1 \;-\; T_i \cdot (f')^{-1}\Big(\langle B(X_i), \lambda^*\rangle + \lambda_0^\dagger\Big)\Big) \\ &\qquad +\; \langle \delta, |\lambda^* + \Delta| - |\lambda^*|\rangle\,. \end{aligned} \tag{3.14}$$

Next, we fix $\tilde{\varepsilon} > 0$ and establish in (3.14) the lower bound $-\tilde{\varepsilon}$ with probability going to 1 as $N \to \infty$. Then we conclude that this holds for all $\tilde{\varepsilon} > 0$. The measurability of $G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger)$ will give us the lower bound 0 in (3.14) with probability going to 1.

In (3.14) we control the **first term** by (what?) and the **second term** by $\|\delta\|_1$.

First Term

We note, that by $\|B(x)\|_2 \leq 1$ for all $x \in \mathcal{X}$ and the Cauchy-Schwarz inequality it holds

$$\left[B(X_i)^\top, 1\right] \cdot \begin{bmatrix} \Delta \\ \Delta_0 \end{bmatrix} \lesssim \|(\Delta, \Delta_0)\| = \varepsilon\,. \tag{3.15}$$

Next, we see that

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N}\left(1 - T_i \cdot (f')^{-1}\left(\langle B(X_i), \lambda^*\rangle + \lambda_0^\dagger\right)\right) \\ &\lesssim \frac{1}{N}\sum_{i=1}^{N}\left|1 - \frac{T_i}{\pi(X_i)}\right| + \frac{1}{N}\sum_{i=1}^{N}\left|\langle B(X_i), \lambda^*\rangle + \lambda_0^\dagger - f'\left(\frac{1}{\pi(X_i)}\right)\right| \\ &=: S_N + M_N\,. \end{aligned} \tag{3.16}$$

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

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

Also

$$\mathbf{E}\left[\left|1 - \frac{T}{\pi(X)}\right|\right] \leq 1 + \mathbf{E}\left[\frac{T}{\pi(X)}\right] = 2\,. \tag{3.17}$$

Thus Etemadi's ($\mathcal{L}_1$ version) strong law of large numbers (cf. [Kle20, Theorem 5.17]) applies to S_N, that is, $S_N \leq \tilde{\varepsilon}/(4\varepsilon)$ with probability going to 1. Next, we bound M_N.

Recall that $\sum_{k=1}^{N} B_k(x) = 1$ for all $x \in \mathcal{X}$. Thus

$$\begin{aligned}
&\left| \langle B(X_i), \lambda^* \rangle \; + \; \lambda_0^\dagger \; - \; f'\left(\frac{1}{\pi(X_i)}\right) \right| \\
&= \left| \sum_{k=1}^{N} B_k(X_i) \left(f'\left(\frac{1}{\pi(X_k)}\right) - \; \lambda_0^\dagger \right) \; + \; \lambda_0^\dagger - f'\left(\frac{1}{\pi(X_i)}\right) \right| \\
&= \left| \sum_{k=1}^{N} B_k(X_i) \left(f'\left(\frac{1}{\pi(X_k)}\right) - f'\left(\frac{1}{\pi(X_i)}\right) \right) \right| \\
&\leq \; \sum_{k=1}^{N} \frac{1_{\{X_k \in A_N(X_i)\}}}{\sum_{j=1}^{N} 1_{\{X_j \in A_N(X_i)\}}} \left| \left(f'\left(\frac{1}{\pi(X_k)}\right) - f'\left(\frac{1}{\pi(X_i)}\right) \right) \right| \\
&\leq \; \omega\left(f' \circ (x \mapsto 1/x) \circ \pi, h_N \right) \to 0 \qquad \text{almost surely}\,,
\end{aligned}$$

where ω is the modulus of continuity. The convergence to 0 is due to f' being continuous, $\pi(x) \in (0,1)$ for all $x \in \mathcal{X}$ and the (assumed) continuity of π. Indeed, by $h_N \to 0$ for $N \to \infty$ it follows $\omega\left(f' \circ (x \mapsto 1/x) \circ \pi, h_N \right) \to 0$. Note, that this works because of the partitioning. We conclude, that $M_N \leq \tilde{\varepsilon}/(4\varepsilon)$ with probability going to 1.

This establishes the desired bound of $\tilde{\varepsilon}/(2\varepsilon)$ in (3.16). Together with (3.15) we conclude that the **first term** in (3.14) 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

$$\begin{aligned}
&\langle \delta, |\lambda^* + \Delta| - |\lambda^*| \rangle \\
&\quad \geq \; -\langle \delta, |\Delta| \rangle \; \geq \; -\|\delta\|_1 \, \|\Delta\|_\infty \; \geq \; -\|\delta\|_1 \, \|(\Delta, \Delta_0)\|_2 \; \geq \; -\|\delta\|_1 \, \varepsilon \; \geq \; -\tilde{\varepsilon}/2\,,
\end{aligned}$$

with probability going to 1 as $N \to \infty$. The convergence is due to $\|\delta\|_1$ converging to 0 in probability.

Conclusion

With the analysis of the **first** and **second term** in (3.14) we conclude

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) \; - \; G(\lambda^*, \lambda_0^\dagger) \geq -\tilde{\varepsilon} \tag{3.18}$$

with probability going to 1 as $N \to \infty$. A closer look reveils the measurability of $G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger)$. Since this holds true for all $\tilde{\varepsilon} > 0$ we get

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0 \tag{3.19}$$

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

$$\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0 \tag{3.20}$$

with probability going to 1 as $N \to \infty$. We see, that this holds for all $\varepsilon > 0$. This finish the proof. □

□

3.1.2 Estimate of the Inverse Propensitiy Score by the Weights

The following theorem is an easy consequence of Theorem 3.1.

Theorem 3.2. *Consider the weights function defined by*

$$w(x) := (f')^{-1}\left(\langle B(x), \lambda^\dagger\rangle + \lambda_0^\dagger\right) \qquad \text{for all } x \in \mathcal{X}\,. \tag{3.21}$$

Under the conditions of Theorem 3.1 it holds $w(X) \xrightarrow{\mathbf{P}} 1/\pi(X)$

Proof. For all $\varepsilon > 0$ it holds

$$\begin{aligned}
\left|w(X) - \frac{1}{\pi(X)}\right| &= \left|(f')^{-1}\left(\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger\right) - \frac{1}{\pi(X)}\right| \\
&\lesssim \left|\langle B(X), \lambda^\dagger - \lambda^*\rangle\right| + \left|\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger - f'\left(\frac{1}{\pi(X)}\right)\right| \\
&\lesssim \left\|\lambda^\dagger - \lambda^*\right\|_2 + \left|\sum_{i=1}^N B_k(X) \cdot f'\left(\frac{1}{\pi_k}\right) - f'\left(\frac{1}{\pi(X)}\right)\right| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \leq \varepsilon\,,
\end{aligned} \tag{3.22}$$

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

Gaussian Bridge

My Contribution

I adapted the error decomposition of [WZ19] to estimates of the distribution function. I found out, that a simple switch from Y_i to $\mathbf{1}_{\{Y_i \leq z\}}$ does this. I learned from [vdV00]

about the functional delta method and application to plug in estimates with the distribution function. This motivated me to make the results uniform in $z \in \mathbb{R}$. I found out, that consistency is enough to bound R_3 and that R_2 can be bounded without concrete learning rates. This works because of the partitioning estimate and concrete oracle parameters and the first two constraints of the primal optimization problem.

Theorem 3.3. (Slutzky's theorem) *Let (E, d) be a metric space and let $X, X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables with values in E. Assume $X_n \to X$ in distribution and $d(X_n, Y_n) \to 0$ in probability. Then $Y_n \to X$ in distribution.*

Proof. [Kle20, Theorem 13.8] □

Theorem 3.4. *Under conditions the stochastic process*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{n} w_i^\dagger \mathbf{1}_{\{Y_i \leq z\}} - \mathbf{P}[Y(1) \leq z]\right)_{z \in \mathbb{R}}. \tag{3.23}$$

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 $F_{Y(1)}(z) = \mathbf{P}[Y(1) \leq z]$.

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

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

Analysis of R_1

By Theorem 2.1 it holds $w_i^\dagger = w(X_i)$ for $i \in \{1, \dots, n\}$, that is, for $i \leq n$ we can identify $w(X_i)$ with the optimal solution to problem 2.1. Thus the constraints of the problem apply.

$$\left| \frac{1}{N} \left(\sum_{i=1}^{n} w(X_i) B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) \right| \leq \delta_k \qquad \text{for all } k \in \{1, \dots, N\} . \tag{3.24}$$

Note, that the first sum goes over $\{1, \dots, n\}$ while the second sum goes over $\{1, \dots, N\}$. A second, equivalent version of the constraints is

$$\left| \frac{1}{N} \left(\sum_{i=1}^{N} T_i w(X_i) B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) \right| \leq \delta_k \qquad \text{for all } k \in \{1, \dots, N\} . \tag{3.25}$$

Now both sums go over $\{1, \dots, N\}$ and the indicator of treatment T_i takes care that in the first sum only the terms with $i \leq n$ are effective. Having this flexibility with the versions helps. I regard the first version as suitable for non-probabilistic computations, although n is of course a random variable. On the other hand, the second version is more honest, exactly telling the dependence on the indicator of treatment. This version is useful in probabilistic computations.

Let's bound R_1.

$$\begin{aligned} \sup_{z \in \mathbb{R}} |R_1(z)| &= \sup_{z \in \mathbb{R}} \left| \sqrt{N} \sum_{k=1}^{N} \left[\frac{1}{N} \left(\sum_{i=1}^{n} w(X_i) B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) F_{Y(1)}(z|X_k) \right] \right| \\ &\leq \sqrt{N} \sum_{k=1}^{N} \left| \frac{1}{N} \left(\sum_{i=1}^{n} w(X_i) B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) \right| \sup_{z \in \mathbb{R}} F_{Y(1)}(z|X_k) \\ &\leq \sqrt{N} \, \|\delta\|_1 \end{aligned} \tag{3.26}$$

Playing around with norm equivalences we discover that $\sqrt{N} \, \|\delta\|_1 \to 0$ for $N \to \infty$ is the weakest (natural) assumption to control R_1. Indeed, other ways to continue the second row in (3.26) are

$$(\cdots) \leq \sqrt{N} \, \|\delta\|_2 \left(\sum_{k=1}^{N} \left(\sup_{z \in \mathbb{R}} F_{Y(1)}(z|X_k) \right)^2 \right)^{1/2} \leq N \, \|\delta\|_2 ,$$

by the Cauchy-Schwarz inequality and $F_{Y(1)} \in [0, 1]$, or

$$(\cdots) \leq \sqrt{N} \, \|\delta\|_\infty \sum_{k=1}^{N} \sup_{z \in \mathbb{R}} F_{Y(1)}(z|X_k) \leq N^{3/2} \, \|\delta\|_\infty .$$

Since $\delta \in \mathbb{R}^N$, however, it holds

$$\sqrt{N}\,\|\delta\|_1 \;\leq\; N\,\|\delta\|_2 \;\leq\; N^{3/2}\,\|\delta\|_\infty\,.$$

With hindsight, the assumption $\sqrt{N}\,\|\delta\|_1 \to 0$ for $N \to \infty$ also suffices to control the second (or first) occurrence of a term, that we control by assumptions on δ. This is the **second term** of (3.14), where we estimate

$$\langle \delta, |\Delta| \rangle \;=\; \sum_{k=1}^{N} \delta_k\,|\Delta_k| \;\leq\; \|\delta\|_1\,\|\Delta\|_\infty \;\leq\; \|\delta\|_1\,\|\Delta\|_2 \;\leq\; \|\delta\|_1\,\varepsilon \;\to\; 0 \quad \text{for } N \to \infty\,.$$

Analysis of R_2

In the original paper [WZ19] the authors derive concrete learning rates for the weights and employ them in bounding this term. They obtain a multiplied learning rate, which is sufficiently fast. Their approach, however, calls for concrete learning rates of the weights. Arguably, the process of deriving such rates is the most complicated part of the paper. I found out, that we don't need concrete rates for the weights. Consistency of the weights is enough and gives us an (arbitrarily slow but sufficient) learning rate to establish the results. We don't even need rates for the weights to control R_2. They only play a role in bounding R_3. Nevertheless, we use the second constraint of Problem (2.1)

$$1 \;=\; \frac{1}{N}\sum_{i=1}^{n} w_i^\dagger \;=\; \frac{1}{N}\sum_{i=1}^{n} w(X_i) \;=\; \frac{1}{N}\sum_{i=1}^{N} T_i w(X_i)\,. \tag{3.27}$$

To this end, we note that

$$\begin{aligned}
&\sup_{z\in\mathbb{R}} \left| F_{Y(1)}(z|X_i) - \sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k) \right| \\
&\leq\; \sum_{k=1}^{N} \frac{\mathbf{1}_{\{X_k \in A_N(X_i)\}}}{\sum_{j=1}^{N} \mathbf{1}_{\{X_j \in A_N(X_i)\}}} \sup_{z\in\mathbb{R}} \left| F_{Y(1)}(z|X_i) - F_{Y(1)}(z|X_k) \right| \\
&\leq\; \sup_{z\in\mathbb{R}} \omega\left(F_{Y(1)}(z|\cdot), h_N \right)\,,
\end{aligned}$$

where ω is the modulus of continuity and h_N is the width of the partition $\mathcal{P}_N = \{A_{1,N}, A_{2,N}, \ldots\}$. There are many (more concrete, yet stronger) assumptions on the regularity of $F_{Y(1)}$ and the width of the partition h_N that give us

$$\sqrt{N} \sup_{z\in\mathbb{R}} \omega\left(F_{Y(1)}(z|\cdot), h_N \right) \to 0 \qquad \text{for } N \to \infty\,. \tag{3.28}$$

But we shall keep this more general (and abstract) assumption. We conclude

$$\begin{aligned}
&\sup_{z\in\mathbb{R}} |R_2(z)| \\
&\leq \sqrt{N} \sum_{i=1}^{N} \left[\frac{T_i \cdot w(X_i) - 1}{N} \sup_{z\in\mathbb{R}} \left| F_{Y(1)}(z|X_i) \; - \; \sum_{k=1}^{N} B_k(X_i) \cdot F_{Y(1)}(z|X_k) \right| \right] \\
&\leq \sqrt{N} \sup_{z\in\mathbb{R}} \omega\left(F_{Y(1)}(z|\cdot), h_N\right) \sum_{i=1}^{N} \frac{T_i \cdot w(X_i) + 1}{N} \\
&= 2\sqrt{N} \sup_{z\in\mathbb{R}} \omega\left(F_{Y(1)}(z|\cdot), h_N\right) \; \to \; 0\,.
\end{aligned}$$

Analysis of R_3

We will apply theory of empirical processes to bound

$$R_3(z) = \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} \left[T_i \left(w(X_i) - \frac{1}{\pi(X_i)} \right) \left(\mathbf{1}_{\{Y_i \leq z\}} - F_{Y(1)}(z|X_i) \right) \right] \right) \tag{3.29}$$

in probability. Why don't we use simple concentration inequalities such as Bernstein's or Markov's inequality? The reason is, that the weights $w(x) := (f')^{-1}\left(\langle B(x), \lambda^\dagger\rangle + \lambda_0^\dagger\right)$ depend (thorough B and $(\lambda^\dagger, \lambda_0^\dagger)$) on the whole data set $D := (T_i, X_i)_{i=1,\dots,N}$. Thus, it is more honest to write $w(x, D)$ instead. This captures the whole dependence on probabilities. Note, that $(Y_i)_{i=1,\dots N}$ are independent of w given D. A standard computation shows

$$\mathbf{E}\left[\frac{T}{\pi(X)} \left(\mathbf{1}_{\{Y(T)\leq z\}} - F_{Y(1)}(z|X) \right) \right] = 0\,. \tag{3.30}$$

Furthermore

$$\begin{aligned}
&\mathbf{E}\left[T w(X, D) \left(\mathbf{1}_{\{Y(T)\leq z\}} - F_{Y(1)}(z|X) \right) \right] \\
&= \mathbf{E}\left[\mathbf{E}\left[w(X, D) \left(\mathbf{1}_{\{Y(1)\leq z\}} - F_{Y(1)}(z|X) \right) | T = 1, X, D \right] \right] \\
&= \mathbf{E}\left[w(X, D) \mathbf{E}\left[\mathbf{1}_{\{Y(1)\leq z\}} - F_{Y(1)}(z|X) | X, D \right] \right] \\
&= \mathbf{E}\left[w(X, D) \mathbf{E}\left[\mathbf{1}_{\{Y(1)\leq z\}} - F_{Y(1)}(z|X) | X \right] \right] \\
&= 0
\end{aligned}$$

The second equality is due to the assumption of $(Y(0), Y(1)) \perp T|X$. The third equality is due to $X \perp D$. Thus

$$R_3(z) = G_N f_D^z\,. \tag{3.31}$$

By the consistency of the weights there exists a learning rate (ε_N) such that

$$\mathbf{P}\left[\left|w(X,D) - \frac{1}{\pi(X)}\right| \leq \varepsilon_N\right] \to 1 \qquad \text{for } N \to \infty\,. \tag{3.32}$$

Let $\mathcal{F}_N := \varepsilon_N B_{\mathcal{F}}$. It holds

$$\mathbf{P}\left[f_D^z \in \mathcal{F}_N \;\forall\, z \in \mathbb{R}\right] = \mathbf{P}\left[\sup_{z\in\mathbb{R}} |f_D^z| \leq \varepsilon_N\right] \to 1 \tag{3.33}$$

Then the lemma applies?.

Lemma 3.3. *Consider a function class $\mathcal{F}$ with unit ball $B_{\mathcal{F}} := \{f \in \mathcal{F}\colon \|f\|_\infty \leq 1\}$. Let (ε_N) be a sequence converging to 0 and let $(\mathcal{F}_N) := (C \cdot \varepsilon_N \cdot B_{\mathcal{F}})$ denote the sequence of scaled unit balls in $\mathcal{F}$. Assume that there exists $k < 2$ such that the covering number of the unit ball in $\mathcal{F}$ satisfies*

$$\log N_{[]}\left(\varepsilon, B_{\mathcal{F}}, \mathrm{L}_2(\mathbf{P})\right) \;\lesssim\; \left(\frac{1}{\varepsilon}\right)^k \qquad \textit{for all } \varepsilon > 0\,. \tag{3.34}$$

Then it holds $\|G_N\|^*_{\mathcal{F}_N} \xrightarrow{\mathbf{P}} 0$ for $N \to \infty$.

Proof. By maximal inequalities it holds

$$\begin{aligned}
\mathbf{E}^*\left[\|G_N\|_{\mathcal{F}_N}\right] &\lesssim J_{[]}\left(\varepsilon_N, \mathcal{F}_N, \mathrm{L}_2(\mathbf{P})\right) \\
&= \int_0^{\varepsilon_N} \sqrt{\log N_{[]}\left(\varepsilon, \mathcal{F}_N, \mathrm{L}_2(\mathbf{P})\right)}\, d\varepsilon \\
&= \int_0^{\varepsilon_N} \sqrt{\log N_{[]}\left(\varepsilon/(C\cdot\varepsilon_N), B_{\mathcal{F}}, \mathrm{L}_2(\mathbf{P})\right)}\, d\varepsilon \\
&\lesssim \int_0^{\varepsilon_N} \left(\frac{\varepsilon_N}{\varepsilon}\right)^{k/2} d\varepsilon \\
&= \varepsilon_N^{k/2} \frac{1}{1-k/2} \varepsilon_N^{1-k/2} \\
&\lesssim \varepsilon_N \\
&\to 0 \qquad \text{for } N \to \infty\,.
\end{aligned}$$

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

$$\mathbf{P}\left[\|G_N\|^*_{\mathcal{F}_N} \geq \varepsilon\right] \leq \varepsilon^{-1}\, \mathbf{E}^*\left[\|G_N\|_{\mathcal{F}_N}\right] \;\to\; 0 \qquad \text{for } N \to \infty\,.$$

□

Lemma 3.4. *Consider the (random) function* f_D^z *given by*

$$f_D^z(T, X, Y(T)) := T\left(w(D, X) - \frac{1}{\pi(X)}\right)\left(\mathbf{1}_{\{Y(T)\leq z\}} - F_{Y(1)}(z|X)\right). \tag{3.35}$$

Assume that there exists a function class $\mathcal{F}$ *satisfying the requirements of Lemma 3.3 and that* $f_D^z \in \mathcal{F}$ *for all* $z \in \mathbb{R}$ *almost surely. It then holds* $\sup_{z\in\mathbb{R}} |G_N f_D^z| \xrightarrow{\mathbf{P}} 0$ *for* $N \to \infty$.

Proof. By the consistency of the weights there exists a learning rate (ε_N) such that

$$\mathbf{P}\left[\left|w(X, D) - \frac{1}{\pi(X)}\right| \leq \varepsilon_N\right] \to 1 \qquad \text{for } N \to \infty. \tag{3.36}$$

Let $\mathcal{F}_N := \varepsilon_N B_{\mathcal{F}}$ as in Lemma 3.3. It holds

$$\sup_{z\in\mathbb{R}} |f_D^z| \lesssim \left|w(X, D) - \frac{1}{\pi(X)}\right| \leq \varepsilon_N \tag{3.37}$$

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

$$\mathbf{P}\left[f_D^z \in \mathcal{F}_N \ \forall\, z \in \mathbb{R}\right] = \mathbf{P}\left[\sup_{z\in\mathbb{R}} |f_D^z| \lesssim \varepsilon_N\right] \to 1 \qquad \text{as } N \to \infty. \tag{3.38}$$

Then it holds for all $\varepsilon > 0$

$$\begin{aligned}
\mathbf{P}\left[\sup_{z\in\mathbb{R}} |G_N f_D^z| \leq \varepsilon\right] &\geq \mathbf{P}\left[\sup_{z\in\mathbb{R}} |G_N f_D^z| \leq \|G_N\|^*_{\mathcal{F}_N} \leq \varepsilon\right] \\
&\geq \mathbf{P}\left[f_D^z \in \mathcal{F}_N \ \forall\, z \in \mathbb{R} \text{ and } \|G_N\|^*_{\mathcal{F}_N} \leq \varepsilon\right] \\
&\geq \mathbf{P}\left[f_D^z \in \mathcal{F}_N \ \forall\, z \in \mathbb{R}\right] - \mathbf{P}\left[\|G_N\|^*_{\mathcal{F}_N} \geq \varepsilon\right] \\
&\to 1.
\end{aligned}$$

The convergence of the second term is due to Lemma 3.3.

□

Since $\mathbf{E}[f_D^z(T, X, Y(T))] = 0$ for all $z \in \mathbb{R}$. We conclude $\sup_{z\in\mathbb{R}} |R_3(z)| \xrightarrow{\mathbf{P}} 0$.

Analysis of R_4

To bound this term we adapt [vdV00, Example 19.6]. To this end, let $\varepsilon > 0$ and $m \in \mathbb{N}$. We can choose

$$-\infty = z_0 < z_1 < \cdots < z_{m-1} < z_m = \infty \tag{3.39}$$

such that

$$\mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\right] \leq \varepsilon \qquad \text{for all } l \in \{1, \ldots, m\} \tag{3.40}$$

and $m \leq 2/\varepsilon$. We define brackets by

$$\underline{f_l}(T,X,Y(T)) := \frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T)\leq z_{l-1}\}} - F_{Y(1)}(z_l|X)\right) + F_{Y(1)}(z_{l-1}|X) - F_{Y(1)}(z_l) \tag{3.41}$$

$$\overline{f_l}(T,X,Y(T)) := \frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T)\leq z_{l}\}} - F_{Y(1)}(z_{l-1}|X)\right) + F_{Y(1)}(z_l|X) - F_{Y(1)}(z_{l-1}) \tag{3.42}$$

Let $\mathcal{G}$ be the space covered by these brackets. An elementary but lengthy calculation shows

$$\left\|\overline{f_l} - \underline{f_l}\right\|_{\mathrm{L}_2(\mathbf{P})} \;\leq\; 2\,\varepsilon^{1/4}\sqrt{\|1/\pi(X)\|_{\mathrm{L}_2(\mathbf{P})} + 2} \tag{3.43}$$

Thus we need $1/\pi(X) \in \mathrm{L}_2(\mathbf{P})$. By this assumption it follows, that

$$N_{[]}(C\varepsilon^{1/4}, \mathcal{G}, \mathrm{L}_2(\mathbf{P})) \;\leq\; \frac{2}{\varepsilon} \tag{3.44}$$

and thus

$$N_{[]}(\varepsilon, \mathcal{G}, \mathrm{L}_2(\mathbf{P})) \;\lesssim\; \frac{1}{\varepsilon^4}\,. \tag{3.45}$$

This covering number is of polynomial order. Thus

$$\log N_{[]}(\varepsilon, \mathcal{G}, \mathrm{L}_2(\mathbf{P})) \lesssim \log(1/\varepsilon)\,. \tag{3.46}$$

But then $\mathcal{G}$ is $\mathbf{P}$-Donsker. Define

$$g^z(T,X,Y(T)) := \frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T)\leq z\}} - F_{Y(1)}(z|X)\right) + F_{Y(1)}(z|X) - F_{Y(1)}(z) \tag{3.47}$$

Then $g^z \in \mathcal{G}$ for all $z \in \mathbb{R}$. Since $\mathbf{E}[g^z] = 0$ it holds

$$R_4(z) = G_N g^z \tag{3.48}$$

By the Donsker Theorem [vdV00, Theorem 19.5] the process R_4 converges in $l^\infty(\mathbb{R})$ to a Gaussian process, called $\mathbf{P}$-Brownian bridge, with mean 0 and covariance?

Conclusion

Computing the covariance and applying Slutzky's Theorem we finish the proof. □

3.2 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 .

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

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** $\operatorname{aff}(\Omega)$ of a set $\Omega \subseteq \mathbb{R}^n$ is the smallest affine set that includes Ω. We define the **relative interior** $\operatorname{ri}\Omega$ of a set $\Omega \subseteq \mathbb{R}^n$ to be the interior relative to the affine hull, that is,

$$\operatorname{ri}(\Omega) := \{x \in \Omega \mid \exists\, \varepsilon > 0 : (x + \varepsilon B_{\mathbb{R}^n}) \cap \operatorname{aff}(\Omega) \subset \Omega\}. \tag{4.1}$$

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] □

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

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) $\operatorname{ri}(C) = \{z \in C$: *for all* $x \in C$ *there exists* $t > 0$ *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 $(z_1, z_2) \in \mathrm{ri}(C_1 \times C_2)$. 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 $(z_1, z_2) \in \mathrm{ri}(C_1 \times C_2)$. If $t_1 > t_2$ similar arguments lead to the same result. We have proven $\mathrm{ri}(C_1 \times C_2) \supseteq \mathrm{ri}\, C_1 \times \mathrm{ri}\, C_2$ and equality. □

Functions

A function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ 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

$$f(\theta x + (1-\theta)y) \;\leq\; \theta f(x) + (1-\theta) f(y) \qquad \text{for all } x, y \in \mathbb{R}^n \text{ and all } \theta \in [0,1]\,. \tag{4.5}$$

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

$$f\left(\sum_{i=1}^m \theta_i x_i\right) \;\leq\; \sum_{i=1}^m \theta_i f(x_i) \qquad \text{for all } x_1, \ldots, x_m \in \mathbb{R}^n\,. \tag{4.6}$$

We call a function **strictly convex** if the inequality in (4.5) is strict.

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

$$\mathrm{dom}\, f \;:=\; \{x \in \mathbb{R}^n : f(x) < \infty\}\,. \tag{4.7}$$

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

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

$$\langle x^*, x - \overline{x} \rangle \leq f(x) - f(\overline{x})\,. \tag{4.8}$$

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

$$\langle \nabla f(\overline{x}), x - \overline{x} \rangle \leq f(x) - f(\overline{x})\,. \tag{4.9}$$

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

$$f(y) - f(x) \;\geq\; \langle \nabla f(x), y - x \rangle \;+\; \frac{m}{2} \|y - x\|^2 \,. \tag{4.10}$$

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

$$\nabla^2 f(x) - m \cdot \mathbf{I} \qquad \text{is positive semi-definite for all } \; x \in \operatorname{dom} f \,, \tag{4.11}$$

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** $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ of f defined by

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} \langle x^*, x \rangle - f(x) \,. \tag{4.12}$$

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.

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 4.1. (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*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \tag{4.13}$$

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***

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*) \tag{4.14}$$

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*

$$(f+g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{4.15}$$

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***

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{4.16}$$

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

$$S_{f,n} : \mathbb{R}^n \to \overline{\mathbb{R}}, \qquad (x_1, \ldots, x_n) \mapsto \sum_{i=1}^{n} f(x_i),$$
$$S_{f,n}^* : \mathbb{R}^n \to \overline{\mathbb{R}}, \qquad (x_1^*, \ldots, x_n^*) \mapsto \sum_{i=1}^{n} f^*(x_i^*)\,.$$

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 $i = 1, \ldots, n$, this is

$$p_i : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto x_i\,. \tag{4.17}$$

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

$$p_i^* : \mathbb{R} \to \mathbb{R}^n, \quad x \mapsto e_i \cdot x\,. \tag{4.18}$$

For the inverse image of the adjoint of p_i it holds

$$(p_i^*)^{-1}\{(x_1^*, \ldots, x_n^*)\} = \begin{cases} \{x_i^*\}\,, & \text{if } x_j^* = 0 \text{ for all } j \neq i\,, \\ \emptyset & \text{else.} \end{cases} \tag{4.19}$$

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

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

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

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

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

$$\left(S_{f,n} \circ p_{\{1,\ldots,n\}}\right)^* (x_1^*, \ldots, x_{n+1}^*) = \begin{cases} S_{f,n}^*(x_1^*, \ldots, x_n^*) & \text{if } x_{n+1}^* = 0\,, \\ \infty & \text{else,} \end{cases} \tag{4.20}$$

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

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

and

$$\operatorname{ri}\left(\operatorname{dom}\left(S_{f,n} \circ p_{\{1,\ldots,n\}}\right)\right) \;\cap\; \operatorname{ri}\left(\operatorname{dom} f_{n+1}\right)$$
$$= \operatorname{ri}\left(\operatorname{dom}\left(S_{f,n} \circ p_{\{1,\ldots,n\}}\right) \;\cap\; \operatorname{dom} f_{n+1}\right) \;=\; \operatorname{ri}\left(\bigcap_{i=1}^{n+1} \operatorname{dom} f_i\right) \;\neq\; \emptyset\,.$$

By the conjugate sum rule it follows

$$(S_{f,n+1})^* = (S_{f,n} \circ p_{\{1,\ldots,n\}} + f_{n+1})^* = (S_{f,n} \circ p_{\{1,\ldots,n\}})^* \square f_{n+1}^*$$
$$= S_{f,n}^* \circ p_{\{1,\ldots,n\}} + f_{n+1}^* = S_{f,n+1}^*\,.$$

$\diamondsuit$

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

My Contribution

I adapt ideas from [TB91] to take also equality constraints. For this, I had to understand the connection to my version of the primal optimization problem. I filled in many details that were omitted in the paper: I derived the Karush-Kuhn-Tucker conditions for the problem from the general result [Roc70, Theorem 28.3]. I prove in detail, that they hold for the adapted problem.

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 4.1).

Theorem 4.4. *Consider the optimization problem*

$$\begin{aligned} &\underset{w\in\mathbb{R}^n}{\text{minimize}} && f(w) \\ &\text{subject to} && \mathbf{U}w \;\geq\; d\,. \\ &&& \mathbf{A}w \;=\; a\,, \end{aligned} \tag{4.21}$$

and its dual problem

$$\begin{aligned} &\underset{\lambda_d\in\mathbb{R}^r,\lambda_a\in\mathbb{R}^s}{\text{maximize}} && \langle\lambda_d, d\rangle \;+\; \langle\lambda_a, a\rangle \;-\; f^*\Big(\mathbf{U}^\top\lambda_d + \mathbf{A}^\top\lambda_a\Big) \\ &\text{subject to} && \lambda_d \;\geq\; 0\,. \end{aligned} \tag{4.22}$$

Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ *be an optimal solution to* (4.22)*. If the objective function* f *of* (4.21) *is strictly convex and its convex conjugate* f^* *is continuously differentiable, then the unique optimal solution to* (4.21) *is given by*

$$w^\dagger = \nabla f^*\Big(\mathbf{U}^\top\lambda_d^\dagger + \mathbf{A}^\top\lambda_a^\dagger\Big)\,. \tag{4.23}$$

Plan of Proof

We show that $w^\dagger$ and $(\lambda_d^\dagger, \lambda_a^\dagger)$ meet the Karush-Kuhn-Tucker conditions for 4.21, that is, **complementary slackness**

$$\langle\lambda_d^\dagger, d - \mathbf{U}w^\dagger\rangle \;=\; 0\,, \tag{4.24}$$

primal and **dual feasibility**

$$\mathbf{U}w^\dagger \;\geq\; d\,, \tag{4.25}$$
$$\mathbf{A}w^\dagger \;=\; a\,,$$
$$\lambda_d^\dagger \;\geq\; 0\,, \tag{4.26}$$

and **stationarity**

$$0_n \;\in\; [\partial f(w^\dagger) \;+\; \partial\,(w \mapsto d - \mathbf{U}w)\,(w^\dagger)\cdot\lambda_d^\dagger \;+\; \partial\,(w \mapsto a - \mathbf{A}w)\,(w^\dagger)\cdot\lambda_a^\dagger\,]\,. \tag{4.27}$$

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 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 (4.22).

Complementary Slackness

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

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

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 (4.24) we will show for all coordinates

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

It is well know that a concave functions g satisfies

$$g(x) - g(y) \geq \nabla g(x)^\top(x-y) \qquad \text{for all } x, y\,. \tag{4.28}$$

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

First, we show

$$\nabla G_i(\lambda_d^\dagger) \;\leq\; 0 \qquad \text{for all } i \in \{1,\ldots,s\}\,. \tag{4.29}$$

Assume towards a contradiction that $\nabla G_i(\lambda_d^\dagger) > 0$ for some $i \in \{1, \dots, s\}$. 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 (4.28)

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

which contradicts the optimality of $\lambda_d^\dagger$ for (4.22). It follows (4.29).

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 ∇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 (4.24), 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 (4.29), $w^\dagger$ satisfies the inequality constraints in (4.21). 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 4.3 and the differentiability of g it holds

$$a \; \in \; \partial g(\lambda_a^\dagger) \; = \; \left\{ \mathbf{A} \cdot \nabla f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \right\} \; = \; \left\{ \mathbf{A} w^\dagger \right\}\,. \tag{4.30}$$

Thus $a = \mathbf{A} w^\dagger$. But then $w^\dagger$ satisfies also the equality constraints. We proved (4.25).

Stationarity

First we show

$$\mathbf{U}^\top \lambda_d^\dagger \; + \; \mathbf{A}^\top \lambda_a^\dagger \; \in \; \partial f(w^\dagger)\,. \tag{4.31}$$

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 (4.31). Next we show

$$-\mathbf{U}^\top \in\; \partial\,(w \mapsto d - \mathbf{U}w)\,(w^\dagger) \qquad \text{and} \qquad -\,\mathbf{A}^\top \in\; \partial\,(w \mapsto d - \mathbf{A}w)\,(w^\dagger)\,. \tag{4.32}$$

To this end, note that

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

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 (4.31) and (4.32) 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 (4.27), that is, stationarity.

Dual Feasibility and Conclusion

Dual feasibility (4.26) follows immediately from the optimality of $\lambda_d^\dagger$ for (4.22). Thus, $(\lambda_d^\dagger, \lambda_a^\dagger)$ and $w^\dagger$ satisfy the Karush-Kuhn-Tucker conditions for (4.21). 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.

[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.

[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.

[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.