Causal Inference in Observational Studies

Universität Stuttgart

Universität Stuttgart

Ioan Scheffel

March 15, 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 treatment?" 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 of a fair coin. You call the procedure 'randomization'.

Would you smoke if a coin tells you to? If you say yes - you 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 my best students celebrate before the exams."

I hope, by now it's clear that we should focus on treatment assignment. The propensity score [RR83], that is, the probability of treatment given (observed) individual characteristics, helps with that.

Theorem. [RR83, Theorem 1] *Observed individual characteristics are independent of treatment assignment given the propensity score.*

In the second example, where you flip a fair coin to assign treatment, the propensity score is $1/2$, despite variation across individual characteristics. The coin ignores everything. What is the propensity score in the other examples? I admit, I don't know. It varies, but we can see trends. In the first example, tall blades had a large propensity score. In the third example, the assistant thinks that students attending parties have a larger propensity score. This is not true, after all, but somehow the best students have a large propensity to celebrate before exams.

The propensity score is a simple concept that works well with potential outcomes. They are potential, because they exist (or we assume they exist) independent of our observation. They live in parallel universes. If we have a binary treatment, that is, you either treat or don't, there are two potential outcomes. One under treatment and one under no treatment. Ideally we would like to compare (for one unit) those two potential outcomes. But that is impossible. Instead people keep asking: "Had it been better if (20 years ago) I made a different decision?" You know what happened but don't know what would have happened. 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 (under similar conditions) again. This works well when the effect of an action is short term.

If the propensity score is known we actually observe one of the potential outcomes. 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 observed or unobserved individual characteristics or the treatment. Then we observe neither of the two potential outcomes, but a biased version. Why then bother?

A simple idea to obtain information about the true potential outcome from its biased version (we also say confounded version) is, to weight it 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**. Let $X \in \mathcal{X}$ be a vector with individual characteristics. We call this the **covariate vector**.

Furthermore, let $(Y(0), Y(1))$ be the **potential outcomes**, that is, $Y(0)$ is the potential outcome under no treatment and $Y(1)$ the potential outcome under treatment. All the quantities we introduce are random variables.

We define the propensity score π with individual characteristics x to be

$$\pi(x) := \mathbf{P}[T = 1 | X = x] \tag{1.1}$$

We observe

$$\text{either} \quad Y(0) \,|\, T = 0 \quad \text{or} \quad Y(1) \,|\, T = 1\,. \tag{1.2}$$

We show in Lemma 3.13, that if treatment assignment is **strongly ignorable** [RR83, (1.3)]

$$(Y(0), Y(1)) \perp T \,|\, X \quad \text{and} \quad 0 < \pi(X) < 1\,, \tag{1.3}$$

that is, potential outcomes are independent of treatment given covariates and every possible set of characteristic has a chance to receive treatment, we get

$$\mathbf{E}\left[\frac{T}{\pi(X)} Y(T)\right] = \mathbf{E}\left[Y(1)\right]\,. \tag{1.4}$$

That is, by weighting the observed outcome under treatment with the inverse propensity score we recover (in expectation) the potential outcome under treatment. This is relevant, because $Y(t)|T = t$ does not have the same distribution as $Y(t)$ for $t \in \{0,1\}$.

In observational studie, the propensity score is unknown. A very popular method is to use estimates of it to create weights. We hope to recover (1.4) from the estimate. In practice, estimating the propensity score is a difficult task. Researchers often compare estimates from different models. They employ metrics called covariate balance. The high level idea is, that the weighted treatment group should be similar to the unweighted control group.

Let's be more specific what we mean by covariate balance. We consider a class of functions B of the covariates, which we call regression basis. A simple example are the (known) moments of the covariats. We will extend the view, but for now this is sufficient. Our measure of balance is

$$\frac{1}{N}\left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i)\right) \tag{1.5}$$

where (w_i) are the (estimated) weights for a (sub-)population of size $n \le N$ and B are (basis-)functions of the covariates. Rather than estimating the propensity score

and then checking for covariate balance, the method of [IR14] tries to solve both tasks simultaneously. Therefore it is called the Covariate Balancing Propensity Score.

We will consider a third method, which only balances covariates but does not (explicitly) model treatment or outcome. This method was introduced in [Hai12]. It is a convex optimization problem with constraints on the balance of moments of the covariates. The method gained popularity by the observation of [ZP17] that it is doubly robust. Graphic?

We consider (1.6) for all basis functions is the regression basis. It is a non-trivial question, which basis to choose in practice. How strictly to enforce covariate balance is another question. It is relevant, because very strict assumptions can render the problem infeasible, whereas loosening can result in bias of the estimator. In [Hai12] they choose the (known) moments of the covariate as basis and enforce strict balance, that is, the quantity in (1.6) has to vanish. In [WZ19] they consider the regression basis of sieve estimators [New97], where the number of basis functions grows with the sample size. Also they loosen the strict constraints on the covariate balance as to vanish only for $N \to \infty$. This takes the for

$$\left| \frac{1}{N} \left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i) \right) \right| \le \delta_B \tag{1.6}$$

for some $\delta_B > 0$ with $\delta_B \to 0$ for $N \to \infty$. The paper [WZ19] also contains theoretical analysis. It shows a surprising connection to propensity score estimation. They show that with the regression basis of sieve estimators their method (implicitly) models the inverse propensity score. Their analysis is in part motivated by the observation, that entropy balancing is doubly robust [ZP17]. That is, if the basis functions estimate one of outcome or treatment well, the weighted mean is consistent. One novelty, introduced in this thesis, is, to choose universally consistent regression basis, such as partitioning estimates [GKKW02] and recover the results of [WZ19]. By the universal consistency we would expect, that both outcome and treatment are estimated well with sufficiently large samples.

A second novelty introduced in this thesis is, to use these weights and the weighted mean to estimate tho distribution function of potential outcomes.

We show, that (under mild assumptions) with the regression basis of partitioning estimates, the weighted mean is asymptotically well behaved in estimating distribution functions. This result is both new by regression basis and estimand. By the functional delta method [vdV00] we immediately get access to a large class of plug-in estimators.

With my thesis I contribute to one of the main purposes in causal inference, that is, reinforcing classical methods of statistical analysis for use in observational studies.

2 Entropy Balancing Weights

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.

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

$$\text{subject to} \qquad w_i \geq 0 \qquad\qquad\qquad\qquad \text{for all } i \in \{1,\ldots,n\} \,,$$

$$\frac{1}{N} \sum_{i=1}^{n} w_i = 1$$

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

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

$$[w_1, \ldots, w_n]^\top \;\mapsto\; \sum_{i=1}^{n} f(w_i) \tag{2.1}$$

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

$$[\lambda_1, \ldots, \lambda_n]^\top \;\mapsto\; \sum_{i=1}^{n} f^*(\lambda_i) \tag{2.2}$$

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

$$f\colon [0, \infty) \to \mathbb{R}, \quad w \mapsto \begin{cases} 0 & \text{if } w = 0, \\ w \log w & \text{else.} \end{cases} \tag{2.3}$$

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

$$(f')^{-1} = \lambda \mapsto e^{\lambda - 1} \tag{2.4}$$

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

$$f\colon \mathbb{R} \to \mathbb{R}, \quad w \mapsto (w - 1/n)^2 \tag{2.5}$$

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

$$(f')^{-1} = \lambda \mapsto \frac{\lambda}{2} + \frac{1}{n} \tag{2.6}$$

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

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

where we keep to the convention "$0/0 = 0$". It holds

$$\sum_{k=1}^N B_k(X_i) = 1 \qquad \text{for all } i \in \{1, \ldots, N\}. \tag{2.7}$$

Since $B_k(x) \in [0,1]$ for all $x \in \mathbb{R}^d$ and all $k \in \{1, \ldots, N\}$, it holds

$$\|B(x)\|_2^2 = \sum_{k=1}^N \left(\frac{\mathbf{1}_{X_k \in A_N(x)}}{\sum_{j=1}^N \mathbf{1}_{X_j \in A_N(x)}} \right)^2 \leq \sum_{k=1}^N \frac{\mathbf{1}_{X_k \in A_N(x)}}{\sum_{j=1}^N \mathbf{1}_{X_j \in A_N(x)}} = 1 \quad \text{for all } x \in \mathbb{R}^d. \tag{2.8}$$

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

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

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

$$\underset{w \in \mathbb{R}^n}{\text{minimize}} \qquad \varphi(w) \qquad\qquad (2.9)$$

$$\text{subject to} \qquad \mathbf{U}w \;\geq\; d\,,$$

$$\qquad\qquad\qquad\qquad \mathbf{A}w \;=\; a\,,$$

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. $\qquad\square$

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. $\diamond$

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

$$\lambda_d := \begin{bmatrix} \rho \\ \lambda^+ \\ \lambda^- \end{bmatrix} \qquad and \qquad \lambda_a := \lambda_0 \in \mathbb{R} \tag{2.10}$$

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

$$\underset{\rho,\lambda^+,\lambda^-,\lambda_0}{\text{maximize}} \qquad G(\rho, \lambda^+, \lambda^-, \lambda_0)$$

$$\text{subject to} \qquad \rho, \lambda^+, \lambda^- \geq 0,$$

where

$$G(\rho, \lambda^+, \lambda^-, \lambda_0) := -\sum_{i=1}^{n} f^* \left(\rho_i + \langle B(X_i), \lambda^+ - \lambda^- \rangle + \lambda_0 \right)$$

$$+ \sum_{i=1}^{N} \left(\langle B(X_i), \lambda^+ - \lambda^- \rangle + \lambda_0 \right)$$

$$- N \langle \delta, \lambda^+ + \lambda^- \rangle.$$

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. $\square$

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,

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

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

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

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

$$\tilde{\lambda} \; := \; \left[\; \lambda_1^+, \ldots, \; \lambda_i^+ - (\lambda_i^+ \wedge \lambda_i^-) \,, \; \ldots, \lambda_N^+, \; \lambda_1^-, \ldots, \lambda_i^- - (\lambda_i^+ \wedge \lambda_i^-) \,, \; \ldots, \lambda_N^-, \lambda_0 \right]^\top \,. \tag{2.11}$$

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

$$G(\tilde{\lambda}, \lambda_0) - G(\lambda, \lambda_0) \; = \; 2N \cdot \delta_i \cdot (\lambda_i^+ \wedge \lambda_i^-) \; > \; 0 \,, \tag{2.12}$$

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

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

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. $\qquad \square$

3 Asymptotic Analysis

3.1 Consistency of Optimal Solutions

3.1.1 Estimate of an Oracle Parameter by the Dual

Assumption 1. *The propensity score function satisfies $\pi(x) \in (0,1)$ for all $x \in \mathcal{X}$.*

Assumption 2. *The propensity score function is continuous in $\mathcal{X}$.*

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 omitted.

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

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

where $\pi(x) = \mathbf{P}[T = 1 | X = x]$ is the **propensity score** at $x \in \mathcal{X}$. Why this choice? With hindsight, we need for all $i \in \{1, \ldots, N\}$

$$\left| \langle B(X_i), \lambda^* \rangle - f'\left(\frac{1}{\pi(X_i)}\right) \right| \leq \omega\left(f' \circ (x \mapsto 1/x) \circ \pi, h_N\right). \tag{3.2}$$

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

$$\left| \langle B(X_i), \lambda^* \rangle - f'\left(\frac{1}{\pi(X_k)}\right) \right| \to 0 \qquad \text{almost surely.} \tag{3.3}$$

We need this to bound the term M_N in (3.12). The first big goal is to prove the following proposition.

Proposition 3.1. *Define the oracle parameter λ^* as in (3.1). Furthermore, assume that the propensity score function is continuous and positive on $\mathcal{X}$. Then for all $\varepsilon > 0$ with probability going to 1 there exists an optimal solution $(\lambda^\dagger, \lambda_0^\dagger)$ to Problem ? with $\left\| (\lambda^\dagger, \lambda_0^\dagger) - (\lambda^*, 0) \right\|_2 \leq \varepsilon$.*

Remark. In the analysis of the next sections will assume existence of an optimal solution to Problem? The purpose of Proposition 3.1 is therefore twofold. First, it ensures the existence of an optimal solution with probability going to 1 as $N \to \infty$. This shows that the assumption of existence is (at least for large N) likely met. Nevertheless, it is beyond the scope of my thesis to investigate feasibility of Problem ? in more detail. Second, if a solution exists it converges to the oracle parameter λ^* in probability. This ensures the consistency of the weights for the inverse propensity score. $\diamondsuit$

We use a hint from the last display of [WZ19, p.22]. The high-level idea is that the existence of the optimal solution and its proximity to the oracle parameter can be analysed by the objective function.

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 $\varepsilon > 0$*

$$\inf_{\|\Delta\| = \varepsilon} g(y + \Delta) - g(y) \geq 0 \tag{3.4}$$

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 (3.4) is true. Then it holds

$$g(x) < g(y^*) \quad \text{for some } x \in \mathbb{R}^m \setminus (y + \varepsilon B). \tag{3.5}$$

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

$$\theta x + (1 - \theta) y^* = y + \Delta_x. \tag{3.6}$$

It follows

$$g(y^*) \leq g(y) \leq g(y + \Delta_x) = g(\theta x + (1 - \theta) y^*)$$
$$\leq \theta g(x) + (1 - \theta) g(y^*) < g(y^*), \tag{3.7}$$

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 (3.4) being true, the equality is due to (3.6), the third inequality is due to the convexity of g and the strict inequality is due to (3.5). Thus every local minimum of g in $y + \varepsilon B$ is also a global minimum. $\qquad\square$

Remark. The hint from [WZ19, page 22] states (3.4) with strict inequality. They base their subsequent analysis on a quadratic Taylor expansion and aim to prove (3.4) with strict inequality. I show, that this approach is inefficient. To do that, I need Lemma 3.1 exactly as stated - with $\geq$ in (3.4). $\qquad\diamond$

Lemma 3.1 motivates the following lemma.

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

$$\mathbf{P}\left[\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \Delta_0) - G(\lambda^*, 0) \geq 0\right] \;\to\; 1 \qquad \text{for } N \to \infty. \qquad (3.8)$$

Proof. Recall the objective function G of Problem?

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

$$= \; g(\lambda, \lambda_0) \;+\; \langle\delta, |\lambda|\rangle$$

with

$$g \;:=\; (\lambda, \lambda_0) \;\mapsto\; \frac{1}{N}\sum_{i=1}^{N}\left[T_i \cdot f^*(\lambda_0 + \langle B(X_i), \lambda\rangle) \;-\; (\lambda_0 + \langle B(X_i), \lambda\rangle)\right]. \qquad (3.9)$$

Since we assume the convex conjugate f^* of f to be differentiable (it always convex), g is differentiable convex function with gradient

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

It is well know that a differentiable convex functions g satisfies

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

Thus

$$G(\lambda^* + \Delta, \Delta_0) \;-\; G(\lambda^*, 0)$$

$$\geq -\frac{1}{N}\sum_{i=1}^{N}\left[B(X_i)^\top, 1\right]\cdot\begin{bmatrix}\Delta \\ \Delta_0\end{bmatrix}\left(1 \;-\; T_i \cdot (f')^{-1}\left(\langle B(X_i), \lambda^*\rangle\right)\right) \qquad (3.10)$$

$$+ \; \langle\delta, |\lambda^* + \Delta| - |\lambda^*|\rangle.$$

Next, we fix $\tilde{\varepsilon} > 0$ and establish in (3.10) 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 $G(\lambda^* + \Delta, \Delta_0) - G(\lambda^*, 0)$ will give us the lower bound 0 in (3.10) with probability going to 1.

In (3.10) we control the **first term** by the law of large numbers and the **second term** by $\|\delta\|_1 \to 0$.

First Term

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

$$\left[B(X_i)^\top, 1 \right] \cdot \begin{bmatrix} \Delta \\ \Delta_0 \end{bmatrix} \;\lesssim\; \|(\Delta, \Delta_0)\| = \varepsilon \,. \tag{3.11}$$

Next, we see that

$$\frac{1}{N} \sum_{i=1}^{N} \left(1 - T_i \cdot (f')^{-1} \left(\langle B(X_i), \lambda^* \rangle \right) \right)$$
$$\lesssim \frac{1}{N} \sum_{i=1}^{N} \left| 1 - \frac{T_i}{\pi(X_i)} \right| + \frac{1}{N} \sum_{i=1}^{N} \left| \langle B(X_i), \lambda^* \rangle - f'\left(\frac{1}{\pi(X_i)} \right) \right| \tag{3.12}$$
$$=: S_N + M_N \,.$$

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

$$\mathbf{E}\left[\left| 1 - \frac{T}{\pi(X)} \right| \right] \leq 1 + \mathbf{E}\left[\frac{T}{\pi(X)} \right] = 2 \,. \tag{3.13}$$

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. By (2.7) it holds

$$
\begin{aligned}
&\left| \langle B(X_i), \lambda^* \rangle \;-\; f'\left(\frac{1}{\pi(X_i)} \right) \right| \\
&= \left| \sum_{k=1}^{N} B_k(X_i) \cdot f'\left(\frac{1}{\pi(X_k)} \right) \;-\; f'\left(\frac{1}{\pi(X_i)} \right) \right| \\
&= \left| \sum_{k=1}^{N} B_k(X_i) \left(f'\left(\frac{1}{\pi(X_k)} \right) - f'\left(\frac{1}{\pi(X_i)} \right) \right) \right| \\
&\leq \sum_{k=1}^{N} \frac{\mathbb{1}_{\{X_k \in A_N(X_i)\}}}{\sum_{j=1}^{N} \mathbb{1}_{\{X_j \in A_N(X_i)\}}} \left| f'\left(\frac{1}{\pi(X_k)} \right) - f'\left(\frac{1}{\pi(X_i)} \right) \right| \\
&\leq \omega\left(f' \circ (x \mapsto 1/x) \circ \pi, h_N \right) \;\to\; 0 \,,
\end{aligned}
\tag{3.14}
$$

where ω is the modulus of continuity. The convergence to 0 is due to the f' being continuous, $\pi(x) \in (0,1)$ for all $x \in \mathcal{X}$, the (assumed) continuity of π and $h_N \to 0$ for $N \to \infty$. 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 (3.12). Together with (3.11) we conclude that the **first term** in (3.10) is bounded below by $-\tilde{\varepsilon}/2$ with probability going to 1 as $N \to \infty$.

Second Term

It holds

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

Since $\delta \geq 0$ we get

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

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 (3.10) we conclude

$$
G(\lambda^* + \Delta, \Delta_0) \;-\; G(\lambda^*, 0) \geq -\tilde{\varepsilon}
\tag{3.15}
$$

with probability going to 1 as $N \to \infty$. But this holds for all $\tilde{\varepsilon} > 0$. A closer look reveals that $G(\lambda^* + \Delta, \Delta_0) \;-\; G(\lambda^*, 0)$ is measurable. Indeed, this holds because

X, T, $B(X)$ and λ^* are measurable and f^* is continuous. Since (3.15) holds true for all $\tilde{\varepsilon} > 0$, it thus follows

$$G(\lambda^* + \Delta, \Delta_0) \; - \; G(\lambda^*, 0) \geq 0 \tag{3.16}$$

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

$$\inf_{\|(\Delta,\Delta_0)\|=\varepsilon} G(\lambda^* + \Delta, \Delta_0) - G(\lambda^*, 0) \geq 0 \tag{3.17}$$

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

Proof. *(Proposition 3.1)* An immediate consequence of Lemma 3.1 is that for all $\varepsilon > 0$ it holds

$\mathbf{P}\left[\text{There exists an optimal solution } (\lambda^\dagger, \lambda_0^\dagger) \text{ to Problem? with } \left\|(\lambda^\dagger, \lambda_0^\dagger) - (\lambda^*, 0)\right\|_2 \leq \varepsilon\right]$

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

Applying Lemma 3.2 finishes the proof. $\qquad\square$

3.1.2 Estimate of the Inverse Propensitiy Score by the Weights

Before we proof the theorem, we proof some lemmas that are useful in the next section.

> **Assumption 3.** *There exists $\underline{N} \in \mathbb{N}$ such that for all $N \geq \underline{N}$ and for all data sets D_N the optimal solution $(\lambda^\dagger, \lambda_0^\dagger)$ to Problem? exists in a (compact) parameter space $\Theta_N \subset \mathbb{R}^{N+1}$, containing the origin, with $\operatorname{diam}\Theta_N < \infty$.*

> **Assumption 4.** *The objective function f to Problem? is strictly convex, with strictly convex and continuously differentiable convex conjugate f^*.*

Definition 3.1. *We define the weight function to be*

$$w\colon \mathcal{X} \times \mathbb{R}^{N+1} \to \mathbb{R}, \qquad (x, \lambda, \lambda_0) \mapsto (f')^{-1}\left(\langle B(x), \lambda^\dagger\rangle + \lambda_0^\dagger\right) .$$

If the optimal solution $(\lambda^\dagger, \lambda_0^\dagger)$ to Problem ? exists, we write

$$w^\dagger(x) = w(x, \lambda^\dagger, \lambda_0^\dagger) \qquad \text{for all } x \in \mathcal{X} .$$

Remark. The (optimal) weight function $w^\dagger$ depends on the data $D_N := (T_i, X_i)_{i \in \{1,\dots,N\}}$ and the partition $\mathcal{P}_N$. $\hfill \diamond$

Lemma 3.3. *Let Assumption 3 and Assumption 4 hold true. Then it holds $w^\dagger(X) \in L^\infty(\mathbf{P})$ for all $N \geq \underline{N}$.*

Proof. Let $N \geq \underline{N}$. By (2.8) it holds for all $x \in \mathcal{X}$ and for all $(\lambda, \lambda_0) \in \mathbb{R}^{N+1}$ it holds

$$|\langle B(x), \lambda \rangle + \lambda_0| \leq \|(B(x), 1)\|_2 \|(\lambda, \lambda_0)\|_2 \leq 2 \|(\lambda, \lambda_0)\|_2 . \tag{3.18}$$

By Assumption 3, there exists a (compact) parameter space Θ_N around the origin, with $\operatorname{diam}\Theta_N < \infty$, such that for all data sets D_N it holds $(\lambda^\dagger, \lambda_0^\dagger) \in \Theta_N$. By Assumption 4, $(f')^{-1}$ is non-decreasing and continuous. Thus by (3.18) it holds

$$\left| w^\dagger(x) \right| \leq \left| (f')^{-1}(-2\operatorname{diam}\Theta_N) \right| + \left| (f')^{-1}(2\operatorname{diam}\Theta_N) \right| < \infty$$

for all $x \in \mathcal{X}$. $\qquad\square$

Lemma 3.4. *Let Assumption 3 and Assumption 4 hold true. Furthermore, let $N \geq \underline{N}$, and let $Z \in L^1(\mathbf{P})$ be a random variable that is independent of D_N with $\mathbf{E}[Z \mid X] = 0$ almost surely. It holds $\mathbf{E}\left[w^\dagger(X) \cdot Z\right] = 0$.*

Proof. Let $N \geq \underline{N}$. By Lemma 3.3 and $Z \in L^1(\mathbf{P})$ it holds

$$\left\| w^\dagger(X) \cdot Z \right\|_{L^1(\mathbf{P})} \leq \left\| w^\dagger(X) \right\|_{L^\infty(\mathbf{P})} \|Z\|_{L^1(\mathbf{P})} < \infty . \tag{3.19}$$

By (3.19), $Z \perp D_N$ and $\mathbf{E}[Z \mid X] = 0$ almost surely it holds

$$\mathbf{E}\left[w^\dagger(X) \cdot Z \mid D_N, X\right] = w^\dagger(X) \cdot \mathbf{E}[Z \mid D_N, X]$$
$$= w^\dagger(X) \cdot \mathbf{E}[Z \mid X] = 0 \qquad \text{almost surely.}$$

Thus

$$\mathbf{E}\left[w^\dagger(X) \cdot Z\right] = \mathbf{E}\left[\mathbf{E}\left[w^\dagger(X) \cdot Z \mid D_N, X\right]\right] = 0 .$$

$\qquad\square$

Theorem 3.1. *Let Assumption 3 hold true. Under the conditions of Proposition 3.1 it holds $w^\dagger(X) \xrightarrow{\mathbf{P}} 1/\pi(X)$. Furthmore, there exists a decreasing sequence*

$(\varepsilon_N) \subset (0,1]$ *such that* $\varepsilon_N \to 0$ *and*

$$\mathbf{P}\left[\left|w^\dagger(X) - \frac{1}{\pi(X)}\right| \le \varepsilon_N\right] \to 0 \qquad \text{for } N \to \infty. \tag{3.20}$$

Proof. Let $N \ge \underline{N}$. For all $\varepsilon > 0$ it holds

$$
\begin{aligned}
\left|w^\dagger(X) - \frac{1}{\pi(X)}\right| &= \left|(f')^{-1}\left(\langle B(X), \lambda^\dagger\rangle + \lambda_0^\dagger\right) - \frac{1}{\pi(X)}\right| \\
&\lesssim \left|\langle (B(X), 1), (\lambda^\dagger, \lambda_0^\dagger) - (\lambda^*, 0)\rangle\right| + \left|\langle B(X), \lambda^*\rangle - f'\left(\frac{1}{\pi(X)}\right)\right| \\
&\lesssim \left\|(\lambda^\dagger, \lambda_0^\dagger) - (\lambda^*, 0)\right\|_2 + \left|\sum_{i=1}^{N} B_k(X) \cdot f'\left(\frac{1}{\pi(X_k)}\right) - f'\left(\frac{1}{\pi(X)}\right)\right| \\
&\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \le \varepsilon,
\end{aligned}
\tag{3.21}
$$

with probability going to 1 as $N \to \infty$. The convergence of the first term follows from Proposition 3.1 and that of the second term from (3.14). The second statement of the theorem follows from the Selection lemma [SC08, A.1.4.]. $\qquad\square$

Gaussian Bridge

$(\lambda^\dagger, \lambda_0^\dagger)$ We adapt the error decomposition in [WZ19, page 27] to estimates of the distribution function $F_{Y(1)}$ of $Y(1)$, that is,

$$F_{Y(1)} : \mathbb{R} \to [0,1], \qquad z \mapsto \mathbf{P}[Y(1) \le z]. \tag{3.22}$$

Assumption 5. *For any decreasing sequence* (ε_N) *with* $\varepsilon_N \to 0$ *for* $N \to \infty$, *there exists a sequence of (measurable) function classes* $(\mathcal{F}_N)$ *with envelope functions* (F_N), *satisfying for some* $k < 2$

$$\|F_N\|_{L^2(\mathbf{P})} \le \varepsilon_N \quad \text{and} \quad \log N_{[]}(\varepsilon, \mathcal{F}_N, \mathrm{L}_2(\mathbf{P})) \lesssim \left(\frac{1}{\varepsilon}\right)^k \quad \text{for all } N \in \mathbb{N},$$

such that for all $(\lambda, \lambda_0) \in \mathbb{R}^{N+1}$ *and for all* $N \in \mathbb{N}$ *the function*

$$\mathcal{X} \to \mathbb{R}, \qquad x \mapsto \mathbf{1}_{\left\{\sup_{y \in A_N(x)}\left|w(y,\lambda,\lambda_0) - \frac{1}{\pi(y)}\right| \le \varepsilon_N\right\}}\left(w(x,\lambda,\lambda_0) - \frac{1}{\pi(x)}\right)$$

$$\tag{3.23}$$

> *is contained in $\mathcal{F}_N$.*

Remark. We can derive Assumption 5 from regularity assumptions on the inverse propensity score $1/\pi(\cdot)$ and the distribution of X. To this end, we want to employ [vdvW13, Corollary 2.7.4]. To do this, the crucial observation is, that

$$w(\cdot, \lambda, \lambda_0) \quad \text{is constant on each cell } A_N \in \mathcal{P}_N.$$

Thus, the regularity of the function (3.23) on each cell $A_N \in \mathcal{P}_N$ is decided by $1/\pi(\cdot)$. Thus, we assume that there exists $\alpha > d/2$, where $\mathcal{X} \subseteq \mathbb{R}^d$, such that for $V := d/\alpha$. and for all $(j, N) \in \mathbb{N}^2$ there exists $M_{N,j} \geq \varepsilon_N$ such that

$$\frac{1}{\pi(\cdot)} \in C^{\alpha}_{M_{N,j}}(A_{N,j}) \quad \text{and} \quad \sum_{j=1}^{\infty} M_{N,j}^{2V/(V+2)} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \lesssim 1.$$

Then the function 3.23, restricted to $A_{N,j}$, is in $C^{\alpha}_{M_{N,j}}(A_{N,j})$ for all $(j, N) \in \mathbb{N}^2$. Then [vdvW13, Corollary 2.7.4] gives us the statement in Assumption 5. $\diamond$

Theorem 3.2. *Under conditions the stochastic process*

$$\sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{n} w_i^{\dagger} \mathbf{1}_{\{Y_i \leq z\}} - F_{Y(1)}(z) \right)_{z \in \mathbb{R}} \tag{3.24}$$

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

$$\begin{aligned}
&\mathbf{Cov}(z_1, z_2) \\
&= \mathbf{E} \left[\frac{F_{Y(1)}(z_1 \wedge z_2 \mid X)}{\pi(X)} - \frac{1 - \pi(X)}{\pi(X)} F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X) \right] \\
&\quad - F_{Y(1)}(z_1) \cdot F_{Y(1)}(z_2)
\end{aligned} \tag{3.25}$$

Lemma 3.5. *It holds*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{n}w(X_i)\mathbf{1}_{\{Y_i\leq z\}} \;-\; F_{Y(1)}(z)\right)_{z\in\mathbb{R}} \;=\; R_1 \;+\; R_2 \;+\; R_3 \;+\; R_4 \quad (3.26)$$

with

$$R_1 \;:=\; \sqrt{N}\sum_{k=1}^{N}\left[\frac{1}{N}\left(\sum_{i=1}^{n}w(X_i)B_k(X_i) - \sum_{i=1}^{N}B_k(X_i)\right)F_{Y(1)}(z|X_k)\right]_{z\in\mathbb{R}},$$

$$R_2 \;:=\; \sqrt{N}\sum_{i=1}^{N}\left[\frac{T_i\cdot w(X_i)-1}{N}\left(F_{Y(1)}(z|X_i) \;-\; \sum_{k=1}^{N}B_k(X_i)\cdot F_{Y(1)}(z|X_k)\right)\right]_{z\in\mathbb{R}},$$

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

$$R_4 \;:=\; \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}_{\{Y_i\leq z\}}-F_{Y(1)}(z|X_i)\right) \;+\; \left(F_{Y(1)}(z|X_i)-F_{Y(1)}(z)\right)\right)_{z\in\mathbb{R}}.$$

Proof. We fix $z \in \mathbb{R}$. It holds

$$\frac{1}{N} \sum_{i=1}^{N} w^\dagger(X_i) \cdot T_i \cdot \mathbf{1}_{\{Y_i(T_i) \leq z\}}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(w^\dagger(X_i) - \frac{1}{\pi(X_i)} \right) T_i \cdot \mathbf{1}_{\{Y_i(T_i) \leq z\}}$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{\pi(X_i)} \mathbf{1}_{\{Y_i(T_i) \leq z\}}$$

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

$$+ \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{\pi(X_i)} \left(\mathbf{1}_{\{Y_i(T_i) \leq z\}} - F_{Y(1)}(z|X_i) \right)$$

$$+ \frac{1}{N} \sum_{i=1}^{N} w^\dagger(X_i) \cdot T_i \cdot F_{Y(1)}(z|X_i)$$

$$= R_3(z)/\sqrt{N}$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{\pi(X_i)} \left(\mathbf{1}_{\{Y_i(T_i) \leq z\}} - F_{Y(1)}(z|X_i) \right) + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z) \right)$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left(w^\dagger(X_i) \cdot T_i - 1 \right) F_{Y(1)}(z|X_i)$$

$$+ F_{Y(1)}(z)$$

$$= R_3(z)/\sqrt{N}$$

$$+ R_4(z)/\sqrt{N}$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left(w^\dagger(X_i) \cdot T_i - 1 \right) \left(F_{Y(1)}(z|X_i) - \sum_{k=1}^{N} B_k(X_i) \cdot F_{Y(1)}(z|X_k) \right)$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left(w^\dagger(X_i) \cdot T_i - 1 \right) \sum_{k=1}^{N} B_k(X_i) \cdot F_{Y(1)}(z|X_k)$$

$$+ F_{Y(1)}(z)$$

$$= R_3(z)/\sqrt{N}$$

$$+ R_4(z)/\sqrt{N}$$

$$+ R_2(z)/\sqrt{N}$$

$$+ \sum_{k=1}^{N} \frac{1}{N} \sum_{i=1}^{N} \left(w^\dagger(X_i) \cdot T_i B_k(X_i) - B_k(X_i) \right) \cdot F_{Y(1)}(z|X_k)$$

$$+ F_{Y(1)}(z)$$

$$= \left(R_3(z) + R_4(z) + R_2(z) + R_1(z) \right)/\sqrt{N} + F_{Y(1)}(z).$$

This holds for all $z \in \mathbb{R}$. Multiplying with $\sqrt{N}$ yields the result. $\qquad\square$

Let $F_{Y(1)}(z|x) := \mathbf{P}[Y(1) \leq z | X = x]$ denote a conditional version of the distribution function of $Y(1)$ at $x \in \mathcal{X}$. We also need the propensity score $\pi(x) := \mathbf{P}[T = 1 | X = x]$ and the weights function $w(x) := (f')^{-1}\left(\langle B(x), \lambda^\dagger \rangle + \lambda_0^\dagger\right)$.

Lemma 3.6. *Let the weights function w satisfy the box constraints in Problem 2.1 and $\sqrt{N}\,\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$. Then it holds $\sup_{z \in \mathbb{R}} |R_1(z)| \xrightarrow{\mathbf{P}} 0$.*

Proof. It holds

$$
\begin{aligned}
\sup_{z \in \mathbb{R}} |R_1(z)| \;&=\; \sup_{z \in \mathbb{R}} \left| \sqrt{N} \sum_{k=1}^{N} \left[\frac{1}{N} \left(\sum_{i=1}^{n} w(X_i) B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) F_{Y(1)}(z|X_k) \right] \right| \\
&\leq\; \sqrt{N} \sum_{k=1}^{N} \left| \frac{1}{N} \left(\sum_{i=1}^{n} w(X_i) B_k(X_i) - \sum_{i=1}^{N} B_k(X_i) \right) \right| \sup_{z \in \mathbb{R}} F_{Y(1)}(z|X_k) \\
&\leq\; \sqrt{N}\,\|\delta\|_1
\end{aligned}
\tag{3.27}
$$

The last inequality is due to $F_{Y(1)} \in [0,1]$ and the assumption that $(w(X_i))$ satisfies the box constraints of Problem 2.1. Since we assume $\sqrt{N}\,\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$ it holds $\sup_{z \in \mathbb{R}} |R_1(z)| \xrightarrow{\mathbf{P}} 0$. $\qquad\square$

Remark. We want to comment on the box constraints of Problem 2.1, that is,

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

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

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

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.

Also we want to comment on the assumption on $\|\delta\|$. Playing around with norm equivalences we discover that $\sqrt{N}\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$ for $N \to \infty$ is the weakest (natural) assumption to control R_1. Indeed, other ways to continue the second row in (3.27) 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 $\sqrt{N}\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$ for $N \to \infty$ also suffices to control the second (or first) occurrence of a term, that we control by assumptions on $\|\delta\|$. This is the **second term** of (3.10), where we estimate

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

$$\diamondsuit$$

> **Lemma 3.7.** *Let the conditions of Theorem 3.1 hold true. Furthermore assume, that the width of the partitioning estimate h_N and a conditional version of the distribution function of $Y(1)$ satisfy*
>
> $$\sqrt{N}\sup_{z\in\mathbb{R}} \omega\big(F_{Y(1)}(z|\cdot), h_N \big) \to 0 \qquad \text{for } N \to \infty \,,$$
>
> *where ω is the modulus of continuity. Then it holds $\sup_{z\in\mathbb{R}}|R_2(z)| \xrightarrow{\mathbf{P}} 0$.*

Proof.

$$\sup_{z\in\mathbb{R}}|R_2(z)|$$

$$\leq\; \sqrt{N}\sum_{i=1}^N \left[\frac{T_i\cdot w(X_i)-1}{N} \sup_{z\in\mathbb{R}} \left| F_{Y(1)}(z|X_i) \;-\; \sum_{k=1}^N B_k(X_i)\cdot F_{Y(1)}(z|X_k) \right| \right]$$

$$\leq\; \sqrt{N}\sup_{z\in\mathbb{R}} \omega\big(F_{Y(1)}(z|\cdot), h_N \big) \sum_{i=1}^N \frac{T_i\cdot w(X_i)+1}{N}$$

$$=\; 2\sqrt{N}\sup_{z\in\mathbb{R}} \omega\big(F_{Y(1)}(z|\cdot), h_N \big) \,.$$

3 Asymptotic Analysis

The equality is due to

$$1 \;=\; \frac{1}{N}\sum_{i=1}^{n} w_i^{\dagger} \;=\; \frac{1}{N}\sum_{i=1}^{n} w(X_i) \;=\; \frac{1}{N}\sum_{i=1}^{N} T_i w(X_i)\,, \qquad (3.28)$$

that is, $w(X_i)$ satisfy the second constraint of Problem 2.1. The second inequality follows from $\sum_{k=1}^{N} B_k(X) = 1$ and the convexity of the absolute value. Indeed,

$$\sup_{z\in\mathbb{R}} \left| F_{Y(1)}(z|X_i) - \sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k) \right|$$

$$\leq \sum_{k=1}^{N} \frac{\mathbf{1}_{\{X_k \in A_N(X_i)\}}}{\sum_{j=1}^{N} \mathbf{1}_{\{X_j \in A_N(X_i)\}}} \sup_{z\in\mathbb{R}} \left| F_{Y(1)}(z|X_i) - F_{Y(1)}(z|X_k) \right|$$

$$\leq \sup_{z\in\mathbb{R}} \omega\left(F_{Y(1)}(z|\cdot), h_N \right)\,.$$

Since we assume

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

it follows $\sup_{z\in\mathbb{R}} |R_2(z)| \xrightarrow{\mathbf{P}} 0$. $\qquad\qquad\square$

Remark. 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.

We also want to comment on the assumption

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

I decided to keep this more general (and abstract) assumption, althogh there are many (more concrete, yet stronger) assumptions on the regularity of $F_{Y(1)}(z|\cdot)$ and the convergence speed of h_N. If for example $F_{Y(1)}(z|\cdot)$ is α-Hölder continuous with $\alpha \in (0,1]$ for all $z \in \mathbb{R}$, it suffices $\sqrt{N}h_N^{\alpha} \to 0$ to control R_2. $\qquad\qquad\diamond$

To control the remaining terms R_3 and R_4 we use empirical processes. We introduce the concept and the results we need in the next paragraphs. For an introduction to empirical processes see [vdV00]. More advanced techniques are in [vdvW13].

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, $(\mathcal{Z}, \Sigma)$ a measurable space, and $\xi_1, \ldots, \xi_N :$ $(\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{Z}, \Sigma)$ a sample of independent and identically-distributed random variables with probability distribution $\mathbf{P}_\xi$. A family $\mathcal{F}$ of measurable functions $f :$ $(\mathcal{Z}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induces a stochastic process by

$$f \ \mapsto \ \mathbb{G}_N f \ := \ \frac{1}{\sqrt{n}} \sum_{i=1}^{N} (f(\xi_i) - \mathbf{E}_\xi[f]) \, . \tag{3.29}$$

We call this the **empirical process** $\mathbb{G}_N$ indexed by $\mathcal{F}$. We define the (random) norm

$$\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}_N f| \, . \tag{3.30}$$

Remark. We stress that $\|\mathbb{G}_n\|_{\mathcal{F}}$ often ceases to be measurable, even in simple situations [vdvW13, page 3]. To deal with this, we introduce the notion of **outer expectation** $\mathbf{E}^*$, that is,

$$\mathbf{E}^*[Z] \ := \ \inf \left\{ \mathbf{E}[U] \mid U \geq Z, \ U : (\Omega, \mathcal{A}, \mathbf{P}) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}})) \text{ measurable and } \mathbf{E}[U] < \infty \right\} \, . \tag{3.31}$$

In our application the technical difficulties halt at this point, because we only consider Z with $\mathbf{E}^*[Z] < \infty$. Then there exists a smallest measurable function Z^* dominating Z with $\mathbf{E}^*[Z] = \mathbf{E}[Z^*]$. Thus, we may assume Z to be measurable in this regard.

$\diamondsuit$

In our application we need concentration inequalities for $\|\mathbb{G}_n\|_{\mathcal{F}}$. One easy way is to use maximal inequalities for the expectation together with Markov's inequality. There are also Bernstein-like inequalities for empirical processes. We need to introduce more concepts.

Given two functions $\underline{f}$ and $\overline{f}$, the bracket $[\underline{f}, \overline{f}]$ is the set of all functions f with $\underline{f} \leq f \leq \overline{f}$. We define a $(\varepsilon, L^r(\mathbf{P}))$-bracket to be a bracket $[\underline{f}, \overline{f}]$ with $\|\overline{f} - \underline{f}\|_{L^r(\mathbf{P})} < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L^r(\mathbf{P}))$ is the minimum number of $(\varepsilon, L^r(\mathbf{P}))$-brackets needed to cover $\mathcal{F}$. For most classes $\mathcal{F}$ the bracketing number grows to infinity for $\varepsilon \to 0$. To measure the speed of convergence we introduce the bracketing integral

$$J_{[]}(\delta, \mathcal{F}, L_r(\mathbf{P})) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_N, L^r(\mathbf{P}))} \, d\varepsilon \, . \tag{3.32}$$

An envelope function F of a class $\mathcal{F}$ satisfies $|f(x)| \leq F(x) < \infty$ for all $f \in \mathcal{F}$ and all x.

Theorem 3.3. (Maximal inequality) *For any class $\mathcal{F}$ of measurable functions with envelope function F,*

$$\mathbf{E}^*_{\mathbf{P}}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J_{[]}(\|F\|_{L^2(\mathbf{P})}, \mathcal{F}, L_2(\mathbf{P})). \tag{3.33}$$

Proof. [vdV00, Corollary 19.35] ☐

Next, we define some auxiliary functions. For $z \in \mathbb{R}$ we define the function

$$f_z \; : \; \{0,1\} \times \mathcal{X} \times \mathcal{Y} \; \to \; \mathbb{R}$$
$$(t,x,y) \; \mapsto \; t\left(\mathbf{1}_{\{y \leq z\}} - F_{Y(1)}(z|x)\right),$$

and the function classes

$$\begin{aligned}
\mathcal{F} &:= \{f_z \mid z \in \mathbb{R}\} \\
\mathcal{G} &:= \left\{\frac{f_z}{\pi(\cdot)} + F_{Y(1)}(z|\cdot) - F_{Y(1)}(z) \; : \; z \in \mathbb{R}\right\}.
\end{aligned} \tag{3.34}$$

Assumption 6. *It holds*

$$(Y(0), Y(1)) \perp T \mid X \qquad \text{and} \qquad 0 < \pi(x) < 1 \quad \text{for all } x \in \mathcal{X}.$$

Lemma 3.8. *It holds $f_z(T, X, Y(T)) \in L^1(\mathbf{P})$ and $f_z(T, X, Y(T)) \perp D_N$ for all $z \in \mathbb{R}$. If also Assumption 6 holds, then for all $z \in \mathbb{R}$*

$$\mathbf{E}\left[f_z\left(T, X, Y(T)\right) \mid X\right] \; = \; 0 \qquad \text{almost surely.}$$

Proof. Since f_z is bounded by 1, it holds $f_z(T, X, Y(T)) \in L^1(\mathbf{P})$. Since

$$(T, X, Y(T)) \; \perp \; D_N \; = \; (T_i, X_i)_{i \in \{1,\dots,N\}}$$

it holds $f_z(T, X, Y(T)) \perp D_N$ for all $z \in \mathbb{R}$. For the third statement, note that

$$\begin{aligned}
\mathbf{E}\left[f_z\left(T, X, Y(T)\right) \mid X\right] &= \mathbf{E}\left[T\left(\mathbf{1}_{\{Y(T) \leq z\}} - F_{Y(1)}(z|X)\right) \mid X\right] \\
&= \mathbf{E}\left[\mathbf{1}_{\{Y(1) \leq z\}} - F_{Y(1)}(z|X) \mid X, T = 1\right]\pi(X) \\
&= \left(\mathbf{E}\left[\mathbf{1}_{\{Y(1) \leq z\}} \mid X\right] - F_{Y(1)}(z|X)\right)\pi(X) \\
&= 0 \qquad \text{almost surely.}
\end{aligned}$$

The third equality is due to Assumption 6. ☐

The next lemma provides bracketing numbers for specific function classes needed to control R_3 and R_4.

Lemma 3.9. *The function class $\mathcal{F}$ and $\mathcal{G}$ defined in (3.34) are measurable. Furthermore,*

$$N_{[]}(\varepsilon, \mathcal{F}, L^2(\mathbf{P})) \;\lesssim\; \left(\frac{1}{\varepsilon}\right)^2 \qquad \text{for all } \varepsilon > 0.$$

If $1/\pi(X) \in L^2(\mathbf{P})$, it also holds

$$N_{[]}(\varepsilon, \mathcal{G}, L^2(\mathbf{P})) \;\lesssim\; \left(\frac{1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}}{\varepsilon}\right)^4 \qquad \text{for all } \varepsilon > 0.$$

Proof. As in [vdV00, Example 19.6] we choose for $\varepsilon > 0$ and $m \in \mathbb{N}$

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

such that

$$\mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\right] \;\leq\; \varepsilon \qquad \text{for all } l \in \{1, \ldots, m\} \tag{3.35}$$

and $m \leq 2/\varepsilon$. Next, we define m brackets by

$$\overline{f_l}(t, x, y) \;:=\; t\left(\mathbf{1}_{\{y \leq z_l\}} - F_{Y(1)}(z_{l-1}|x)\right),$$
$$\underline{f_l}(t, x, y) \;:=\; t\left(\mathbf{1}_{\{y \leq z_{l-1}\}} - F_{Y(1)}(z_l|x)\right),$$

for $l \in \{1, \ldots, m\}$. These brackets cover $\mathcal{F}$. Indeed,

for all $z \in \mathbb{R}$ there exists $l \in \{1, \ldots, m\}$ such that $z_{l-1} \leq z \leq z_l$.

By the monotonicity of $\mathbf{1}_{\{y \leq (\cdot)\}}$ and $F_{Y(1)}(\cdot|x)$ and the non-negativity of T it follows

for all $z \in \mathbb{R}$ there exists $l \in \{1, \ldots, m\}$ such that $\underline{f_l} \leq f_z \leq \overline{f_l}$.

Thus, the m brackets $[\underline{f_l}, \overline{f_l}]$ cover $\mathcal{F}$.

Let's calculate the size of the brackets. It holds

$$\mathbf{E}\left[T \cdot \left(\mathbf{1}_{\{Y(T) \leq z_l\}} - F_{Y(1)}(z_{l-1}|X) \;-\; \mathbf{1}_{\{Y(T) \leq z_{l-1}\}} + F_{Y(1)}(z_l|X)\right)\right]$$
$$= \; \mathbf{E}\left[T \cdot \left(\mathbf{1}_{\{Y(T) \in [z_{l-1}, z_l]\}} \;+\; \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right)\right]$$
$$\leq \; \mathbf{E}\left[\pi(X) \cdot \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right] \;+\; \varepsilon$$
$$\leq \; 2\varepsilon.$$

We used (3.35), $0 \leq T, \pi(X) \leq 1$ and Lemma 3.13. It follows

$$\left\|\left(\overline{f_l} - \underline{f_l}\right)(T, X, Y(T))\right\|_{L^2(\mathbf{P})}$$
$$\lesssim \; \mathbf{E}\left[T \cdot \left(\mathbf{1}_{\{Y(T) \in [z_{l-1}, z_l]\}} \;+\; \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right)\right]^{1/2} \;\lesssim\; \varepsilon^{1/2}.$$

Since $m \leq 2/\varepsilon$ it holds

$$N_{[\,]}\left(\varepsilon^{1/2}, \mathcal{F}, L^2(\mathbf{P})\right) \lesssim \frac{1}{\varepsilon}$$

and thus

$$N_{[\,]}(\varepsilon, \mathcal{F}, L^2(\mathbf{P})) \lesssim \left(\frac{1}{\varepsilon}\right)^2.$$

Next, we look at $\mathcal{G}$. To this end, we define m brackets by

$$\overline{g_l}(t, x, y) := \frac{t}{\pi(x)}\left(\mathbf{1}_{\{y \leq z_l\}} - F_{Y(1)}(z_{l-1}|x)\right) + F_{Y(1)}(z_l|x) - F_{Y(1)}(z_{l-1}),$$

$$\underline{g_l}(t, x, y) := \frac{t}{\pi(x)}\left(\mathbf{1}_{\{y \leq z_{l-1}\}} - F_{Y(1)}(z_l|x)\right) + F_{Y(1)}(z_{l-1}|x) - F_{Y(1)}(z_l),$$

for $l \in \{1, \ldots, m\}$. With the same arguments as before, we see that these brackets cover $\mathcal{G}$. Let's calculate the size. It holds

$$\left\|\frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T) \in [z_{l-1}, z_l]\}} + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\,|\,X\right]\right)\right\|_{L^2(\mathbf{P})}$$

$$\lesssim \left(\mathbf{E}\left[\frac{1}{\pi(X)}\frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T) \in [z_{l-1}, z_l]\}} + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\,|\,X\right]\right)\right]\right)^{1/2}$$

$$\lesssim \left(\mathbf{E}\left[\frac{1}{\pi(X)}\mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\,|\,X\right]\right]\right)^{1/2}$$

$$\lesssim \left(\|1/\pi(X)\|_{L^2(\mathbf{P})}\sqrt{\varepsilon}\right)^{1/2} = \varepsilon^{1/4}\|1/\pi(X)\|_{L^2(\mathbf{P})}^{1/2}$$

and

$$\left\|\mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\,|\,X\right] + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\right]\right\|_{L^2(\mathbf{P})} \lesssim \varepsilon^{1/2}.$$

Thus

$$\left\|\left(\overline{g_l} - \underline{g_l}\right)(T, X, Y(T))\right\|_{L^2(\mathbf{P})} \lesssim \varepsilon^{1/4}\left(1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}^{1/2}\right)$$

$$\lesssim \varepsilon^{1/4}\left(1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}\right).$$

As before, it follows

$$N_{[\,]}(\varepsilon, \mathcal{G}, L^2(\mathbf{P})) \lesssim \left(\frac{1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}}{\varepsilon}\right)^4.$$

$\square$

Lemma 3.10. *Let $(\mathcal{H}_N)$ be a sequence of measurable function classes with envelope functions (H_N). If*

$$J_{[]}\left(\|F_N\|_{L^2(\mathbf{P})}, \mathcal{H}_N, L^2(\mathbf{P})\right) \;\to\; 0 \qquad \text{for } N \to \infty \tag{3.36}$$

it holds $\|\mathbb{G}_N\|^*_{\mathcal{H}_N} \xrightarrow{\mathbf{P}} 0$.

Proof. By Markov's inequality and Theorem 3.3 it holds for all $\varepsilon > 0$

$$\mathbf{P}[\|\mathbb{G}_N\|^*_{\mathcal{H}_N} \geq \varepsilon]$$
$$\leq\; \varepsilon^{-1}\mathbf{E}[\|\mathbb{G}_N\|^*_{\mathcal{H}_N}] \;=\; \varepsilon^{-1}\mathbf{E}^*[\|\mathbb{G}_N\|_{\mathcal{H}_N}] \;\lesssim\; \varepsilon^{-1}J_{[]}\left(\|H_N\|_{L^2(\mathbf{P})}, \mathcal{H}_N, L^2(\mathbf{P})\right) \;\to\; 0$$

for $N \to \infty$. $\qquad\square$

A technical lemma for products of function classes.

Define the product of two function classes

$$\mathcal{F} \cdot \mathcal{G} := \{f \cdot g \colon f \in \mathcal{F}, g \in \mathcal{G}\}\,.$$

Lemma 3.11. *Let $\mathcal{F}$ and $\mathcal{G}$ be two function classes with envelope functions F and G satisfying $\|F\|_\infty, \|G\|_\infty \leq 1$. For all $\varepsilon > 0$ and all $r \in [1, \infty)$ it holds*

$$N_{[]}(2\varepsilon, \mathcal{F} \cdot \mathcal{G}, \mathrm{L}_r(\mathbf{P})) \;\leq\; N_{[]}(\varepsilon, \mathcal{F}, \mathrm{L}_r(\mathbf{P})) \cdot N_{[]}(\varepsilon, \mathcal{G}, \mathrm{L}_r(\mathbf{P}))\,.$$

Proof. Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$. We can choose two $(\varepsilon, L^r(\mathbf{P}))$ brackets $[\underline{f}, \overline{f}]$ and $[\underline{g}, \overline{g}]$ containing f and g with $\|\underline{f}\|_\infty, \|\overline{f}\|_\infty \leq \|F\|_\infty \leq 1$ and $\|\underline{g}\|_\infty, \|\overline{g}\|_\infty \leq \|G\|_\infty \leq 1$. We the get an $(2\varepsilon, L^r(\mathbf{P}))$ $[\underline{h}, \overline{h}]$ bracket, containing $f \cdot g$, by $\qquad\square$

Lemma 3.12. *Let $(\varepsilon_N) \subset (0, 1]$ be a decreasing sequence with $\varepsilon_N \to 0$ for $N \to \infty$ and let Assumption 5 hold true for a sequence of function classes $\mathcal{F}_N$. Then*

$$J_{[]}(\|F_N\|_{L^2(\mathbf{P})}, \mathcal{F}_N \cdot \mathcal{F}, \mathrm{L}_2(\mathbf{P})) \to 0 \quad \text{and} \quad \|\mathbb{G}_N\|^*_{\mathcal{F}_N \cdot \mathcal{F}} \xrightarrow{\mathbf{P}} 0 \qquad \text{for } N \to \infty\,.$$

Proof. By Assumption 5 and Lemma 3.9 it holds for some $k < 2$

$$\|F_N\|_{L^2(\mathbf{P})} \;\leq\; \varepsilon_N \quad \text{and} \quad \log N_{[]}(\varepsilon, \mathcal{F}_N, \mathrm{L}_2(\mathbf{P})) \;\lesssim\; \left(\frac{1}{\varepsilon}\right)^k \qquad \text{for all } N \in \mathbb{N},$$

and

$$N_{[]}(\varepsilon, \mathcal{F}, L^2(\mathbf{P})) \lesssim \left(\frac{1}{\varepsilon}\right)^2 \qquad \text{for all } \varepsilon > 0 \,.$$

Since $\mathcal{F}_N$ and $\mathcal{F}$ have envelope function smaller 1, we can apply Lemma 3.11 to get

$$\log N_{[]}(\varepsilon, \mathcal{F}_N \cdot \mathcal{F}, \mathrm{L}_2(\mathbf{P})) \lesssim \left(\frac{1}{\varepsilon}\right)^k + \log(1/\varepsilon) \lesssim \left(\frac{1}{\varepsilon}\right)^k \qquad \text{for all } \varepsilon > 0 \,.$$

Since $k/2 \in (0, 1)$ it holds

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

The second statement follows from Lemma 3.10 for $\mathcal{H}_N := \mathcal{F}_N \cdot \mathcal{F}$ and $H_N := F_N$. $\quad\square$

Lemma 3.13. *Let $g_1 \colon \mathcal{X} \to \mathbb{R}$ and $g_2 \colon \mathcal{Y} \to \mathbb{R}$ be a measurable functions. It holds*

$$\mathbf{E}\left[\frac{T}{\pi(X)} g_1(X)\right] = \mathbf{E}\left[g_1(X)\right] \,.$$

If Assumption 6 holds true, then

$$\mathbf{E}\left[\frac{T}{\pi(X)} g_2(Y(T))\right] = \mathbf{E}\left[f(Y(1))\right] \,.$$

Lemma 3.14. *Under conditions it holds* $\sup_{z \in \mathbb{R}} |R_3(z)| \xrightarrow{\mathbf{P}} 0$.

Proof. Let $N \geq \underline{N}$, $z \in \mathbb{R}$, and let $g^\dagger$ denote the function (3.23) with $(\lambda^\dagger, \lambda_0^\dagger)$. If

$$\left| w^\dagger(X) - \frac{1}{\pi(X)} \right| \leq \varepsilon_N$$

it holds

$$g^\dagger(X) \cdot f_z(T, X, Y(T)) = \left(w^\dagger(X) - \frac{1}{\pi(X)}\right) T \left(\mathbf{1}_{\{Y(T) \leq z\}} - F_{Y(1)}(z|X)\right) \,.$$

By Lemma 3.8 it holds

$$f_z(T, X, Y(T)) \in L^1(\mathbf{P}),$$
$$f_z(T, X, Y(T)) \perp D_N,$$
$$\mathbf{E}[f_z(T, X, Y(T))|X] = 0.$$

Thus, it follows from Lemma 3.4

$$\mathbf{E}\left[w^\dagger(X) \cdot f_z(T, X, Y(T))\right] = 0.$$

Since

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

by Lemma 3.13, it follows

$$\mathbf{E}\left[g^\dagger(X) f_z(T, X, Y(T))\right]$$
$$= \mathbf{E}\left[w^\dagger(X) \cdot f_z(T, X, Y(T))\right] - \mathbf{E}\left[\frac{T}{\pi(X)} f_z(T, X, Y(T))\right] = 0.$$

But then

$$R_3(z) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N}\left[\left(w(X_i) - \frac{1}{\pi(X_i)}\right) T_i \left(\mathbf{1}_{\{Y_i \leq z\}} - F_{Y(1)}(z|X_i)\right)\right] = \mathbb{G}_N\left(g^\dagger \cdot f_z\right).$$

It follows

$$\mathbf{P}\left[\sup_{z \in \mathbb{R}} |R_3(z)| \geq \varepsilon\right] \leq \mathbf{P}\left[\sup_{z \in \mathbb{R}} |R_3(z)| \geq \varepsilon \text{ and } |w^\dagger(X) - 1/\pi(X)| \leq \varepsilon_N\right]$$
$$+ \mathbf{P}\left[|w^\dagger(X) - 1/\pi(X)| > \varepsilon_N\right]$$
$$\leq \mathbf{P}\left[\|\mathbb{G}_N\|_{\mathcal{F}_N \cdot \mathcal{F}}^* \geq \varepsilon\right] + \mathbf{P}\left[|w^\dagger(X) - 1/\pi(X)| > \varepsilon_N\right]$$
$$\to 0.$$

The convergence of the first term follows from Lemma 3.12. The convergence of the second term follows from Theorem 3.1.

$$\square$$

Until now, all parts of the error decomposition converge to 0. The last term R_4 will decide the profile of the limiting process. To this end we need the following concept.

Definition 3.2. *We call a class $\mathcal{F}$ of measurable functions $\mathbf{P}$-Donsker if the sequence of processes $\{G_N f : f \in \mathcal{F}\}$ converges in $l^\infty(\mathcal{F})$ to a tight limit process.*

Theorem 3.4. *Every class $\mathcal{F}$ of measurable functions with $J_{[]}(1, \mathcal{F}, L_2(\mathbf{P})) < \infty$ is*

    $\mathbf{P}$*-Donsker, that is, the sequence of processes $\{G_N f \colon f \in \mathcal{F}\}$ converges in $l^\infty(\mathcal{F})$ to a Gaussian process with mean 0 and covariance function given by*

$$\operatorname{Cov}(f, g) := \mathbf{E}[fg] - \mathbf{E}[f]\mathbf{E}[g]. \qquad (3.37)$$

Proof. [vdV00, Theorem 19.5]   □

Lemma 3.15. *Let $1/\pi(X) \in L^2(\mathbf{P})$. R_4 converges converges in $l^\infty(\mathbb{R})$ to a Gaussian process with mean 0 and covariance*

$\mathbf{Cov}(z_1, z_2)$
$$= \mathbf{E}\left[\frac{F_{Y(1)}(z_1 \wedge z_2 \mid X)}{\pi(X)} - \frac{1 - \pi(X)}{\pi(X)} F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right] - F_{Y(1)}(z_1) \cdot F_{Y(1)}(z_2)$$

Proof. By Lemma 3.8 it follows

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

Thus

$$\begin{aligned}
R_4(z) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{T_i}{\pi(X_i)} \left(\mathbf{1}_{\{Y_i \leq z\}} - F_{Y(1)}(z|X_i)\right) + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{f_z(T_i, X_i, Y_i)}{\pi(X_i)} + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right) \\
&= \mathbb{G}_N \left(\frac{f_z}{\pi(\cdot)} + F_{Y(1)}(z|\cdot) - F_{Y(1)}(z)\right).
\end{aligned}$$

By Lemma 3.9 it holds

$$\begin{aligned}
&\log N_{[]}(\varepsilon, \mathcal{G}, L^2(\mathbf{P})) \\
&\lesssim \log\left(\frac{1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}}{\varepsilon}\right) \lesssim \frac{1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}}{\varepsilon} \qquad \text{for all } \varepsilon \in (0, 1).
\end{aligned}$$

Thus

$$J_{[]}(1, \mathcal{G}, L^2(\mathbf{P})) \lesssim \int_0^1 \sqrt{\frac{1 + \|1/\pi(X)\|_{L^2(\mathbf{P})}}{\varepsilon}} \, d\varepsilon \lesssim 1 + \|1/\pi(X)\|_{L^2(\mathbf{P})} < \infty.$$

But then $\mathcal{G}$ is $\mathbf{P}$-Donsker. 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. We now calculate the covariance of the limiting process.

Covariance

$$\mathbf{E}\left[\left(f_{1/\pi}^{z_1} + F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(f_{1/\pi}^{z_2} + F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]$$

$$= \ \mathbf{E}\left[f_{1/\pi}^{z_1} \cdot f_{1/\pi}^{z_2}\right]$$

$$+ \ \mathbf{E}\left[f_{1/\pi}^{z_1}\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \ + \ \mathbf{E}\left[f_{1/\pi}^{z_2}\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\right]$$

$$+ \ \mathbf{E}\left[\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]$$

$$=: \ C_0 \quad + \quad C_1 + C_2 \quad + \quad C_3\,.$$

It holds

$$C_0 \ = \ \mathbf{E}\left[f_{1/\pi}^{z_1} \cdot f_{1/\pi}^{z_2}\right]$$

$$= \ \mathbf{E}\left[\frac{1}{\pi(X)}\frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T) \le z_1\}} - F_{Y(1)}(z_1|X)\right)\left(\mathbf{1}_{\{Y(T) \le z_2\}} - F_{Y(1)}(z_2|X)\right)\right]$$

$$= \ \mathbf{E}\left[\frac{1}{\pi(X)}\left(\mathbf{1}_{\{Y(1) \le z_1\}} - F_{Y(1)}(z_1|X)\right)\left(\mathbf{1}_{\{Y(1) \le z_2\}} - F_{Y(1)}(z_2|X)\right)\right]$$

$$= \ \mathbf{E}\left[\frac{1}{\pi(X)}\left(F_{Y(1)}(z_1 \wedge z_2|X) \ - \ F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right)\right]\,.$$

$$C_1 \ = \ \mathbf{E}\left[f_{1/\pi}^{z_1}\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]$$

$$= \ \mathbf{E}\left[\frac{T}{\pi(X)}\left(\mathbf{1}_{\{Y(T) \le z_1\}} - F_{Y(1)}(z_1|X)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]$$

$$= \ \mathbf{E}\left[\left(\mathbf{1}_{\{Y(1) \le z_1\}} - F_{Y(1)}(z_1|X)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]$$

$$= \ 0\,.$$

In the same way we see $C_2 = 0\,.$

$$C_3 \ = \ \mathbf{E}\left[\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]$$

$$= \ \mathbf{E}\left[F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right] \ - \ F_{Y(1)}(z_1) \cdot F_{Y(1)}(z_2)\,.$$

Adding up the results gives us (3.25). □

We have gathered all the results to prove Theorem 3.2.

Proof. *(Theorem 3.2)* We connect the statement of the theorem to the error decomposition by Lemma 3.5. By Lemma 3.6, Lemma 3.7, Lemma 3.14 it follows $\sup_{z \in \mathbb{R}} |R_i(z)| \xrightarrow{\mathbf{P}} 0$ for $i = 1, 2, 3$. Thus, by Slutzky's theorem (cf. [Kle20, Theorem 13.18]) the behaviour of the limiting process is the one of Lemma 3.15. □

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

My Contribution

I present the relevant facts from Convex analysis. I prove some results that I did not find in the literature, but likely are folklore.

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 Ω. 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,

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

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

Proof. [Roc70, Theorem 11.3] $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

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) $\text{ri}(C) \neq \emptyset$ *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* $\text{ri}\left(\bigcap_{i \in I} C_i\right) = \bigcap_{i \in I} \text{ri}(C_i)$.

(v) *Let* $L : \mathbb{R}^n \to \mathbb{R}^m$ *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

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

Using (iii) again, we get $\mathrm{ri}(C_1 \times C_2) \subseteq \mathrm{ri}\,C_1 \times \mathrm{ri}\,C_2$. Suppose $(z_1, z_2) \in \mathrm{ri}\,C_1 \times \mathrm{ri}\,C_2$. 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. $\qquad\square$

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

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

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

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

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

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

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

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

$$\Theta_1 \; := \; C_1 \times [0, \infty) \,,$$
$$\Theta_2(x^*) \; := \; \{(x, \lambda) \in \mathbb{R}^n \; : \; x \in C_2 \text{ and } \lambda \le \langle x^*, x \rangle \; - \; \sigma_{C_1 \cap C_2}(x^*) \}$$

can by properly separated.

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

Convexity of Θ_1 and $\Theta_2(x^*)$

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

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

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

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

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

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

We start by calculating the relative interiors. It holds

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

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

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

Then it follows

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

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

Applying Theorem 4.1 finishes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = (\sigma_{C_1} \square \, \sigma_{C_2})(x^*) \qquad \text{for all } x^* \in \mathbb{R}^n. \tag{4.10}$$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = \sigma_{C_1}(x_1^*) + \sigma_{C_2}(x_2^*). \tag{4.11}$$

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

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

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. [MMN22, Proposition 4.2]). Like in differential calculus, there exist sum and chain rule for computing the convex conjugate.

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***

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*) \tag{4.13}$$

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*

$$(f + g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{4.14}$$

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***

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{4.15}$$

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, $dom\, f \neq \emptyset$ 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

$$p_i : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto x_i. \tag{4.16}$$

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

$$p_i^* : \mathbb{R} \to \mathbb{R}^n, \quad x \mapsto e_i \cdot x\,. \tag{4.17}$$

For the inverse image of the adjoint of p_i it holds

$$(p_i^*)^{-1}\left\{(x_1^*, \ldots, x_n^*)\right\} \;=\; \begin{cases} \{x_i^*\}, & \text{if } x_j^* = 0 \text{ for all } j \neq i\,, \\ \emptyset & \text{else.} \end{cases} \tag{4.18}$$

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

$$\left(S_{f,n} \circ p_{\{1,\ldots,n\}}\right)^* (x_1^*, \ldots, x_{n+1}^*) = \begin{cases} S_{f,n}^*(x_1^*, \ldots, x_n^*) & \text{if } x_{n+1}^* = 0\,, \\ \infty & \text{else,} \end{cases} \tag{4.19}$$

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

$$\mathrm{ri}\left(\mathrm{dom}\left(S_{f,n} \circ p_{\{1,\dots,n\}}\right)\right) \ \cap \ \mathrm{ri}\left(\mathrm{dom}\, f_{n+1}\right)$$

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

By the conjugate sum rule it follows

$$(S_{f,n+1})^* = \left(S_{f,n} \circ p_{\{1,\dots,n\}} + f_{n+1}\right)^* = \left(S_{f,n} \circ p_{\{1,\dots,n\}}\right)^* \square f_{n+1}^*$$

$$= S_{f,n}^* \circ p_{\{1,\dots,n\}} + f_{n+1}^* = S_{f,n+1}^*\,.$$

$\Diamond$

4.2 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 **??**).

Theorem 4.4. *Consider the optimization problem*

$$\underset{w\in\mathbb{R}^n}{\text{minimize}} \qquad\qquad f(w) \qquad\qquad\qquad (4.20)$$

$$\text{subject to} \qquad\qquad \mathbf{U}w \ \geq \ d\,.$$

$$\mathbf{A}w \ = \ a\,,$$

and its dual problem

$$\underset{\lambda_d\in\mathbb{R}^r,\lambda_a\in\mathbb{R}^s}{\text{maximize}} \qquad \langle\lambda_d,d\rangle \ + \ \langle\lambda_a,a\rangle \ - \ f^*\left(\mathbf{U}^\top\lambda_d + \mathbf{A}^\top\lambda_a\right) \qquad (4.21)$$

$$\text{subject to} \qquad \lambda_d \ \geq \ 0\,.$$

Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ be an optimal solution to (4.21). If the objective function f of (4.20) is strictly convex and its convex conjugate f^* is continuously differentiable, then the unique optimal solution to (4.20) is given by

$$w^\dagger = \nabla f^*\left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger\right). \tag{4.22}$$

Plan of Proof

We show that $w^\dagger$ and $(\lambda_d^\dagger, \lambda_a^\dagger)$ meet the Karush-Kuhn-Tucker conditions for 4.20, that is, **complementary slackness**

$$\langle \lambda_d^\dagger, d - \mathbf{U}w^\dagger \rangle \;=\; 0\,, \tag{4.23}$$

primal and **dual feasibility**

$$\mathbf{U}w^\dagger \;\geq\; d\,, \tag{4.24}$$
$$\mathbf{A}w^\dagger \;=\; a\,,$$
$$\lambda_d^\dagger \;\geq\; 0\,, \tag{4.25}$$

and **stationarity**

$$0_n \;\in\; [\partial f(w^\dagger) \;+\; \partial\left(w \mapsto d - \mathbf{U}w\right)(w^\dagger) \cdot \lambda_d^\dagger \;+\; \partial\left(w \mapsto a - \mathbf{A}w\right)(w^\dagger) \cdot \lambda_a^\dagger]\,. \tag{4.26}$$

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 4.2. [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 (4.21).

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 (4.23) we will show for all coordinates

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

It is well know that a concave functions g satisfies

$$g(x) - g(y) \geq \nabla g(x)^\top (x - y) \qquad \text{for all } x, y \,. \tag{4.27}$$

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

First, we show

$$\nabla G_i(\lambda_d^\dagger) \ \leq \ 0 \qquad \text{for all } i \in \{1, \ldots, s\} \,. \tag{4.28}$$

Assume towards a contradiction that $\nabla G_i(\lambda_d^\dagger) > 0$ for some $i \in \{1, \ldots, s\}$. By the continuity of ∇G there exists $\varepsilon > 0$ such that $\nabla G_i(\lambda_d^\dagger + e_i \cdot \varepsilon) > 0$. It follows from (4.27)

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

which contradicts the optimality of $\lambda_d^\dagger$ for (4.21). It follows (4.28).

Next, we assume that $\lambda_{d,i}^\dagger > 0$ and $\nabla G_i(\lambda_d^\dagger) < 0$ for some $i \in \{1, \ldots, s\}$. 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 (4.23), 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 (4.28), $w^\dagger$ satisfies the inequality constraints in (4.20). 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 4.2 and the differentiability of g it holds

$$a \ \in \ \partial g(\lambda_a^\dagger) \ = \ \left\{ \mathbf{A} \cdot \nabla f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \right) \right\} \ = \ \left\{ \mathbf{A} w^\dagger \right\}. \tag{4.29}$$

Thus $a = \mathbf{A} w^\dagger$. But then $w^\dagger$ satisfies also the equality constraints. We proved (4.24).

Stationarity

First we show

$$\mathbf{U}^\top \lambda_d^\dagger \ + \ \mathbf{A}^\top \lambda_a^\dagger \ \in \ \partial f(w^\dagger). \tag{4.30}$$

By Proposition 4.2 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 (4.30). Next we show

$$-\mathbf{U}^\top \in \ \partial \left(w \mapsto d - \mathbf{U} w \right)(w^\dagger) \qquad \text{and} \qquad -\mathbf{A}^\top \in \ \partial \left(w \mapsto d - \mathbf{A} w \right)(w^\dagger). \tag{4.31}$$

To this end, note that

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

Thus $-\mathbf{U}^\top \in \partial \left(w \mapsto d - \mathbf{U} w \right)(w^\dagger)$. In the same way it follows $-\mathbf{A}^\top \in \partial \left(w \mapsto d - \mathbf{A} w \right)(w^\dagger)$. From (4.30) and (4.31) we conclude

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

We have proved (4.26), that is, stationarity.

Dual Feasibility and Conclusion

Dual feasibility (4.25) follows immediately from the optimality of $\lambda_d^\dagger$ for (4.21). Thus, $(\lambda_d^\dagger, \lambda_a^\dagger)$ and $w^\dagger$ satisfy the Karush-Kuhn-Tucker conditions for (4.20). Applying [Roc70, Theorem 28.3] finishes the proof. $\qquad\square$

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.

[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.

[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.

[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.

Bibliography

[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.