A Novel Weighted Mean Approach to Estimate the Distribution Function in Observational Studies

Asymptotic Analysis

Ioan Scheffel

A thesis presented for the degree of
Master of Science Mathematics

supervised by PD Dr. Jürgen Dippon
Institute for Stochastics and Applications
Faculty 8: Mathematics and Physics
University of Stuttgart
submitted at

Eigenständigkeitserklärung

Ich erkläre mit meiner Unterschrift, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen dieser Arbeit, die dem Wortlaut, dem Sinn oder der Argumentation nach anderen Werken entnommen sind (einschließlich des World Wide Web und anderer elektronischer Text- und Datensammlungen), habe ich unter Angabe der Quellen vollständig kenntlich gemacht.

Abstract

english

This thesis is a master thesis.

Abstract

german

Dies ist ein Master Arbeit.

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

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

We show in Lemma 6.3, 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 $n \leq N$ 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

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

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 The Optimization Problem behind the Weights

There are different ways to generate weights for covariate balance. We discussed this in the introduction. Now, we introduce the balancing weights framework of [WZ19]. It is a (generic) convex optimization problem that enforces covariate balance by constraints on the search space. Similar to classical propensity score estimates, it only extracts from the data information about treatment status and individual characteristics. It ignores the outcome. This gives the additional option to balance covariates before observing outcomes.

The primary optimization task is to minimize an objective function over a predefined search space. From a practical point of view, the objective function instils additional goodness in the weights, for example, low sample variance [Zub15, Introduction]. More important, however, are the constraints that enforce covariate balance. Both objective function and design of the constraints distinguish the method.

From a mathematical point of view, we have slightly different requirements. The proofs should be as clear an short as possible. The mathematical objects involved should help with that (or at least not prevent it). Therefore, we focus on theoretical properties of the method. As a by-product, we create new ideas that wait for testing in practice.

The notion of what to balance is defined by the basis functions of the covariates. It is common that they form a regression basis. We shall use this flexibility to introduce basis functions with random design to this framework. This makes the proofs easier.

2.1 Introduction

Let $(T_1, X_1), \ldots, (T_N, X_N)$ be independent and identically-distributed copies of T and X (see Introduction page). We gather them in the (random) data set

$$D_N \;:=\; \{\, (T_i, X_i) \;:\; i \in \{1, \ldots, N\} \,\}\;.$$

Furthermore, let

$$n \;:=\; \#\{i \in \{1,\dots,N\} \;:\; T_i = 1\}$$

be the number of treated units. This is a random variable. We assume the order $T_i = 1$ for all $i \leq n$. For a

$$\text{(proper) convex function} \qquad \varphi \;:\; \mathbb{R} \;\to\; \mathbb{R} \cup \{+\infty\}\;,$$

a vector of N basis functions of the covariates

$$B \;:=\; [B_1,\dots,B_N]^\top \qquad \text{with} \qquad B_k \;:\; \mathbb{R}^d \;\to\; \mathbb{R} \qquad \text{for all } k \in \{1,\dots,N\}\;,$$

and a (random) constraints vector

$$\delta \;:=\; [\delta_1,\dots,\delta_N]^\top \qquad \text{with} \qquad \delta_k \;:\; (\Omega, \sigma(D_N), \mathbf{P}) \;\to\; \mathbb{R} \qquad \text{for all } k \in \{1,\dots,N\}\;,$$

we consider the (random) convex optimization problem

Problem 1.

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

What is random in Problem 1? First, the dimension of the search space ($w \in \mathbb{R}^n$) depends on the random variable n. Thus, we only compute weights for the treated units (the ones with $T_i = 1$). Next, consider the **objective function**

$$w \;\mapsto\; \sum_{i=1}^{n}\varphi(w_i)\,.$$

The number of summands is random (again n). Note, that sometimes we use the equivalent notation

$$w \;\mapsto\; \sum_{i=1}^{N} T_i\cdot\varphi(w_i)\,,$$

where we set the weights of the untreated (the ones with $T_i = 0$) to some arbitrary value in the domain of φ. Let's consider the **constraints**. There is no randomness in the first two constraints.

$$w_i \geq 0 \qquad \text{for all } i\in\{1,\ldots,n\} \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{n} w_i = 1\,.$$

They only make sure, that the weights (divided by N) form a convex combination. If, for example, the outcome space $\mathcal{Y}$ is convex we make sure that a weighted-mean-estimate of $\mathbf{E}[Y(1)]$ satisfies

$$\widehat{Y}(1) \;:=\; \frac{1}{N}\sum_{i=1}^{n} w_i\cdot Y_i \;\in\; \mathcal{Y}$$

or that a weighted-mean-estimate of the distribution function of $Y(1)$ satisfies

$$\widehat{F}_{Y(1)} \;:=\; \frac{1}{N}\sum_{i=1}^{n} w_i\cdot \mathbf{1}\{Y_i\leq z\} \;\in\; [0,1]\,.$$

We talked about the covariate balancing constraint in the introduction (we shall call them the **box constraints**, because of the absolute value).

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

They are crucial - we shall discuss their implications as the analysis unfolds. For now, note that the number of summands in

$$\sum_{i=1}^{n} w_i \cdot B_k(X_i)$$

is random again, and sometimes we switch to

$$\sum_{i=1}^{N} T_i \cdot w_i \cdot B_k(X_i) \, .$$

In section?, we shall specify the vector of basis functions B. Instead of sieve estimators as in [WZ19], where the number of basis functions grows slower than N to ∞ and the basis functions have fixed design, we shall choose the basis of partitioning estimates as in [GKKW02, §4], which depends on the whole data set D_N and therefore has random design. We shall see that this choice greatly simplifies the consistency proofs. Finally, note that [WZ19, Algorithm 1 on page 11] is a (random) algorithm to specify δ based on D_N.

2.2 Objective Function

The formulation of Problem 1 allows for great flexibility. To obtain clear and short proofs, however, we have to restrict it.

Definition 2.1. We define φ in Problem 1 by

$$\varphi \; : \; \mathbb{R} \; \to \; [0, \infty) \, , \qquad x \mapsto (x-1)^2 \, .$$

Remark. If we plug this choice in Problem 1, we observe

$$\sum_{i=1}^{n} \varphi(w_i) \; = \; \sum_{i=1}^{N} T_i \left(T_i \cdot w_i - 1 \right)^2 \; = \; \sum_{i=1}^{N} T_i \left(T_i \cdot w_i - \frac{1}{N} \sum_{i=1}^{N} T_i \cdot w_i \right)^2 \, .$$

Thus Problem 1 minimizes the sample variance of the weights $(T_i \cdot w_i)$. This is in line with the objective function in [Zub15]. ♢

Next, we derive theoretical properties of φ that we will use in the subsequent analysis.

Lemma 2.1. *The function* φ *of Definition 2.1 satisfies*

(i) φ *is strictly convex and continuously differentiable on* $\mathbb{R}$*, with derivative* φ'

(ii) The inverse of the derivative $(\varphi')^{-1}$ *exists and is continuously differentiable*

(iii) Both φ' *and* $(\varphi')^{-1}$ *are uniformly continuous*

Proof. The proof is easy. We omit the details. □

The next lemma prepares a link to the assumptions of Theorem 7.4.

Lemma 2.2. *The convex conjugate of* φ *(see* (7.12)*) is*

$$\varphi^* : \mathbb{R} \;\to\; \mathbb{R}\,, \quad x^* \;\mapsto\; x^* \cdot (\varphi')^{-1}(x^*) \;-\; \varphi\Big((\varphi')^{-1}(x^*)\Big)\,.$$

Furthermore, φ^* *is strictly convex and continuously differentiable on* $\mathbb{R}$*.*

Proof. We define

$$\phi : \mathbb{R} \times \mathbb{R} \;\to\; \mathbb{R}\,, \quad (x, x^*) \;\mapsto\; x \cdot x^* \;-\; \varphi(x)\,.$$

Let $x^* \in \mathbb{R}$. By Lemma 2.1.*(i)*, φ is continuously differentiable on $\mathbb{R}$ with derivative φ'. The same holds for $\phi(\cdot, x^*)$ with derivative satisfying

$$\frac{\partial}{\partial x}\phi(x, x^*) \;=\; x^* - \varphi'(x) \qquad \text{for all } x \in \mathbb{R}\,.$$

By Lemma 2.1.*(ii)*, it holds that

$$z \;:=\; (\varphi')^{-1}(x^*)$$

is an extreme point of $\phi(\cdot, x^*)$. Since φ is strictly convex by Lemma 2.1.*(i)*, $\phi(\cdot, x^*)$ is strictly concave. Thus, z is the unique maximum point of $\phi(\cdot, x^*)$ on $\mathbb{R}$. Thus

$$\begin{aligned}
\varphi^*(x^*) \;&=\; \sup_{x \in \mathbb{R}} x \cdot x^* - \varphi(x) \;=\; \sup_{x \in \mathbb{R}} \phi(x, x^*) \\
&=\; \phi(z, x^*) \\
&=\; x^* \cdot (\varphi')^{-1}(x^*) \;-\; \varphi\Big((\varphi')^{-1}(x^*)\Big) \qquad \text{for all } x^* \in \mathbb{R}\,.
\end{aligned}$$

Now we proof the second statement. Since $(\varphi')^{-1}$ is continuously differentiable by Lemma 2.1.*(ii)*, it holds

$$\begin{aligned}\frac{\partial}{\partial x^*}\varphi^*(x^*) &= (\varphi')^{-1}(x^*) + x^* \cdot \frac{\partial}{\partial x^*}(\varphi')^{-1}(x^*) - \varphi'\left((\varphi')^{-1}(x^*)\right) \cdot \frac{\partial}{\partial x^*}(\varphi')^{-1}(x^*) \\ &= (\varphi')^{-1}(x^*) + x^* \cdot \frac{\partial}{\partial x^*}(\varphi')^{-1}(x^*) - x^* \cdot \frac{\partial}{\partial x^*}(\varphi')^{-1}(x^*) \\ &= (\varphi')^{-1}(x^*) \qquad \text{for all } x^* \in \mathbb{R}\,.\end{aligned} \tag{2.1}$$

Since φ is strictly convex and continuously differentiable, φ' is continuous and strictly non-decreasing. Thus $(\varphi')^{-1}$ is continuous and strictly non-decreasing. It follows from (2.1) that φ^* is strictly convex and continuously differentiable. □

With Lemma 2.2 we are ready to complete the link.

Lemma 2.3. *The function*

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

satisfies Assumption 4.

Proof. By Example 7.1 the convex conjugate of Φ is

$$\Phi^* : \mathbb{R}^n \to \mathbb{R}\,, \qquad [\lambda_1, \dots, \lambda_n]^\top \mapsto \sum_{i=1}^{n} \varphi^*(\lambda_i)\,,$$

where φ^* is the convex conjugate of φ. By Lemma 2.1, φ is strictly convex. Thus, Φ is strictly convex. By Lemma 2.2, φ^* continuously differentiable on $\mathbb{R}$. Thus, Φ is continuously differentiable on $\mathbb{R}^n$. It follows the statement of Assumption 4 for Φ. □

Takeaways The choice of Definition 2.1 introduces the sample variance to Problem 1. It has good practical and theoretical properties. Among the theoretical are strict convexity that allows linking Problem 1 to the theory of convex analysis.

2.3 Dual Problem

In the previous section we have expounded our choice of φ - and with it the objective function of Problem 1. Now, we want to apply Theorem 7.4 to Problem 1. To this end, we provide its proper formulation.

Lemma 2.4. *A matrix formulation of Problem 1 is*

$$\begin{aligned} &\underset{w\in\mathbb{R}^n}{\text{minimize}} && \Phi(w) \\ &\text{subject to} && \mathbf{U}w \;\geq\; d\,, \\ &&& \mathbf{A}w \;=\; a\,, \end{aligned} \tag{2.2}$$

with objective function

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

inequality matrix and vector

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

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 that the box constraints of Problem 1 are

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

Put differently, it holds both

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

for all $k\in\{1,\ldots,N\}$. In matrix notation this is

$$\pm\mathbf{B}(\mathbf{X})w \;\geq\; [d_{n+1},\ldots,d_{n+2N}]^\top\,.$$

Proving the rest of the statements is straightforward. We omit the details. □

Remark. The inequality constraints of Lemma 2.4 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 Lemma2.4 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.

I was surprised to find the (exact) same mistake in the earlier paper [CYZ16, page 35 second display]. There is no reference in [WZ19, Proof of Lemma 1] to [CYZ16]. Yet the formulation and the mistake are the same. Did the authors of [WZ19] (inadvertently?) plagiarize the mathematical analysis of [CYZ16] ? ♢

In the next lemma we apply Theorem 7.4 to Problem 1.

Lemma 2.5. *Consider the optimization problem*

$$\begin{aligned} \underset{\substack{\rho\,,\lambda^+,\lambda^-\geq 0\\ \lambda_0\in\mathbb{R}}}{\text{maximize}} \quad & -\sum_{i=1}^{n}\varphi^*\big(\rho_i+\lambda_0+\langle B(X_i),\lambda^+-\lambda^-\rangle\big) \\ & +\sum_{i=1}^{N}\big(\lambda_0+\langle B(X_i),\lambda^+-\lambda^-\rangle\big) \;-\; \langle\delta,\lambda^++\lambda^-\rangle\,. \end{aligned} \tag{2.3}$$

If there exists the optimal solution $(\rho^\dagger,\lambda_0^\dagger,\lambda^{+,\dagger},\lambda^{-,\dagger})$ *then the unique optimal solutions to Problem 1 are*

$$w_i^\dagger \;:=\; (\varphi')^{-1}\Big(\rho_i^\dagger \;+\; \lambda_0^\dagger \;+\; \langle B(X_i),\lambda^{+,\dagger}-\lambda^{-,\dagger}\rangle\Big) \qquad \textit{for all } i\in\{1,\dots,n\}\,.$$

Proof. First, note that by the strict convexity of φ^* (see Lemma 2.2), a solution to Problem (2.3) is unique (if it exists). By Lemma 2.4, Problem 1 has the form required in Theorem 7.4. By Lemma 2.3, the objective function Φ of Problem 1 satisfies Assumption 4. Thus we can apply Theorem 7.4 to Problem 1. Calculations yield the result. □

With the next theorem we merge $\lambda^+\,,\lambda^-\geq 0$ to $\lambda=\lambda^+-\lambda^-\in\mathbb{R}\,.$

Theorem 2.1. *Consider the optimization problem*

$$\begin{aligned} \underset{\substack{\rho\;\in\;\mathbb{R}^N\\ \lambda_0\;\in\;\mathbb{R}\\ \lambda\;\in\;\mathbb{R}^N}}{\text{minimize}} \quad & \frac{1}{N}\sum_{i=1}^{N}\Big[T_i\cdot\varphi^*(\rho_i+\lambda_0+\langle B(X_i),\lambda\rangle) \;-\; \lambda_0-\langle B(X_i),\lambda\rangle\Big] \;+\; \langle\delta,|\lambda|\rangle\,, \\ \text{subject to} \quad & \rho_i\geq 0 \quad \text{for all } i\leq n \qquad \text{and} \qquad \rho_i=0 \quad \text{for all } i>n\,. \end{aligned} \tag{2.4}$$

If there exists the optimal solution $(\rho^\dagger,\lambda_0^\dagger,\lambda^\dagger)$ *then the unique optimal solutions to*

Problem 1 are

$$w_i^\dagger \;:=\; (\varphi')^{-1}\Big(\rho_i^\dagger \;+\; \lambda_0^\dagger \;+\; \langle B(X_i), \lambda^\dagger\rangle\Big) \qquad \textit{for all } i \in \{1,\dots,n\}\;.$$

Proof. Assume that $(\rho^\dagger, \lambda_0^\dagger, \lambda^{+,\dagger}, \lambda^{-,\dagger})$ is an optimal solution to Problem 2.3. We write

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

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

$$\begin{aligned} \text{either} \qquad & \lambda_i^{+,\dagger} > 0 \\ \text{or} \qquad & \lambda_i^{-,\dagger} > 0\,. \end{aligned} \tag{2.5}$$

Assume towards a contradiction that

$$\text{there exists } i \in \{1,\dots,N\} \text{ such that} \qquad \lambda_i^{+,\dagger} > 0 \qquad \text{and} \qquad \lambda_i^{-,\dagger} > 0\,. \tag{2.6}$$

Consider

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

and

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

Since

$$\lambda_i^{\pm,\dagger} \;-\; (\lambda_i^{+,\dagger} \wedge \lambda_i^{-,\dagger}) \;\geq\; 0\,,$$

the perturbed vectors $\tilde\lambda^{\pm,\dagger}$ are in the domain of the optimization problem. By Assumption (2.6) and $\delta > 0$ it follows

$$G\left(\rho^\dagger, \lambda_0^\dagger, \tilde\lambda^{+,\dagger}, \tilde\lambda^{-,\dagger}\right) \;-\; G\left(\rho^\dagger, \lambda_0^\dagger, \lambda^{+,\dagger}, \lambda^{-,\dagger}\right) \;=\; 2 \cdot \delta_i \cdot (\lambda_i^{+,\dagger} \wedge \lambda_i^{-,\dagger}) \;>\; 0\,,$$

which contradicts the optimality of $(\rho^\dagger, \lambda^{+,\dagger}, \lambda^{-,\dagger}, \lambda_0^\dagger)$ (it is supposed to be a maximum in the domain of the optimization problem) . It follows (2.5). But then $\lambda_i^{\pm,\dagger} \geq 0$ collapses to $\lambda_i^\dagger \in \mathbb{R}$ for all $i \in \{0,\dots,N\}$, that is, we set

$$\lambda_i^\dagger \;=\; \lambda_i^{+,\dagger} \;-\; \lambda_i^{-,\dagger} \qquad \text{and} \qquad |\lambda_i^\dagger| \;=\; \lambda_i^{+,\dagger} \;+\; \lambda_i^{-,\dagger}\,.$$

Thus, we can extend the domain of Problem 2.3 to $\lambda \in \mathbb{R}^N$ and update the objective function in the following way (without changing the optimal solution).

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

Multiplying G with $-1/N$ doesn't change the solution either (if we search instead for a minimum). To finish the proof, we choose the notation with T_i instead of n. This extends the domain of ρ to $\mathbb{R}^N_{\geq 0}$, but the new ρ_i are not effective because of $T_i = 0$ for all $i > n$. Thus we may set them to 0. □

Remark. This is the final form of the dual of Problem 1. Since the constraints in the dual problem are elementary, a result such as Lemma 4.1 keeps the initiative going. The dual variables $(\rho, \lambda_0, \lambda)$ are connected to the constraints of Problem 1, that is, $\rho \in \mathbb{R}^N_{\geq 0}$ to $T_i \cdot w_i \geq 0$ for all $i \in \{1, \ldots, N\}$, $\lambda_0 \in \mathbb{R}$ to $\frac{1}{N} \sum_{i=1}^{N} T_i \cdot w_i - 1 = 0$, and $\lambda \in \mathbb{R}^N$ to the N box constraints. ♢

Takeaways We derive a dual formulation of Problem 1 that is easier to analyse. Theorem 2.1 provides a functional relationship of optimal dual solutions and optimal weights.

3 Constructing the Weights Process

In the formulation of Theorem 2.4 we encounter "If there exists the optimal solution $(\rho^\dagger, \lambda_0^\dagger, \lambda)$ ... " . To be able to study asymptotic properties of the weights, we shall assume that Problem 2.4 is feasible, construct a measurable dual solution, and plug it in $(\varphi')^{-1}$. Before we formulate concrete assumptions, we provide tools from functional analysis to obtain measurability. Afterwards, we tailor the feasibility assumptions to the capability of this tools. Then, we interpose a section on basis functions before we construct the weights process - the theoretical analogy of optimal weights.

3.1 Argmax Measurability Theorem

We follow [AB07]. A **correspondence** ψ from a set S_1 to a set S_2 assigns to each $s_1 \in S_1$ a subset $\psi(s_1) \subset S_2$. To clarify that we map s_1 to a set, we use the double arrow, that is, $\psi\colon S_1 \twoheadrightarrow S_2$. Let $(\mathcal{Z}, \Sigma_\mathcal{Z})$ be a measurable space and $\mathcal{S}$ a topological space. We say, that a correspondence $\psi\colon \mathcal{Z} \twoheadrightarrow \mathcal{S}$ is **weakly measurable** , if

$$\{z \in \mathcal{Z} \mid \psi(z) \cap O \neq \emptyset\} \in \Sigma_\mathcal{Z} \qquad \text{for all open subsets } O \subset \mathcal{S}\,.$$

A **selector** from a correspondence $\psi\colon \mathcal{Z} \twoheadrightarrow \mathcal{S}$ is a function $s\colon \mathcal{Z} \to \mathcal{S}$ that satisfies

$$s(z) \in \psi(z) \qquad \text{for all } z \in \mathcal{Z}\,.$$

Definition 3.1. Let $(\mathcal{Z}, \Sigma_\mathcal{Z})$ be a measurable space, and let $\mathcal{S}_1$ and $\mathcal{S}_2$ be topological space. A function $f\colon \mathcal{Z} \times \mathcal{S}_1 \to \mathcal{S}_2$ is a **Caratheodory function** if

$$f(\cdot, s_1)\colon \mathcal{Z} \to \mathcal{S}_2 \qquad \text{is } (\Sigma_\mathcal{Z}, \mathcal{B}(\mathcal{S}_2)) - \textit{measurable} \text{ for all } s_1 \in \mathcal{S}_1\,,$$

and

$$f(z, \cdot)\colon \mathcal{Z} \to \mathcal{S}_2 \qquad \text{is continuous for all } z \in \mathcal{Z}\,.$$

Theorem 3.1. *Let $\mathcal{S}$ be a separable metrizable space and $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$ a measurable space. Let $\psi\colon \mathcal{Z} \twoheadrightarrow \mathcal{S}$ be a weakly measurable correspondence with non-empty compact values, and suppose $f\colon \mathcal{Z} \times \mathcal{S} \to \mathbb{R}$ is a Caratheodory function. Define the value function $m\colon \mathcal{Z} \to \mathbb{R}$ by*

$$m(z) := \max_{s \in \psi(z)} f(z, s)\,,$$

and the correspondence $\mu\colon \mathcal{Z} \twoheadrightarrow \mathcal{S}$ of maximizers by

$$\mu(z) := \{s \in \psi(z) | f(z, s) = m(z)\}\ .$$

Then the value function m is measurable, the argmax correspondence μ has non-empty and compact values, is measurable and admits a measurable selector.

Proof. [AB07, Theorem 18.19] □

3.2 Measurable Dual Solution

Next, we formulate the feasibility assumption. Note that we assume compactness to be able to apply Theorem 3.1.

Assumption 1. *For all $N \in \mathbb{N}$ there exists a compact and deterministic parameter space $\Theta_N \subset \mathbb{R}^N_{\geq 0} \times \mathbb{R} \times \mathbb{R}^N$ such that the optimal solution $\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)$ of Problem 2.4 are contained in Θ_N.*

Based on this assumption it is easy to derive measurability for the dual solutions $\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)$. To this end, we take a closer look on the objective function.

Definition 3.2. We define the (random) objective function of (the maximize version of) Problem 2.4 by

$$G\ :\ \ (\Omega, \sigma(D_N)) \times \left(\mathbb{R}^N_{\geq 0} \times \mathbb{R} \times \mathbb{R}^N\right)\ \to\ \overline{\mathbb{R}}$$

with

$$G(\omega, (\rho, \lambda_0, \lambda))\ =\ \infty \qquad \text{if } \rho_i \neq 0 \text{ for some } i > n\,,$$

and else

$$\begin{aligned} & G(\omega, (\rho, \lambda_0, \lambda)) \\ &= \frac{1}{N} \sum_{i=1}^{N} \Big[T_i(\omega) \cdot \varphi^*(\rho_i + \lambda_0 + \langle B(X_i)(\omega), \lambda \rangle) \; - \; \lambda_0 - \langle B(X_i)(\omega), \lambda \rangle \Big] \\ &\quad + \; \langle \delta(\omega), |\lambda| \rangle \, . \end{aligned}$$

Lemma 3.1. *The function G of Definition 3.2 is Caratheodory.*

Proof. This follows from continuity of φ^* and the measurability of all random variables included. □

In the proof of the lemma gathers the arguments and applies Theorem 3.1.

Lemma 3.2. *For all $N \in \mathbb{N}$ the dual solution*

$$\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger \right) \; : \; \Omega \; \to \; \mathbb{R}_{\geq 0}^N \times \mathbb{R} \times \mathbb{R}^N$$

to Problem 2.4 is

$$\left(\sigma \left((T_i, X_i)_{i \in \{1, \ldots, N\}} \right), \mathcal{B} \left(\mathbb{R}_{\geq 0}^N \times \mathbb{R} \times \mathbb{R}^N \right) \right) - \textit{measurable} \, .$$

Proof. Since Θ_N is deterministic (by Assumption 1) we can define the (constant) correspondence $\omega \mapsto \Theta_N$. Clearly, this is weakly-measurable. Next, we consider the (random) objective function of (the maximize version of) Problem 2.4, that is, $-G$ (see Definition 3.2). By Lemme 3.1, $-G$ is a Caratheodory function. Since $-G$ is also strictly concave, it has a unique argmax in Θ_N. By Assumption 1 this is $\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger \right)$. By Theorem 3.1 this is

$$\left(\sigma(D_N), \mathcal{B} \left(\mathbb{R}_{\geq 0}^N \times \mathbb{R} \times \mathbb{R}^N \right) \right) - \text{measurable} \, .$$

□

3.3 Basis Functions

Going back to the functional relationship of optimal dual solution and optimal weights (see Theorem 2.1), we see that the basis vector of the covariates plays an important

role. Now, we present our choice. To the best of our knowledge, this is a novelty in the framework of balancing weights.

Let $(\mathcal{P}_N)$ denote a sequence of countable, $\mathcal{B}$-measurable partitions

$$\mathcal{P}_N = \{A_{N,1}, A_{N,2}, \ldots\} \subset \mathcal{B}(\mathbb{R}^d)$$

of $\mathbb{R}^d$, that is,

$$A_{N,i} \cap A_{N,j} = \emptyset \qquad \text{if } i \neq j \qquad \text{and} \qquad \bigcap_{i \in \mathbb{N}} A_{N,i} \;=\; \mathbb{R}^d .$$

We define $A_N(x)$ to be the cell of $\mathcal{P}_N$ containing x, that is,

$$A_N \colon \mathbb{R}^d \;\twoheadrightarrow\; \mathbb{R}^d , \qquad x \;\mapsto\; A_N(x) ,$$

where $A_N(x)$ is the only cell containing x.

Lemma 3.3. *The relation*

$$x \sim y \qquad :\Leftrightarrow \qquad x \in A_N(y)$$

is an equivalence relation.

Proof. The proof is simple. We omit it. □

Before we define the basis vector, we assume uniform partition width such that

$$\lambda(A_N) \;=:\; h_N^d \;\to\; 0 \qquad \text{for } N \to \infty .$$

Next, we define the (empirical) basis functions vector

$$B \colon \mathbb{R}^d \times \mathbb{R}^{d \cdot N} \;\to\; \mathbb{R} , \qquad (x, (x_1, \ldots, x_N)) \;\mapsto\; \frac{\big[\mathbf{1}_{A_N(x)}(x_k)\big]_{k \in \{1,\ldots,N\}}}{\sum_{j=1}^N \mathbf{1}_{A_N(x)}(x_j)} , \tag{3.1}$$

where we keep to the convention "$0/0 = 0$". We shall extend B to depend on the random vectors $X, X_1, \ldots, X_N$. The next lemma studies the measurability of the extensions.

Lemma 3.4.

(i) $B(\cdot, (X_1, \ldots, X_N))(\omega)$ *is* $\big(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^N)\big)$*-measurable and constant on each cell* $A_N \in \mathcal{P}_N$ *for all* $\omega \in \Omega$.

(ii) $B(X, (X_1, \ldots, X_N))$ *is* $\big(\sigma(X, D_N), \mathcal{B}(\mathbb{R}^N)\big)$*-measurable.*

Proof. Consider for $k \in \{1, \dots, N\}$ and $\omega \in \Omega$ the indicator function

$$\mathbf{1}_{A_N(X_k(\omega))} \colon \mathbb{R}^d \to \{0, 1\} \,. \tag{3.2}$$

Since $A_N(X_k(\omega)) \in \mathcal{B}(\mathbb{R}^d)$ this is a $\big(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R})\big)$-measurable function. From the definition of B (3.1) it follows the first part of (i). Since the indicator function in (3.2) is 1 if $x \in A_N(X_k(\omega))$ and 0 else, it is also constant on each cell $A_N \in \mathcal{P}_N$. It follows (i). To prove (ii), note that

$$\mathbf{1}_{A_N(X_k(\omega))}(X(\omega)) \;=\; \mathbf{1}_{\bigcup\limits_{i \in \mathbb{N}} \{X, X_k \in A_{N,i}\}}(\omega) \qquad \text{for all } \omega \in \Omega \,,$$

and $\bigcup_{i \in \mathbb{N}} \{X, X_k \in A_{N,i}\} \in \sigma(X, D_N)$. □

Now we gather some useful properties of the (empirical) basis vector.

Lemma 3.5. *Let* $(x, x_1, \dots, x_N) \in \mathbb{R}^{d(N+1)}$.

(i) $\sum_{k=1}^N B_k(x, x_1, \dots, x_N) \in \{0, 1\}$. *In particular,* $x_1, \dots, x_N \notin A_N(x)$ *is equivalent to* $\sum_{k=1}^N B_k(x, x_1, \dots, x_N) = 0$

(ii) $\sum_{k=1}^N B_k(x_i, x_1, \dots, x_N) \;=\; 1 \qquad$ *for all* $i \in \{1, \dots, N\}$.

(iii) $\|B(x, x_1, \dots, x_N)\|_2 \;\leq 1$

(iv) $B_k(x_i, x_1, \dots, x_N) \;=\; B_i(x_k, x_1, \dots, x_N) \qquad$ *for all* $i, k \in \{1, \dots, N\}$

Proof. Let $(x, x_1, \dots, x_N) \in \mathbb{R}^{d(N+1)}$. We prove *(i)*. Then *(ii)* is a direct consequence of *(i)*. If $x_1, \dots, x_N \notin A_N(x)$, then

$$B_k(x, x_1, \dots, x_N) \;=\; \frac{\mathbf{1}_{A_N(x)}(x_k)}{\sum_{j=1}^N \mathbf{1}_{A_N(x)}(x_j)} \;=\; 0 \qquad \text{for all } k \in \{1, \dots, N\} \,.$$

On the other hand, if the sum is 0 it holds

$$\mathbf{1}_{A_N(x)}(x_k) \;=\; 0 \qquad \text{for all } k \in \{1, \dots, N\} \,.$$

It follows the desired equivalence. If

$$\mathbf{1}_{A_N(x)}(x_k) \;=\; 1 \qquad \text{for some } k \in \{1, \dots, N\} \,,$$

then $\sum_{j=1}^N \mathbf{1}_{A_N(x)}(x_j) \geq 1$ and thus " $0/0$ " doesn't occure. It follows

$$\sum_{k=1}^N B_k(x, x_1, \dots, x_N) \;=\; \frac{\sum_{k=1}^N \mathbf{1}_{A_N(x)}(x_k)}{\sum_{j=1}^N \mathbf{1}_{A_N(x)}(x_j)} \;=\; 1 \,.$$

To prove *(iii)*, note that by *(i)*

$$\|B(x, x_1, \dots, x_N)\|_2^2 \;=\; \sum_{k=1}^{N} B_k(x, x_1, \dots, x_N)^2 \;\le\; \sum_{k=1}^{N} B_k(x, x_1, \dots, x_N) \;\le\; 1\,.$$

To prove *(iv)*, note that by Lemma 3.3 and by symmetry and transitivity of the equivalence relation $x \in A_N(y)$ it holds

$$\begin{aligned} B_k(x_i, x_1, \dots, x_N) \;&=\; \frac{\mathbf{1}\{x_k \in A_N(x_i)\}}{\sum_{j=1}^{N} \mathbf{1}\{x_j, x_k \in A_N(x_i)\}} \;=\; \frac{\mathbf{1}\{x_i \in A_N(x_k)\}}{\sum_{j=1}^{N} \mathbf{1}\{x_j \in A_N(x_k)\}} \\ &=\; B_i(x_k, x_1, \dots, x_N)\,. \end{aligned}$$

□

Now we show that the basis vector plays well with uniformly continuous functions. The result seems simple, yet the consequence are great. It allows us later on to specify an oracle parameter instead of assuming its existence (see [WZ19, Assumption 1.6]). This greatly clarifies the proofs.

Lemma 3.6. *Let* $(x, x_1, \dots, x_N) \in \mathbb{R}^{d(N+1)}$*. For all uniformly continuous functions* $f\colon \mathbb{R}^d \to \mathbb{R}$ *it holds*

$$\left|\sum_{k=1}^{N} B_k(x_i, x_1, \dots, x_N) \cdot f(x_k) - f(x_i)\right| \;\le\; \omega\left(f, h_N^d\right) \qquad \text{for all } i \in \{1, \dots, N\}\,,$$

where $\omega(f, \cdot)$ *is the uniform modulus of continuity of* f*.*

Proof. It follows from Lemma 3.5*(ii)*

$$\begin{aligned} &\left|\sum_{k=1}^{N} B_k(x_i, x_1, \dots, x_N) \cdot f(x_k) - f(x_i)\right| \\ &\le\; \left|\sum_{k=1}^{N} B_k(x_i, x_1, \dots, x_N)\,(f(x_k) - f(x_i))\right| \\ &\le\; \sum_{k=1}^{N} B_k(x_i, x_1, \dots, x_N) \cdot \mathbf{1}\{x_k \in A_N(x_i)\}\,|f(x_k) - f(x_i)| \\ &\le\; \omega\left(f, h_N^d\right)\,. \end{aligned}$$

□

Next, we bring forward the applications of Lemma 3.6 that we need.

Lemma 3.7. *Let* $(x, x_1, \ldots, x_N) \in \mathcal{X}^{N+1}$ *. It holds for* $N \to \infty$

(i)

$$\frac{1}{N} \sum_{i,k=1}^{N} \left| B_k(x_i, x_1, \ldots, x_N) \cdot \varphi' \left(\frac{1}{\pi(x_k)} \right) \; - \; \varphi' \left(\frac{1}{\pi(x_i)} \right) \right| \; \to \; 0 \,,$$

(ii)

$$\sqrt{N} \sup_{z \in \mathbb{R}} \max_{i \in \{1,\ldots,N\}} \sum_{k=1}^{N} \left| B_k(x_i, x_1, \ldots, x_N) \cdot F_{Y(1)}(z|x_k) \; - \; F_{Y(1)}(z|x_i) \right| \; \to \; 0 \,.$$

Proof. By Lemma 3.6, the uniform continuity of φ' it holds

$$\frac{1}{N} \sum_{i,k=1}^{N} \left| B_k(x, x_1, \ldots, x_N) \cdot \varphi' \left(\frac{1}{\pi(x_k)} \right) \; - \; \varphi' \left(\frac{1}{\pi(x_i)} \right) \right| \; \leq \; \omega \left(\varphi', h_N^d \right) \; \to \; 0$$

for $N \to \infty$. Likewise

$$\begin{aligned} &\sqrt{N} \sup_{z \in \mathbb{R}} \max_{i \in \{1,\ldots,N\}} \sum_{k=1}^{N} \left| B_k(x_i, x_1, \ldots, x_N) \cdot F_{Y(1)}(z|x_k) \; - \; F_{Y(1)}(z|x_i) \right| \\ &\leq \; \sqrt{N} \sup_{z \in \mathbb{R}} \omega \left(F_{Y(1)}(z|\cdot), h_N^d \right) \; \to \; 0 \qquad \text{for } N \to \infty \,. \end{aligned}$$

□

Remark. We want to comment on the assumption

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

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 $\sqrt{N} h_N^{\alpha \cdot d} \to 0$.

♢

Takeaways Basis functions of non-parametric partitioning estimates are new to the framework of balancing weights. They play well with uniformly continuous functions and promise to simplify the analysis. This choice of basis functions waits to be tested in practice.

3.4 Weights Process

Based on Theorem 2.1 we want to use the dual solution $\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)$ to construct weights. To this end, we define the (empirical) weights function

$$\begin{aligned} w \,:\; & \left(\mathbb{R}^d \times \mathbb{R}^{d\cdot N}\right) \times \left(\mathbb{R}^N_{\geq 0} \times \mathbb{R} \times \mathbb{R}^N\right) \to \mathbb{R}^N \\ & ((x, x_1, \ldots, x_N), (\rho, \lambda_0, \lambda)) \;\mapsto\; \left[(\varphi')^{-1}\left(\rho_i + \lambda_0 + \langle B(x, x_1, \ldots, x_N), \lambda\rangle\right)\right]_{i\in\{1,\ldots,N\}} . \end{aligned}$$

Definition 3.3. Let $\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)$ be the dual solution of Lemma 3.2. We define the weights process $\left\{w^\dagger(x) | x \in \mathbb{R}^d\right\}$ by

$$w^\dagger(x) \;:=\; w\left((x, X_1, \ldots, X_N,), \left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)\right) \qquad \text{for all } x \in \mathbb{R}^d .$$

Lemma 3.8.

(i) $w^\dagger(\cdot)(\omega)$ *is* $\left(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^N)\right)$*-measurable and constant on each cell* $A_N \in \mathcal{P}_N$ *for all* $\omega \in \Omega$ *.*

(ii) $w^\dagger(X)$ *is* $\left(\sigma(X, D_N), \mathcal{B}(\mathbb{R}^N)\right)$*-measurable.*

Proof. This is a direct consequence of Lemme 3.4, Lemma 3.2 and the (assumed) continuity of $(\varphi')^{-1}$. □

Lemma 3.9. *It holds* $w_i^\dagger(X) \in L^\infty(\mathbf{P})$ *for all* $i \in \{1, \ldots, N\}$ *.*

Proof. By Lemma 3.5.*(iii)* it holds

$$\left|\rho_i^\dagger + \lambda_0^\dagger + \langle B(x, x_1, \ldots, x_N), \lambda^\dagger\rangle\right| \;\lesssim\; \left\|\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)\right\|_2 \qquad \text{for all } i \in \{1, \ldots, N\} .$$

Since $\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)$ is contained in the deterministic and compact parameter space Θ_N , it holds

$$\left\|\left(\rho^\dagger, \lambda_0^\dagger, \lambda^\dagger\right)\right\|_2 \in L^\infty(\mathbf{P}) .$$

By the (assumed) uniform continuity of $(\varphi')^{-1}$ on $\mathbb{R}$, it follows $w_i^\dagger(X) \in L^\infty(\mathbf{P})$ for all $i \in \{1, \ldots, N\}$. □

Lemma 3.10. *Let* $Z \in L^1(\mathbf{P})$ *be a random variable that is independent of* $D_N = (T_i, X_i)_{i\in\{1,\dots,N\}}$ *with* $\mathbf{E}[Z \mid X] = 0$ *almost surely. It holds*

$$\mathbf{E}\left[w_i^\dagger(X)\cdot Z\right] = 0 \qquad \text{for all } i \in \{1,\dots,N\}\,.$$

Proof. We write

$$w_i^\dagger(X) \;=\; w^\dagger(X)$$

and ignore the index i. By Lemma 3.9 it holds

$$\left\|w^\dagger(X)\cdot Z\right\|_{L^1(\mathbf{P})} \;\le\; \left\|w^\dagger(X)\right\|_{L^\infty(\mathbf{P})} \|Z\|_{L^1(\mathbf{P})} \;<\; \infty\,. \tag{3.3}$$

By (3.3), $Z \perp D_N$ and $\mathbf{E}[Z \mid X] = 0$ almost surely it holds

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

Note, that $w^\dagger(X)$ is $(\sigma(D_N, X), \mathcal{B}(\mathbb{R}))$-measurable. Thus

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

□

Theorem 3.2. *The weights process satisfies the constraints of Problem 1, that is,*

(i) $T_i\cdot w_i^\dagger(X_i) \;\ge\; 0 \qquad$ *for all* $i \in \{1,\dots,N\}$

(ii) $\frac{1}{N}\sum_{i=1}^N T_i\cdot w_i^\dagger(X_i) \;=\; 1$

(iii) For all $k \in \{1,\dots,N\}$ *it holds*

$$\left|\frac{1}{N}\left(\sum_{i=1}^N T_i\cdot w_i^\dagger(X_i)\cdot B_k(X_i, X_1,\dots,X_N) \;-\; \sum_{i=1}^N B_k(X_i, X_1,\dots,X_N)\right)\right| \;\le\; \delta_k$$

Proof. This follows from Theorem 2.1 and the construction of the weights process. □

4 Consistency of the Weights Process

The goal of this section is to establish consistency of the weights process for the inverse propensity score. To this end, we first show that asymptotically there exists an optimal solution $(\rho^\dagger, \lambda^\dagger, \lambda_0^\dagger)$ to Problem 2.4 that converges to the oracle parameter

$$(0_N, 0, \lambda^*) \qquad \text{where} \qquad \lambda^* \;:=\; \left[\varphi'\left(\frac{1}{\pi(X_k)}\right)\right]_{k\in\{1,\ldots,N\}}$$

in probability (see Theorem). This result justifies Assumption 1. Furthermore, we will identify the dual solution from Lemma 3.2 with the consistent dual solution to derive consistency of the weights process for the inverse propensity score.

4.1 Consistency of the Dual Solution

We get a grip by the following lemma. The high-level idea is that the existence of the optimal dual solution and its proximity to the oracle parameter can be analysed by the objective function.

Lemma 4.1. *Let $m\,, N \in \mathbb{N}$ and let $g : \mathbb{R}^N_{\geq 0} \times \mathbb{R}^m \to \overline{\mathbb{R}}$ be a continuous and proper convex function. Consider*

$$\tilde{S}(\varepsilon) := \left\{(\Delta_\rho, \Delta) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^m \;:\; \|(\Delta_\rho, \Delta)\|_2 = \varepsilon\right\} \qquad \text{for } \varepsilon > 0\,.$$

Then for all $y \in \mathbb{R}^m$ and $\varepsilon > 0$

$$\inf\Big\{g(\Delta_\rho, y+\Delta) - g(0,y) \;:\; (\Delta_\rho, \Delta) \in \tilde{S}(\varepsilon)\Big\} \;\geq\; 0 \tag{4.1}$$

implies the existence of a global minimum

$$(y^*_\rho, y^*) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^m \quad \text{of } g \text{ such that} \quad \big\|(y^*_\rho, y^*) - (0,y)\big\|_2 \leq \varepsilon\,.$$

Proof. We start by defining the convex set

$$\tilde{B}(\varepsilon) \;:=\; \left\{(\Delta_\rho, \Delta) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^m \;:\; \|(\Delta_\rho, \Delta)\|_2 \leq \varepsilon\right\} \qquad \text{for } \varepsilon > 0\,.$$

Then the translation $(0,y)+\tilde{B}(\varepsilon)$ is also convex. Assume towards a contradiction that it holds (4.1) and that there exists

$$(x_\rho^*, x^*) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^m \setminus \Big((0,y)+\tilde{B}(\varepsilon)\Big) \quad \text{such that} \quad g(x_\rho^*, x^*) < g(0,y)\,. \tag{4.2}$$

Since $(0,y)+\tilde{B}(\varepsilon)$ is bounded, the line segment between (x_ρ^*, x^*) and $(0,y)$ crosses its boundary. The boundary consists of two disjoint sets

$$S_0(\varepsilon) := \{(0, y+\Delta) : \Delta \in \mathbb{R}^m \text{ and } \|\Delta\|_2 < \varepsilon\} \qquad \text{and} \qquad \tilde{S}(\varepsilon)\,.$$

Clearly, if the line segment does not cross $\tilde{S}(\varepsilon)$ it leaves $\mathbb{R}^N_{\geq 0} \times \mathbb{R}^m$. But this is not possible. Thus, there exists $(\Delta_\rho, \Delta) \in \tilde{S}(\varepsilon)$ and $\theta \in (0,1)$ such that

$$\theta \cdot (x_\rho^*, x^*) + (1-\theta)\cdot(0,y) = (\Delta_\rho, y+\Delta)\,. \tag{4.3}$$

It follows

$$\begin{aligned} g(0,y) \leq g(\Delta_\rho, y+\Delta) &= g\big(\theta\cdot(x_\rho^*, x^*) + (1-\theta)\cdot(0,y)\big) \\ &\leq \theta\cdot g(x_\rho^*, x^*) + (1-\theta)\cdot g(0,y) < g(0,y)\,, \end{aligned}$$

which is a contradiction. The first inequality is due to (4.1), the equality is due to (4.3), the second inequality is due to the convexity of g, and the strict inequality is due to assumption (4.2). Thus, all values outside $(0,y)+\tilde{B}(\varepsilon)$ are greater or equal $(0,y)$. Since $(0,y)+\tilde{B}(\varepsilon)$ is also compact, the continuous function g has a local minimum

$$(y_\rho^*, y^*) \in (0,y)+\tilde{B}(\varepsilon)\,.$$

But then it holds

$$g(y_\rho^*, y^*) \leq g(0,y) \leq g(x_\rho, x) \qquad \text{for all} \qquad (x_\rho, x) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^m \setminus \Big((0,y)+\tilde{B}(\varepsilon)\Big)$$

and

$$g(y_\rho^*, y^*) \leq g(z_\rho, z) \qquad \text{for all} \qquad (z_\rho, z) \in (0,y)+\tilde{B}(\varepsilon)\,.$$

Thus, (y_ρ^*, y^*) is also a global minimum in $\mathbb{R}^N_{\geq 0} \times \mathbb{R}^m$. Since $(y_\rho^*, y^*) \in (0,y)+\tilde{B}(\varepsilon)$ there exists $(\Delta_\rho, \Delta) \in \tilde{B}(\varepsilon)$ such that

$$(y_\rho^*, y^*) = (\Delta_\rho, y+\Delta) \qquad \text{for some } (\Delta_\rho, \Delta) \in \tilde{B}(\varepsilon)\,.$$

Thus

$$\big\|(y_\rho^*, y^*) - (0,y)\big\|_2 = \|(\Delta_\rho, \Delta)\|_2 \leq \varepsilon\,.$$

This finish the proof. □

Remark. I learned of the high-level idea from [WZ19, page 22]. I adapted it to the needs of the subsequent analysis and provided the details by myself. Note, that the hint in [WZ19, page 22] uses strict inequality in the statement. I found out that this can be relaxed. It is crucial to my further approach that this holds (only) with inequality, because I use measurability properties to obtain convergence. ♢

On the basis of the (random) objective function G of Problem 2.4 (see Definition 3.2) we define, for $\varepsilon > 0$, an auxiliary function

$$\begin{aligned}\underline{\Delta G^*_\varepsilon}\colon\ (\Omega, \sigma(D_N), \mathbf{P}) \ &\to\ \overline{\mathbb{R}} \\ \omega \ &\mapsto\ \inf\left\{G\left(\omega, (\Delta_\rho, \Delta_0, \lambda^*(\omega) + \Delta)\right) - G\left(\omega, (0_N, 0, \lambda^*(\omega))\right) \ :\ \left\|\Delta_\rho, \Delta_0, \Delta\right\|_2 = \varepsilon\right\}\end{aligned}$$

Lemma 4.2. *For all $\varepsilon > 0$ the function $\underline{\Delta G^*_\varepsilon}$ is $\left(\sigma(D_N), \mathcal{B}(\overline{\mathbb{R}})\right)$-measurable.*

Proof. Let $\varepsilon > 0$. By Lemma 3.1, the function

$$\begin{aligned}\Delta G_\varepsilon\colon \Omega \times \left(\mathbb{R}^N \times \left(\mathbb{R}^N_{\geq 0} \times \mathbb{R} \times \mathbb{R}^N\right)\right) \ &\to\ \overline{\mathbb{R}} \\ \left(\omega, (\lambda, (\Delta_\rho \Delta_0 \Delta))\right) \ &\mapsto\ G\left(\omega, (\Delta_\rho, \Delta_0, \lambda + \Delta)\right) - G\left(\omega, (0_N, 0, \lambda)\right)\end{aligned}$$

is Caratheodory. Since $\left\{\left\|\Delta_\rho \Delta_0 \Delta\right\|_2 = \varepsilon\right\}$ is compact in $\mathbb{R}^N_{\geq 0} \times \mathbb{R} \times \mathbb{R}^N$, the function

$$\begin{aligned}\underline{\Delta G_\varepsilon}\colon \Omega \times \mathbb{R}^N \ &\to\ \overline{\mathbb{R}} \\ (\omega, \lambda) \ &\mapsto\ \inf\left\{G\left(\omega, (\Delta_\rho, \Delta_0, \lambda + \Delta)\right) - G\left(\omega, (0_N, 0, \lambda)\right) \ :\ \left\|\Delta_\rho \Delta_0 \Delta\right\|_2 = \varepsilon\right\}\end{aligned}$$

is Caratheodory. Since λ^* is $\left(\sigma(D_N), \mathcal{B}(\mathbb{R}^N)\right)$- measurable it follows the statement. □

Lemma 4.3. *It holds for all $\varepsilon > 0$*

$$\mathbf{P}\left[\underline{\Delta G^*_\varepsilon} \geq 0\right] \ \to\ 1 \qquad \text{for } N \to \infty\,. \tag{4.4}$$

Proof. Let $\varepsilon > 0$ and $\left\|\Delta_\rho, \Delta_0, \Delta\right\|_2 = \varepsilon$. We show

$$\mathbf{P}\left[\underline{\Delta G^*_\varepsilon} \geq -\tilde{\varepsilon}\right] \ \to\ 1 \qquad \text{for } N \to \infty \text{ for all } \tilde{\varepsilon} > 0\,. \tag{4.5}$$

Then the result follows from the measurability of $\underline{\Delta G^*_\varepsilon}$ (see Lemma 4.2). To this end, note, that

$$G(\rho, \lambda_0, \lambda) \ =\ g(\rho, \lambda_0, \lambda) \ +\ \langle \delta, |\lambda| \rangle \qquad \text{for all } (\rho, \lambda_0, \lambda) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R} \times \mathbb{R}^N\,,$$

with

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

Since we assume φ^* to be continuously differentiable (it is always convex), g is a continuously differentiable convex function with gradient

$$\begin{aligned}&(\rho, \lambda_0, \lambda)\\ &\mapsto\; \frac{1}{N}\left(\sum_{i=1}^{N} T_i \cdot (\varphi')^{-1}(\rho_i + \lambda_0 + \langle B(X_i), \lambda\rangle)\left[e_i^\top, 1, B(X_i)^\top\right]^\top \;-\; \left[0_N^\top, 1, B(X_i)^\top\right]^\top\right).\end{aligned}$$

Thus, by (7.9), it holds

$$\begin{aligned}&G\left(\Delta_\rho, \Delta_0, \lambda^* + \Delta\right) \;-\; G\left(0_N, 0, \lambda^*\right)\\
&\geq\; \frac{1}{N}\left(\sum_{i=1}^{N} T_i \cdot (\varphi')^{-1}(\langle B(X_i), \lambda^*\rangle)\left[e_i^\top, 1, B(X_i)^\top\right] \;-\; \left[0_N^\top, 1, B(X_i)^\top\right]\right)\begin{bmatrix}\Delta_\rho\\ \Delta_0\\ \Delta\end{bmatrix}\\
&\quad+\; \langle \delta, |\lambda^* + \Delta| - |\lambda^*|\rangle\\
&\geq\; \frac{1}{N}\sum_{i=1}^{N}\Big(T_i \cdot (\varphi')^{-1}(\langle B(X_i), \lambda^*\rangle) \;-\; 1\Big)\left[e_i^\top, 1, B(X_i)^\top\right]\cdot\begin{bmatrix}\Delta_\rho\\ \Delta_0\\ \Delta\end{bmatrix} + \langle e_i, \Delta_\rho\rangle\\
&\quad+\; \langle \delta, |\lambda^* + \Delta| - |\lambda^*|\rangle\\
&\geq\; -\frac{1}{N}\sum_{i=1}^{N}\left|\Big(T_i \cdot (\varphi')^{-1}(\langle B(X_i), \lambda^*\rangle) \;-\; 1\Big)\left[e_i^\top, 1, B(X_i)^\top\right]\cdot\begin{bmatrix}\Delta_\rho\\ \Delta_0\\ \Delta\end{bmatrix}\right|\\
&\quad-\; \langle \delta, |\Delta|\rangle\\
&=:\; -I_1\\
&\quad-\; I_2\end{aligned}$$

I_1

By the Cauchy-Schwarz inequality, Lemma 3.5.*(iii)* it holds

$$\left|\left[e_i^\top, 1, B(X_i)^\top\right]\cdot\begin{bmatrix}\Delta_\rho\\ \Delta_0\\ \Delta\end{bmatrix}\right| \;\leq\; \|\Delta_\rho, \Delta_0, \Delta\|_2 \;\leq\; \varepsilon\,.$$

Furthermore,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left|\left(T_i\cdot(\varphi^{'})^{-1}(\langle B(X_i),\lambda^*\rangle)\ -\ 1\right)\right| \\
&\leq\ \frac{1}{N}\sum_{i=1}^{N}\left|1-\frac{T_i}{\pi(X_i)}\right| \\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\omega\left((\varphi^{'})^{-1},\left|\sum_{k=1}^{N}B_k(X_i)\cdot\varphi^{'}\left(\frac{1}{\pi(X_k)}\right)-\varphi^{'}\left(\frac{1}{\pi(X_i)}\right)\right|\right) \\
&=:\ J_1 \\
&\quad+\ J_2
\end{aligned}$$

J_1

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

$$\mathbf{E}\left[\left|1-\frac{T}{\pi(X)}\right|\right]\leq 1+\mathbf{E}\left[\frac{T}{\pi(X)}\right]=2\,. \tag{4.6}$$

Thus Etemadi's ($\mathcal{L}_1$ version) strong law of large numbers (cf. [Kle20, Theorem 5.17]) applies to J_1, that is, $J_1 \xrightarrow{\mathbf{P}} 0$.

J_2

By Lemma 3.7.*(i)* and the uniform continuity of $(\varphi^{'})^{-1}$ it holds

$$\begin{aligned}
\omega\left((\varphi^{'})^{-1},\left|\sum_{k=1}^{N}B_k(X_i)\cdot\varphi^{'}\left(\frac{1}{\pi(X_k)}\right)-\varphi^{'}\left(\frac{1}{\pi(X_i)}\right)\right|\right)\ &\leq\ \omega\left((\varphi^{'})^{-1},\omega\left(\varphi^{'},h_N^d\right)\right) \\
&\to\ 0\,.
\end{aligned}$$

Thus $J_2\to 0$.

Conclusion I_1

It follows from the analysis of J_1 and J_2

$$\mathbf{P}\left[I_1\leq\tilde{\varepsilon}\right]\to 1\qquad\text{for all }\tilde{\varepsilon}>0\,.$$

I_2

Since $\delta > 0$ we get

$$\langle \delta, |\Delta| \rangle \;\leq\; \|\delta\|_1 \, \|\Delta\|_\infty \;\leq\; \|\delta\|_1 \, \varepsilon \,,$$

Since $\|\delta\|_1$ converges to 0 in probability we get

$$\mathbf{P}\left[I_2 \leq \tilde{\varepsilon}\right] \;\to\; 1 \qquad \text{for all } \tilde{\varepsilon} > 0 \,.$$

Conclusion

By (4.1) we get

$$\mathbf{P}\left[G\left(\Delta_\rho, \Delta_0, \lambda^* + \Delta\right) \;-\; G\left(0_N, 0, \lambda^*\right) \geq -\tilde{\varepsilon}\right] \;\to\; 1 \qquad \text{for all } \tilde{\varepsilon} > 0 \,.$$

Thus

$$\mathbf{P}\left[\underline{\Delta G^*_\varepsilon} \geq -\tilde{\varepsilon}\right] \;\to\; 1 \qquad \text{for all } \tilde{\varepsilon} > 0 \,.$$

From the measurability of $\underline{\Delta G^*_\varepsilon}$ (see Lemma 4.2) it follows

$$\mathbf{P}\left[\underline{\Delta G^*_\varepsilon} \geq 0\right] \;\to\; 1 \,.$$

□

Theorem 4.1. *With probability going to 1 Problem 2.4 is feasible. Furthermore, if the solution* $(\rho^\dagger, \lambda^\dagger, \lambda_0^\dagger)$ *exists, it converges in probability to* $(0_N, 0, \lambda^*)$ *.*

Proof. By Lemma 4.1 and Lemma 4.3 it holds for all $\varepsilon > 0$

$$\begin{aligned} &\mathbf{P}\left[\text{Problem 2.4 is feasible and } \left\|(\rho^\dagger, \lambda^\dagger, \lambda_0^\dagger) - (0_N, 0, \lambda^*)\right\|_2 \leq \varepsilon\right] \\ &\geq\; \mathbf{P}\left[\underline{\Delta G^*_\varepsilon} \geq 0\right] \;\to\; 1 \end{aligned}$$

□

Corollary 4.1.1. *If Problem 2.4 is feasible it holds* $\left\|(\rho^\dagger, \lambda_0^\dagger)\right\|_2 \xrightarrow{\mathbf{P}} 0$ *and* $\left\|\lambda^* - \lambda^\dagger\right\| \xrightarrow{\mathbf{P}} 0$ *.*

Proof. This follows from Theorem 4.1 and Slutzky's Theorem. □

4.2 Main Result

Theorem 4.2. *If Problem 2.4 is feasible it holds* $w_0^\dagger(X) \xrightarrow{\mathbf{P}} 1/\pi(X)$.

Proof. It holds

$$\left| w_0^\dagger(X) \; - \; \frac{1}{\pi(X)} \right| \; \leq \; \omega \left((\varphi^{'})^{-1}, \left| \sum_{k=1}^{N} B_k(X) \cdot \lambda_k^\dagger - \varphi^{'} \left(\frac{1}{\pi(X)} \right) \right| \right) \; + \; \frac{1}{\pi(X)} \mathbf{1} \bigcap_{k=1}^{N} \{X \neq X_k\} \; .$$

Since

$$\left| \sum_{k=1}^{N} B_k(X) \cdot \lambda_k^\dagger - \varphi^{'} \left(\frac{1}{\pi(X)} \right) \right| \leq \left\| \lambda^* - \lambda^\dagger \right\|_2 + \left| \sum_{k=1}^{N} B_k(X) \cdot \varphi^{'} \left(\frac{1}{\pi(X_k)} \right) - \varphi^{'} \left(\frac{1}{\pi(X)} \right) \right|$$

$$\to 0$$

and

$$\mathbf{P}\left[X \neq X_k \text{ for all } k \in \{1, \ldots, N\} \right] = \mathbf{P}\left[X \neq X_1 \right]^N \to 0$$

by Slutzky's Theorem

$$\frac{1}{\pi(X)} \cdot \mathbf{1} \bigcap_{k=1}^{N} \{X \neq X_k\} \xrightarrow{\mathbf{P}} 0$$

□

5 Convergence of the Weighted Mean

5.1 Tools

For the subsequent analysis we need the theory of empirical processes. For an introduction to empirical processes see [vdV00, §19]. For a thorough treatment see [vdVW13, §2].

5.1.1 Empirical Processes - Definition

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, $(\mathcal{Z}, \Sigma)$ a measurable space, and

$$\xi_1, \dots, \xi_N : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{Z}, \Sigma) \quad \text{independent and identically-distributed}$$

random variables with probability distribution $\mathbf{P}_\xi$. Let $\mathcal{F}$ be a class of measurable functions $f : (\mathcal{Z}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel-σ-algebra on $\mathbb{R}$. Then $\mathcal{F}$ induces a stochastic process by

$$f \;\mapsto\; \mathbb{G}_N f \;:=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \left(f(\xi_i) - \mathbf{E}_\xi[f] \right) , \tag{5.1}$$

where $\mathbf{E}_\xi[f] := \int_{\mathcal{Z}} f \, d\mathbf{P}_\xi$. We call $\mathbb{G}_N$ the **empirical process** indexed by $\mathcal{F}$. The purpose of this construction is, to study the behaviour of a centered, scaled arithmetic mean uniformly over $\mathcal{F}$. To this end, we define the (random) norm

$$\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}_N f| . \tag{5.2}$$

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** $\mathbf{E}^*$ (see [vdVW13, page 6])

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

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 $\mathbf{E}^*[Z] = \mathbf{E}[Z^*]$ (see [vdVW13, Lemma 1.2.1]).

An **envelope function** F of a class $\mathcal{F}$ satisfies

$$|f(z)| \;\leq\; F(z) < \infty \qquad \text{for all } f \in \mathcal{F} \text{ and all } z \in \mathcal{Z} .$$

5.1.2 Bracketing Numbers and Integral

To control empirical processes - apart from strong theorems - we need the notion of bracketing number and integral (see [vdV00, page 270]). Given two functions $\underline{f} \leq \overline{f}$,

$$\text{the bracket} \quad [\underline{f}, \overline{f}] \quad \text{is the set of all functions } f \text{ with} \quad \underline{f} \;\leq\; f \;\leq\; \overline{f}\,.$$

For $\varepsilon > 0$ we define a

$$(\varepsilon, L^r(\mathbf{P}))\text{ -bracket to be a bracket} \quad [\underline{f}, \overline{f}] \quad \text{with} \quad \left\|\overline{f} - \underline{f}\right\|_{L^r(\mathbf{P})} \;<\; \varepsilon\,.$$

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 growth we introduce for $\delta > 0$ the **bracketing integral**

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

Next we give a technical lemma to bound the bracketing numbers of products of two function classes, that is,

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

Lemma 5.1. *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. The proof is simple. We omit the details. □

The following has the advantage of being both example (for the interested reader) and helpful for the subsequent analysis.

For $z \in \mathbb{R}$ we define the function

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

Next we define the function classes

$$\begin{aligned} \mathcal{F} \;&:=\; \{f_z \mid z \in \mathbb{R}\} \\ \mathcal{G} \;&:=\; \left\{ \frac{f_z}{\pi(\cdot)} + F_{Y(1)}(z|\cdot) - F_{Y(1)}(z) :\; z \in \mathbb{R} \right\}. \end{aligned} \tag{5.3}$$

Next, we provide bracketing numbers for these classes.

Lemma 5.2. *The function class* $\mathcal{F}$ *and* $\mathcal{G}$ *defined in* (6.13) *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

$$\mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\right] \leq \varepsilon \qquad \text{for all } l \in \{1, \ldots, m\} \tag{5.4}$$

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

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

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

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

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

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

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

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

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

We used (6.14), $0 \leq T, \pi(X) \leq 1$ and Lemma 6.3. It follows

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

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

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

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

$$\begin{aligned}
&\left\| \frac{T}{\pi(X)} \left(\mathbf{1}\{Y(T) \in [z_{l-1}, z_l]\} + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right) \right\|_{L^2(\mathbf{P})} \\
&\lesssim \left(\mathbf{E}\left[\frac{1}{\pi(X)} \frac{T}{\pi(X)} \left(\mathbf{1}\{Y(T) \in [z_{l-1}, z_l]\} + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right) \right] \right)^{1/2} \\
&\lesssim \left(\mathbf{E}\left[\frac{1}{\pi(X)} \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right] \right] \right)^{1/2} \\
&\lesssim \left(\|1/\pi(X)\|_{L^2(\mathbf{P})} \sqrt{\varepsilon} \right)^{1/2} = \varepsilon^{1/4} \|1/\pi(X)\|_{L^2(\mathbf{P})}^{1/2}
\end{aligned}$$

and

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

Thus

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

As before, it follows

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

□

Before we give another example, we fix some useful properties of f_z.

Lemma 5.3. *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 3 holds, then for all* $z \in \mathbb{R}$

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

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

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

The third equality is due to Assumption 3. □

Next, consider the stochastic process (indexed over $x \in \mathbb{R}^d$)

$$\mathbf{1}\left\{\sup_{y\in A_N(x)}\left|w_0^\dagger(y) - \frac{1}{\pi(y)}\right| \;\le\; \varepsilon_N\right\}\left(w_0^\dagger(x) - \frac{1}{\pi(x)}\right)\cdot \mathbf{1}\bigcup_{k=1}^{N}\{x = X_k\}\,. \tag{5.5}$$

We show, that under mild regularity conditions on the inverse propensity score function all paths of (6.3) are contained in shrinking function classes $(\mathcal{F}_N)$ - and provide bracketing numbers. To be more precise, we need theory from [vdVW13, §2.7.1].

Let for any vector $k \in \mathbb{N}_0^d$ $(d \in \mathbb{R})$

$$D^k \;:=\; \frac{\partial^{\|k\|_1}}{\partial^{k_1}x_1 \cdots \partial^{k_d}x_d}\,,$$

and let $\lfloor a \rfloor$ be the greatest integer smaller than $a > 0$. For $\alpha > 0$, a bounded set $\mathcal{Z} \subset \mathbb{R}^d$ $(d \in \mathbb{N})$ and $M > 0$, we define $C_M^\alpha(\mathcal{Z})$ to be the space of all continuous functions $f\colon \mathcal{Z} \to \mathbb{R}$ with

$$\max_{\|k\|_1\le\alpha}\sup_{x\in\mathcal{Z}}\left|D^k f(x)\right| \;+\; \max_{\|k\|_1=\lfloor\alpha\rfloor}\sup_{x,y}\frac{\left|D^k f(x) - D^k f(y)\right|}{\|x-y\|_2^{\alpha-\lfloor\alpha\rfloor}} \;\le\; M\,.$$

where the suprema in the second term are taken over all x, y in the interior of $\mathcal{Z}$ with $x \neq y$. Furthermore, let

$$\mathcal{Z}^1 := \left\{y \in \mathbb{R}^d\colon\; \|x-y\|_2 < 1 \text{ for some } x \in \mathcal{Z}\right\}\,.$$

Lemma 5.4. *Let $\mathcal{P} = \{A_1, A_2, \ldots\}$ be a partition of $\mathbb{R}^d$ into bounded, convex sets with non-empty interior, and let $\mathcal{F}$ be a class of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that the restrictions $\mathcal{F}_{|A_j}$ belong to $C^{\alpha}_{M_j}(A_j)$ for all $j \in \mathbb{N}$. Then there exists a constant K, depending only on α, V, r and d such that*

$$\log N_{[]}(\varepsilon, \mathcal{F}, L^r(\mathbf{Q})) \leq K \left(\frac{1}{\varepsilon}\right)^V \left(\sum_{j=1}^{\infty} \lambda(A_j^1)^{r/(V+r)} M_j^{Vr/(V+r)} \mathbf{Q}(A_j)^{V/(V+r)}\right)^{(V+r)/r} \tag{5.6}$$

for every $\varepsilon > 0$, $V \geq d/\alpha$, and probability measure $\mathbf{Q}$.

Proof. *[vdVW13, Corollary 2.7.4]* □

5.1.3 Maximal Inequality

In our application we need concentration inequalities for $\|\mathbb{G}_n\|^*_{\mathcal{F}}$. One easy way to obtain this is, to use a maximal inequality (see Theorem 6.2) to control the expectation, together with Markov's inequality. There are also Bernstein-like inequalities for empirical processes (see [vdVW13, §2.14.2]).

Theorem 5.1. (Maximal inequality) *For any class $\mathcal{F}$ of measurable functions with envelope function F,*

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

Proof. [vdV00, Corollary 19.35] □

Lemma 5.5. *Let $(\mathcal{H}_N)$ be a sequence of measurable function classes with envelope functions (H_N). If*

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

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

Proof. By Markov's inequality and Theorem 6.2 it holds for all $\varepsilon > 0$

$$\begin{aligned}\mathbf{P}[\|\mathbb{G}_N\|^*_{\mathcal{H}_N} \geq \varepsilon] &\leq \varepsilon^{-1}\mathbf{E}[\|\mathbb{G}_N\|^*_{\mathcal{H}_N}] = \varepsilon^{-1}\mathbf{E}^*[\|\mathbb{G}_N\|_{\mathcal{H}_N}] \\ &\lesssim \varepsilon^{-1} J_{[]}\left(\|H_N\|_{L^2(\mathbf{P})}, \mathcal{H}_N, L^2(\mathbf{P})\right) \\ &\to 0 \qquad \text{for } N \to \infty\,.\end{aligned}$$

□

Donseker Theorem

There is a powerful theorem - a central limit theorem for $\mathbb{G}_N$ uniform in $\mathcal{F}$ - that we now introduce.

Definition 5.1. We call a class $\mathcal{F}$ of measurable functions $\mathbf{P}$-Donsker if the sequence of processes $\{\mathbb{G}_N f\colon f \in \mathcal{F}\}$ converges in $l^\infty(\mathcal{F})$ to a tight limit process.

Theorem 5.2. *Every class $\mathcal{F}$ of measurable functions with*

$$J_{[]}(1, \mathcal{F}, L_2(\mathbf{P})) < \infty$$

is $\mathbf{P}$-Donsker. Furthermore, 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

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

Proof. [vdV00, Theorem 19.5] □

5.1.4 Propensity Score Weights

The next lemma shows what effect the **propensity score weights** $T/\pi(X)$ have on other functions.

Lemma 5.6. *Let $g_1\colon \mathcal{X} \to \mathbb{R}$ and $g_2\colon \mathcal{Y} \to \mathbb{R}$ be a measurable functions.*

(i) It holds

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

(ii) If Assumption 3 holds true, then

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

5.2 Main Result

Theorem 5.3. *The stochastic process*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} T_i \cdot w_i^{\dagger}(X_i) \cdot \mathbf{1}\left\{Y_i \,\leq\, z\right\} \;-\; F_{Y(1)}(z)\right)_{z\in\mathbb{R}} \tag{5.7}$$

converges in $l^{\infty}(\mathbb{R})$ *to a Gaussian process with mean 0 and covariance function satisfying for all* $z_1, z_2 \in \mathbb{R}$

$$\begin{aligned}&\mathbf{Cov}(z_1, z_2)\\ &= \mathbf{E}\left[\frac{F_{Y(1)}(z_1 \wedge z_2 \,|\, X)}{\pi(X)} \;-\; \frac{1-\pi(X)}{\pi(X)} F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right]\\ &\quad -\; F_{Y(1)}(z_1) \cdot F_{Y(1)}(z_2)\,.\end{aligned} \tag{5.8}$$

5.3 Error Decomposition

Lemma 5.7. *It holds*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} T_i \cdot w_i^{\dagger}(X_i) \cdot \mathbf{1}\left\{Y_i \leq z\right\} \; - \; F_{Y(1)}(z)\right)_{z\in\mathbb{R}} \; = \; R_1 \; + \; R_2 \; + \; R_3 \; + \; R_4 \; + \; R_5 \tag{5.9}$$

with

$$\begin{aligned}
R_1 \; &:= \; \sqrt{N}\sum_{k=1}^{N}\left[\frac{1}{N}\left(\sum_{i=1}^{N} T_i \cdot w_i^{\dagger}(X_i)\cdot B_k(X_i) \; - \; \sum_{i=1}^{N} B_k(X_i)\right)\cdot F_{Y(1)}(z|X_k)\right]_{z\in\mathbb{R}}, \\
R_2 \; &:= \; \sqrt{N}\sum_{i=1}^{N}\frac{1}{N}\left[\left(T_i \cdot w_i^{\dagger}(X_i) \; - \; 1\right)\left(F_{Y(1)}(z|X_i) \; - \; \sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k)\right)\right]_{z\in\mathbb{R}}, \\
R_3 \; &:= \; \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left[T_i \cdot \left(w_i^{\dagger}(X_i) \; - \; w_0^{\dagger}(X_i)\right)\cdot\left(\mathbf{1}\{Y_i \leq z\} \; - \; F_{Y(1)}(z|X_i)\right)\right]\right)_{z\in\mathbb{R}}, \\
R_4 \; &:= \; \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left[T_i \cdot \left(w_0^{\dagger}(X_i) \; - \; \frac{1}{\pi(X_i)}\right)\cdot\left(\mathbf{1}\{Y_i \leq z\} \; - \; F_{Y(1)}(z|X_i)\right)\right]\right)_{z\in\mathbb{R}}, \\
R_5 \; &:= \; \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}\{Y_i \leq z\} - F_{Y(1)}(z|X_i)\right) \; + \; \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right)\right)_{z\in\mathbb{R}}.
\end{aligned}$$

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

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} w_i^\dagger(X_i)\cdot T_i \cdot \mathbf{1}\{Y_i \le z\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(w_i^\dagger(X_i) - w_0^\dagger(X_i) + w_0^\dagger(X_i) - \frac{1}{\pi(X_i)}\right) T_i \cdot \mathbf{1}\{Y_i \le z\} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\mathbf{1}\{Y_i \le z\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(w_i^\dagger(X_i) - w_0^\dagger(X_i) + w_0^\dagger(X_i) - \frac{1}{\pi(X_i)}\right) T_i \left(\mathbf{1}\{Y_i \le z\} - F_{Y(1)}(z|X_i)\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}\{Y_i \le z\} - F_{Y(1)}(z|X_i)\right) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N} w_i^\dagger(X_i)\cdot T_i \cdot F_{Y(1)}(z|X_i) \\
&= (R_3(z) + R_4(z))/\sqrt{N} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}\{Y_i \le z\} - F_{Y(1)}(z|X_i)\right) + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}\Big(w_i^\dagger(X_i)\cdot T_i \;-\; 1\Big) F_{Y(1)}(z|X_i) \\
&\qquad\quad + F_{Y(1)}(z) \\
&= (R_3(z) + R_4(z))/\sqrt{N} \\
&\quad + R_5(z)/\sqrt{N} \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}\Big(w_i^\dagger(X_i)\cdot T_i \;-\; 1\Big)\left(F_{Y(1)}(z|X_i) - \sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k)\right) \\
&\qquad\quad + \frac{1}{N}\sum_{i=1}^{N}\Big(w_i^\dagger(X_i)\cdot T_i \;-\; 1\Big)\sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k) \\
&\qquad\qquad + F_{Y(1)}(z) \\
&= (R_3(z) + R_4(z))/\sqrt{N} \\
&\quad + R_5(z)/\sqrt{N} \\
&\qquad + R_2(z)/\sqrt{N} \\
&\qquad\quad + \sum_{k=1}^{N}\frac{1}{N}\sum_{i=1}^{N}\Big(w_i^\dagger(X_i)\cdot T_i B_k(X_i) \;-\; B_k(X_i)\Big)\cdot F_{Y(1)}(z|X_k) \\
&\qquad\qquad + F_{Y(1)}(z) \\
&= \left(R_3(z) \;+\; R_4(z) \;+\; R_5(z) \;+\; R_2(z) \;+\; R_1(z)\right)/\sqrt{N} \;+\; F_{Y(1)}(z)\,.
\end{aligned}
$$

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

5.4 Analysis of the Terms

5.4.1 R_1

Lemma 5.8. *Let* $\sqrt{N}\,\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$. *Then it holds* $\sup_{z\in\mathbb{R}} |R_1(z)| \xrightarrow{\mathbf{P}} 0$.

Proof. By Theorem 3.2 $(w_i^\dagger(X_i))$ satisfy the box constraints of Problem 1 (in the form with the T_i instead of n). Thus

$$\begin{aligned}
\sup_{z\in\mathbb{R}} |R_1(z)| &= \sqrt{N} \sup_{z\in\mathbb{R}} \sum_{k=1}^{N} \left[\frac{1}{N} \left(\sum_{i=1}^{N} T_i \cdot w_i^\dagger(X_i) \cdot B_k(X_i) \;-\; \sum_{i=1}^{N} B_k(X_i) \right) \cdot F_{Y(1)}(z|X_k) \right] \\
&\leq \sqrt{N} \sum_{k=1}^{N} \left| \frac{1}{N} \left(\sum_{i=1}^{N} T_i \cdot w_i^\dagger(X_i) \cdot B_k(X_i) \;-\; \sum_{i=1}^{N} B_k(X_i) \right) \right| \cdot \sup_{z\in\mathbb{R}} F_{Y(1)}(z|X_k) \\
&\leq \sqrt{N}\,\|\delta\|_1
\end{aligned} \tag{5.10}$$

The last inequality is due to $F_{Y(1)} \in [0,1]$. Since we assume $\sqrt{N}\,\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$ it holds $\sup_{z\in\mathbb{R}} |R_1(z)| \xrightarrow{\mathbf{P}} 0$. □

Remark. We want to comment on the box constraints of Problem 1, that is,

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

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

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

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 (6.11) are

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

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

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

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

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

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 (**??**), where we estimate

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

♢

5.4.2 R_2

Lemma 5.9. *Assume*

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

Then $\sup_{z\in\mathbb{R}} |R_2(z)| \xrightarrow{\mathbf{P}} 0$.

Proof.

$$\sup_{z\in\mathbb{R}} |R_2(z)| \;\le\; \sqrt{N} \sup_{z\in\mathbb{R}} \max_{i\in\{1,\ldots,N\}} \sum_{k=1}^{N} \left|B_k(X_i, X_1, \ldots, X_N)\cdot F_{Y(1)}(z|X_k) \;-\; F_{Y(1)}(z|X_i)\right|$$
$$\cdot\; \frac{1}{N}\sum_{i=1}^{N}\left|T_i \cdot w_i^\dagger(X_i) \;-\; 1\right|$$

Note, that by Theorem 3.2.*(i)-(ii)* it holds

$$\frac{1}{N}\sum_{i=1}^{N}\left|T_i \cdot w_i^\dagger(X_i) \;-\; 1\right| \;\le\; 1 \;+\; \frac{1}{N}\sum_{i=1}^{N} T_i \cdot w_i^\dagger(X_i) \;=\; 2\,.$$

The statement follows from Lemma 3.7.*(ii)* □

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 and R_4. ♢

5.4.3 R_3

Lemma 5.10. *It holds* $R_3 \overset{l^\infty(\mathbb{R})}{\to} 0$.

Proof. Note that

$$|R_3| \;\leq\; \omega\left((\varphi')^{-1}, \left\|(\rho^\dagger, \lambda_0^\dagger)\right\|_2\right) \cdot \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i \cdot \left(\mathbf{1}\{Y_i(T_i) \leq z\} \;-\; F_{Y(1)}(z|X_i)\right)\right)_{z\in\mathbb{R}} \tag{5.11}$$

By Corollary 4.1.1 and the uniform continuity of $(\varphi')^{-1}$ it follows

$$\omega\left((\varphi')^{-1}, \left\|(\rho^\dagger, \lambda_0^\dagger)\right\|_2\right) \overset{\mathbf{P}}{\to} 0\,.$$

By Lemma 5.3 it holds $\mathbf{E}\left[f_z(T, X, Y(T))\right] = 0$. Thus

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} T_i \cdot \left(\mathbf{1}\{Y_i(T_i) \leq z\} \;-\; F_{Y(1)}(z|X_i)\right) \;=\; \mathbb{G}_N f_z$$

Since $\mathcal{F}$ is $\mathbf{P}$-Donsker, by the Donsker Theorem the process $(\mathbb{G}_N f_z)_{z\in\mathbb{R}}$ converges in $l^\infty(\mathbb{R})$. Thus by Slutzky's Theorem, the upper bound in (5.11) converges to 0 in $l^\infty(\mathbb{R})$. It follows the statement. □

5.4.4 R_5

Lemma 5.11. *Let* $1/\pi(X) \in L^2(\mathbf{P})$. R_5 *converges converges in* $l^\infty(\mathbb{R})$ *to a*

Gaussian process with mean 0 and covariance

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

Proof. By Lemma 6.10 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

$$\begin{aligned}R_5(z) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}\{Y_i \le z\} - F_{Y(1)}(z|X_i)\right) + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right)\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{f_z(T_i,X_i,Y_i)}{\pi(X_i)} + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right)\\ &= \mathbb{G}_N\left(\frac{f_z}{\pi(\cdot)} + F_{Y(1)}(z|\cdot) - F_{Y(1)}(z)\right).\end{aligned}$$

By Lemma 6.11 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_5 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

$$\begin{aligned}&\mathbf{E}\left[\left(f^{z_1}_{1/\pi} + F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(f^{z_2}_{1/\pi} + F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]\\ &= \mathbf{E}\left[f^{z_1}_{1/\pi} \cdot f^{z_2}_{1/\pi}\right]\\ &\quad + \mathbf{E}\left[f^{z_1}_{1/\pi}\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] + \mathbf{E}\left[f^{z_2}_{1/\pi}\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\right]\\ &\quad + \mathbf{E}\left[\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right]\\ &=: C_0 \quad + \quad C_1 + C_2 \quad + \quad C_3\,.\end{aligned}$$

It holds

$$\begin{aligned}
C_0 &= \mathbf{E}\left[f_{1/\pi}^{z_1} \cdot f_{1/\pi}^{z_2}\right] \\
&= \mathbf{E}\left[\frac{1}{\pi(X)}\frac{T}{\pi(X)}\left(\mathbf{1}\{Y(T) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(\mathbf{1}\{Y(T) \le z_2\} - F_{Y(1)}(z_2|X)\right)\right] \\
&= \mathbf{E}\left[\frac{1}{\pi(X)}\left(\mathbf{1}\{Y(1) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(\mathbf{1}\{Y(1) \le z_2\} - F_{Y(1)}(z_2|X)\right)\right] \\
&= \mathbf{E}\left[\frac{1}{\pi(X)}\left(F_{Y(1)}(z_1 \wedge z_2|X) \;-\; F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right)\right].
\end{aligned}$$

$$\begin{aligned}
C_1 &= \mathbf{E}\left[f_{1/\pi}^{z_1}\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[\frac{T}{\pi(X)}\left(\mathbf{1}\{Y(T) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[\left(\mathbf{1}\{Y(1) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= 0\,.
\end{aligned}$$

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

$$\begin{aligned}
C_3 &= \mathbf{E}\left[\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right] \;-\; F_{Y(1)}(z_1) \cdot F_{Y(1)}(z_2)\,.
\end{aligned}$$

Adding up the results gives us (6.9). □

6 Discussion and Outlook

Gaussian Bridge

6.0.1 Tools and Assumptions

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, $(\mathcal{Z}, \Sigma)$ a measurable space, and

$$\xi_1, \dots, \xi_N : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{Z}, \Sigma) \quad \text{independent and identically-distributed}$$

random variables with probability distribution $\mathbf{P}_\xi$. Let $\mathcal{F}$ be a class of measurable functions $f : (\mathcal{Z}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel-σ-algebra on $\mathbb{R}$. Then $\mathcal{F}$ induces a stochastic process by

$$f \;\mapsto\; \mathbb{G}_N f \;:=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \left(f(\xi_i) - \mathbf{E}_\xi[f] \right) , \tag{6.1}$$

where $\mathbf{E}_\xi[f] := \int_{\mathcal{Z}} f \, d\mathbf{P}_\xi$. We call $\mathbb{G}_N$ the **empirical process** indexed by $\mathcal{F}$. Often, the purpose of this construction is, to study the behaviour of a centered, scaled arithmetic mean uniformly over $\mathcal{F}$. To this end, we define the (random) norm

$$\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}_N f| . \tag{6.2}$$

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** $\mathbf{E}^*$ (see [vdVW13, page 6])

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

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 $\mathbf{E}^*[Z] = \mathbf{E}[Z^*]$ (see [vdVW13, Lemma 1.2.1]).

To control empirical processes - apart from strong theorems - we need the notion of bracketing number and integral (see [vdV00, page 270]). Given two functions $\underline{f} \leq \overline{f}$,

$$\text{the bracket} \quad [\underline{f}, \overline{f}] \quad \text{is the set of all functions } f \text{ with } \quad \underline{f} \;\leq\; f \;\leq\; \overline{f} .$$

For $\varepsilon > 0$ we define a

$$(\varepsilon, L^r(\mathbf{P}))\text{ -bracket to be a bracket } \quad [\underline{f}, \overline{f}] \quad \text{with} \quad \left\|\overline{f} - \underline{f}\right\|_{L^r(\mathbf{P})} < \varepsilon\,.$$

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 growth we introduce for $\delta > 0$ the **bracketing integral**

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

Before we give some results, we introduce the notion of envelope function. An envelope function F of a class $\mathcal{F}$ satisfies $|f(z)| \leq F(z) < \infty$ for all $f \in \mathcal{F}$ and all $z \in \mathcal{Z}$.

There is a powerful theorem - a central limit theorem for $\mathbb{G}_N$ uniform in $\mathcal{F}$ - that we now introduce.

Definition 6.1. We call a class $\mathcal{F}$ of measurable functions $\mathbf{P}$-Donsker if the sequence of processes $\{\mathbb{G}_N f \colon f \in \mathcal{F}\}$ converges in $l^\infty(\mathcal{F})$ to a tight limit process.

Theorem 6.1. *Every class $\mathcal{F}$ of measurable functions with*

$$J_{[]}(1, \mathcal{F}, L_2(\mathbf{P})) < \infty$$

is $\mathbf{P}$-Donsker. Furthermore, 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

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

Proof. [vdV00, Theorem 19.5] □

In our application we need concentration inequalities for $\|\mathbb{G}_n\|^*_{\mathcal{F}}$. One easy way to obtain this is, to use a maximal inequality (see Theorem 6.2) to control the expectation, together with Markov's inequality. There are also Bernstein-like inequalities for empirical processes (see [vdVW13, §2.14.2]).

Theorem 6.2. (Maximal inequality) *For any class $\mathcal{F}$ of measurable functions with envelope function F,*

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

Proof. [vdV00, Corollary 19.35] □

Lemma 6.1. *Let $(\mathcal{H}_N)$ be a sequence of measurable function classes with envelope functions (H_N). If*

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

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

Proof. By Markov's inequality and Theorem 6.2 it holds for all $\varepsilon > 0$

$$\begin{aligned}\mathbf{P}[\|\mathbb{G}_N\|^*_{\mathcal{H}_N} \geq \varepsilon] &\leq \varepsilon^{-1}\mathbf{E}[\|\mathbb{G}_N\|^*_{\mathcal{H}_N}] = \varepsilon^{-1}\mathbf{E}^*[\|\mathbb{G}_N\|_{\mathcal{H}_N}] \\ &\lesssim \varepsilon^{-1} J_{[]}\left(\|H_N\|_{L^2(\mathbf{P})}, \mathcal{H}_N, L^2(\mathbf{P})\right) \\ &\to 0 \qquad \text{for } N \to \infty\,.\end{aligned}$$

□

Next we give a technical lemma to bound the bracketing numbers of products of two function classes, that is,

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

Lemma 6.2. *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 □

This is as much theory of empirical processes as we need for the moment. The next lemma shows what effect the weights $T/\pi(X)$ have on other functions.

Lemma 6.3. *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 3 holds true, then

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

Lemma 6.4. *Let $(\varepsilon_N) \subset (0, 1]$ be a decreasing sequence with $\varepsilon_N \to 0$ for $N \to \infty$*

and let Assumption 2 hold true for a sequence of function classes $\mathcal{F}_N$ *. Then*

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

Proof. By Assumption 2 and Lemma 6.11 it holds for some $k < 2$

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

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 6.2 to get

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

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 6.1 for $\mathcal{H}_N := \mathcal{F}_N \cdot \mathcal{F}$ and $H_N := F_N$. □

Assumption 2. *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* $k < 2$

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

for all $N \in \mathbb{N}$ *the function*

$$\mathcal{X} \;\to\; \mathbb{R}\,, \qquad x \;\mapsto\; \mathbf{1}\left\{ \sup_{y \in A_N(x)} \left| w^\dagger(y) - \frac{1}{\pi(y)} \right| \;\le\; \varepsilon_N \right\} \left(w^\dagger(x) - \frac{1}{\pi(x)} \right) \tag{6.3}$$

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

The next Lemma and the ensuing examples show, what assumptions on the regularity of the inverse propensity score and the distribution of X imply Assumption 2. To this end, we need notation from [vdVW13, §2.7.1]. To this end, let for any vector $k \in \mathbb{N}_0^d$ ($d \in \mathbb{R}$)

$$D^k := \frac{\partial^{\|k\|_1}}{\partial^{k_1} x_1 \cdots \partial^{k_d} x_d},$$

and let $\lfloor a \rfloor$ be the greatest integer smaller than $a > 0$. For $\alpha > 0$, a bounded set $\mathcal{X} \subset \mathbb{R}^d$ ($d \in \mathbb{N}$) and $M > 0$, we define $C_M^\alpha(\mathcal{X})$ to be the space of all continuous functions $f \colon \mathcal{X} \to \mathbb{R}$ with

$$\max_{\|k\|_1 \leq \alpha} \sup_{x \in \mathcal{X}} \left| D^k f(x) \right| + \max_{\|k\|_1 = \lfloor \alpha \rfloor} \sup_{x,y} \frac{\left| D^k f(x) - D^k f(y) \right|}{\|x - y\|_2^{\alpha - \lfloor \alpha \rfloor}} \leq M.$$

where the suprema in the second term are taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$. Furthermore, let

$$\mathcal{X}^1 := \left\{ y \in \mathbb{R}^d \colon \|x - y\|_2 < 1 \text{ for some } x \in \mathcal{X} \right\}.$$

Lemma 6.5. *Let $\mathcal{P} = \{A_1, A_2, \ldots\}$ be a partition of $\mathbb{R}^d$ into bounded, convex sets with non-empty interior, and let $\mathcal{F}$ be a class of functions $f \colon \mathbb{R}^d \to \mathbb{R}$ such that the restrictions $\mathcal{F}_{|A_j}$ belong to $C_{M_j}^\alpha(A_j)$ for all $j \in \mathbb{N}$. Then there exists a constant K, depending only on α, V, r and d such that*

$$\log N_{[]}(\varepsilon, \mathcal{F}, L^r(\mathbf{Q})) \leq K \left(\frac{1}{\varepsilon}\right)^V \left(\sum_{j=1}^{\infty} \lambda(A_j^1)^{r/(V+r)} M_j^{Vr/(V+r)} \mathbf{Q}(A_j)^{V/(V+r)} \right)^{(V+r)/r} \tag{6.4}$$

for every $\varepsilon > 0$, $V \geq d/\alpha$, and probability measure $\mathbf{Q}$.

Proof. *[vdVW13, Corollary 2.7.4]* □

Lemma 6.6. *Let* $(\mathcal{P}_N)$ *denote a sequence of qubic partitions* $\mathcal{P}_N = \{A_{N,1}, A_{N,2}, \ldots\}$ *of* $\mathbb{R}^d$ *with decreasing width* $(h_N) \subset (0,1]$ *such that* $h_N \to 0$ *for* $N \to \infty$. *Furthermore, 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* $M_{N,j} \geq 1$ *such that*

$$\frac{1}{\pi(\cdot)} \in C^{\alpha}_{M_{N,j}}(A_{N,j}) \quad \textit{and} \quad \sum_{j=1}^{\infty} M_{N,j}^{2V/(V+2)} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \lesssim 1\,. \tag{6.5}$$

It then holds the statement of Assumption 2.

Proof. We want to employ Lemma 6.5. To do this, the crucial observation is, that

$$w^{\dagger}(\omega) \quad \text{is constant on each cell } A_N \in \mathcal{P}_N \text{ for all } \omega \in \Omega\,.$$

Thus, the regularity of the function (6.3) on each cell $A_N \in \mathcal{P}_N$ is decided by $1/\pi(\cdot)$. Indeed, (6.3) is either 0 if the threshold of ε_N is exceeded somewhere in the cell, or has the form constant-minus-smooth-function. In any case, it is continuous and bounded by ε_N. All its derivatives are 0 (if the threshold is exceeded) or are governed by $1/\pi(\cdot)$. Thus, it follows from (6.5)

$$(6.3) \in C^{\alpha}_{M_{N,j}}(A_{N,j}) \quad \text{and} \quad \sum_{j=1}^{\infty} M_{N,j}^{2V/(V+2)} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \lesssim 1\,. \tag{6.6}$$

To bound the right-hand-side in (6.4) we note that $\lambda(A_{N,j}) = h_N^d$ and thus $\lambda(A^1_{N,j}) \lesssim 1$ for all $(j,N) \in \mathbb{N}^2$. Thus

$$\sum_{j=1}^{\infty} \lambda(A^1_{N,j})^{2/(V+2)} M_{N,j}^{2V/(V+2)} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \lesssim 1\,.$$

We get $(6.3) \in \mathcal{F}_N$, where $\mathcal{F}_N$ restricted to $A_{N,j}$ is $C^{\alpha}_{M_{N,j}}(A_{N,j})$ and satisfies the requirements of Lemma 6.5. Since $V = d/\alpha \in (0,2)$ by $\alpha > d/2$, applying Lemma 6.5 finishes the proof. □

Remark. Note, that we only get $L^2(\mathbf{P}_X)$ bracketing numbers in this way. If we assume, that all functions in $\mathcal{F}$ are independent of (T,Y) we readily obtain $L^2(\mathbf{P})$ bracketing numbers. Note, that $w^{\dagger}(X)$ and $1/\pi(X)$ are independent of (T,Y). ♢

In the next examples we show the concrete applications of this Lemma.

Example 6.1. Let $d \in \mathbb{R}$ and assume $\pi(\cdot)$ follows a logistic regression model. Then there exist $(\beta_0, \beta) \in \mathbb{R}^{d+1}$ such that

$$\pi(x) \;=\; \frac{1}{1+\exp\left(-\beta_0 - \langle \beta, x\rangle\right)} \quad \text{and} \quad \frac{1}{\pi(x)} \;=\; 1+\exp\left(-\beta_0 - \langle \beta, x\rangle\right) \quad \text{for all } x \in \mathbb{R}^d\,.$$

Clearly, by the smoothness of the exponential function, for all $\alpha > 0$ there exist $(M_{N,j}) \geq 1$ such that $1/\pi(\cdot) \in C^{\alpha}_{M_{N,j}}(A_{N,j})$. Assume $\#\mathcal{X} < \infty$, that is, X can take only finitely many values with positive probability. We write

$$J_N := \{j \in \mathbb{N} \colon \mathbf{P}[X \in A_{N,j}] > 0\} \ .$$

It holds $\#J_N \leq \#\mathcal{X} < \infty$. Thus, the following maximum is attained

$$\max_{j \in J_N} M_{N,j} =: M_N^* \, .$$

But the partitions increasingly better fit the support of X. Thus M_N^* is decreasing in N, that is, $\infty > M_1^* \geq M_N^*$. It follows

$$\sum_{j=1}^{\infty} M_{N,j}^{2V/(V+2)} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \ \leq \ (M_1^*)^{2V/(V+2)} \cdot \#J_N \ \lesssim \ 1 \, .$$

$\diamondsuit$

The next question is if we can extend this to countable sets $\mathcal{X}$. To focus our attention, we assume (the best case) $M_{N,j} = 1$.

Example 6.2. We give three examples. Let $\propto$ denote the equal-up-to-a-constant order. Without loss of generality we assume $\mathcal{X} = \mathbb{N}$. First consider for some $\varepsilon \in (0,1)$

$$\mathbf{P}[X = k] \propto k^{-(1+\varepsilon)} \, .$$

This (barely) is a distribution. But scaling with $V/(V+2) \in (0,1)$ we cause trouble. Indeed, by $V < 2$ it holds

$$(1+\varepsilon) \cdot \frac{V}{V+2} < 1 \, ,$$

and thus

$$\begin{aligned} \sum_{j=1}^{\infty} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \ &= \ \lim_{N \to \infty} \sum_{j \in J_N} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \\ &= \ \sum_{k=1}^{\infty} \mathbf{P}[X = k]^{V/(V+2)} \\ &\propto \ \sum_{k=1}^{\infty} \left(\frac{1}{k}\right)^{(1+\varepsilon) \cdot V/(V+2)} = \infty \, . \end{aligned}$$

On the other hand, the same arguments applied for

$$\mathbf{P}[X = k] \propto k^{-((V+2)/V+\varepsilon)} \, ,$$

with $\varepsilon > 0$ yield

$$\sum_{j=1}^{\infty} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \lesssim 1\,.$$

Finally, let X be Poisson distributed with parameter $\lambda \in (0,1)$. Then

$$\begin{aligned}\sum_{j=1}^{\infty} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} &= \sum_{k=0}^{\infty} \mathbf{P}[X=k]^{V/(V+2)} \\ &= e^{-\lambda \cdot V/(V+2)} \sum_{k=0}^{\infty} \left(\frac{\lambda^k}{k!}\right)^{V/(V+2)} \\ &\leq \sum_{k=0}^{\infty} \left(\lambda^{V/(V+2)}\right)^k = \frac{1}{1-\lambda^{V/(V+2)}} \lesssim 1\,.\end{aligned}$$

♢

In the next example we show, that

$$\sum_{j=1}^{\infty} \mathbf{P}[X \in A_{N,j}]^{V/(V+2)} \to \infty$$

for all continuous distributions of X.

Example 6.3. Let f_X be the probability density of X. Then there exists a compact set $K \subset \mathcal{X} \subset \mathbb{R}^d$, such that $\inf_{x\in K} f_X(x) > 0$. Since $\mathcal{P}_N$ are cubic partitions, it holds for

$$I_N := \{i \in \mathbb{N}\colon A_{N,i} \subset K\} \qquad \text{that} \qquad \bigcup_{i\in I_N} A_{N,i} \nearrow K\,.$$

Thus

$$\begin{aligned}\sum_{i=1}^{\infty} \mathbf{P}[X \in A_{N,i}]^{V/(V+2)} &\geq \sum_{i\in I_N} \mathbf{P}[X \in A_{N,i}]^{V/(V+2)} \\ &\geq \inf_{x\in K} f_X(x)^{V/(V+2)} \cdot h_N^{d\cdot(V/(V+2)-1)} \sum_{i\in I_N} \lambda\,(A_{N,i}) \\ &\to \infty\,.\end{aligned}$$

This follows from $\sum_{i\in I_N} \lambda\,(A_{N,i}) \to \lambda(K) > 0$, $\inf_{x\in K} f_X(x) > 0$, $V/(V+2) - 1 < 0$ and $h_N \to 0$.

♢

Notation

Throughout this section we use the following notation. Let $F_{Y(1)}$ denote the distribution function of $Y(1)$, that is,

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

Let $F_{Y(1)}(\cdot|x)$ denote the distribution function of $Y(1)$ conditional on $X = x \in \mathcal{X}$, that is,

$$F_{Y(1)}(z|x) = \mathbf{P}[Y(1) \leq z \,|\, X = x] \qquad \text{for all } (z,x) \in \mathbb{R} \times \mathcal{X}.$$

We recall from Definition **??** the weight function

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

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

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

We introduced the basis functions (B_k) in ?. Let $\pi(\cdot)$ denote the propensity score function, that is,

$$\pi\colon \mathcal{X} \to [0,1], \qquad x \mapsto \mathbf{P}[T = 1 \,|\, X = x].$$

Let the symbol $\lesssim$ denote the lesser-equal order multiplied by a generic constant $C > 1$ that is independent of N or any other quantitative assumption. We always take C large enough. For example

$$17 \cdot f(x) \lesssim f(x).$$

This helps keeping the complexity of notation at a (necessary) minimum.

Goal of this section

We illustrate the flexibility of the weighted mean estimator by extending the method of [WZ19] to estimates of the distribution function of $Y(1)$, that is, $F_{Y(1)}$. For the asymptotic analysis of estimating the mean $\mathbf{E}[Y(1)]$ see [WZ19, Proof of Theorem 3]. To make this extension, the central observation is, that we can adapt the error decomposition in [WZ19, page 27] to estimates of the distribution function $F_{Y(1)}$ of $Y(1)$. We do this in Lemma 6.7. With this modification, we aim at proving the convergence of

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{n} w^\dagger(X_i)\mathbf{1}\{Y_i(T_i) \leq z\} - F_{Y(1)}(z)\right)_{z\in\mathbb{R}} \tag{6.7}$$

in $l^\infty(\mathbb{R})$ to a Gaussian process with mean 0 and covariance specified in Theorem 6.3. We show, that this is possible under mild assumptions. We discuss this in the next section.

Theorem 6.3. *Under conditions the stochastic process*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{n} w_i^{\dagger}\mathbf{1}_{\{Y_i \le z\}} \;-\; F_{Y(1)}(z)\right)_{z\in\mathbb{R}} \tag{6.8}$$

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

$$\begin{aligned}&\mathbf{Cov}(z_1, z_2)\\ &= \mathbf{E}\left[\frac{F_{Y(1)}(z_1 \wedge z_2 \,|\, X)}{\pi(X)} \;-\; \frac{1-\pi(X)}{\pi(X)} F_{Y(1)}(z_1|X)\cdot F_{Y(1)}(z_2|X)\right]\\ &\quad -\; F_{Y(1)}(z_1)\cdot F_{Y(1)}(z_2)\end{aligned} \tag{6.9}$$

Lemma 6.7. *It holds*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{n} w(X_i)\mathbf{1}\{Y_i \leq z\} \;-\; F_{Y(1)}(z)\right)_{z\in\mathbb{R}} \;=\; R_1 \;+\; R_2 \;+\; R_3 \;+\; R_4 \tag{6.10}$$

with

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

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

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} w^{\dagger}(X_i)\cdot T_i \cdot \mathbf{1}\{Y_i(T_i) \le z\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(w^{\dagger}(X_i) - \frac{1}{\pi(X_i)}\right) T_i \cdot \mathbf{1}\{Y_i(T_i) \le z\} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\mathbf{1}\{Y_i(T_i) \le z\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(w^{\dagger}(X_i) - \frac{1}{\pi(X_i)}\right) T_i \left(\mathbf{1}\{Y_i(T_i) \le z\} - F_{Y(1)}(z|X_i)\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}\{Y_i(T_i) \le z\} - F_{Y(1)}(z|X_i)\right) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N} w^{\dagger}(X_i)\cdot T_i \cdot F_{Y(1)}(z|X_i) \\
&= R_3(z)/\sqrt{N} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(X_i)}\left(\mathbf{1}\{Y_i(T_i) \le z\} - F_{Y(1)}(z|X_i)\right) + \left(F_{Y(1)}(z|X_i) - F_{Y(1)}(z)\right) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}\Big(w^{\dagger}(X_i)\cdot T_i \;-\; 1\Big) F_{Y(1)}(z|X_i) \\
&\qquad\quad + F_{Y(1)}(z) \\
&= R_3(z)/\sqrt{N} \\
&\quad + R_4(z)/\sqrt{N} \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}\Big(w^{\dagger}(X_i)\cdot T_i \;-\; 1\Big)\left(F_{Y(1)}(z|X_i) - \sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k)\right) \\
&\qquad\quad + \frac{1}{N}\sum_{i=1}^{N}\Big(w^{\dagger}(X_i)\cdot T_i \;-\; 1\Big)\sum_{k=1}^{N} B_k(X_i)\cdot F_{Y(1)}(z|X_k) \\
&\qquad\qquad + F_{Y(1)}(z) \\
&= R_3(z)/\sqrt{N} \\
&\quad + R_4(z)/\sqrt{N} \\
&\qquad + R_2(z)/\sqrt{N} \\
&\qquad\quad + \sum_{k=1}^{N}\frac{1}{N}\sum_{i=1}^{N}\Big(w^{\dagger}(X_i)\cdot T_i B_k(X_i) \;-\; B_k(X_i)\Big)\cdot F_{Y(1)}(z|X_k) \\
&\qquad\qquad + F_{Y(1)}(z) \\
&= \left(R_3(z) \;+\; R_4(z) \;+\; R_2(z) \;+\; R_1(z)\right)/\sqrt{N} \;+\; F_{Y(1)}(z)\,.
\end{aligned}
$$

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

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 6.8. *Let the weights function w satisfy the box constraints in Problem 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

$$\begin{aligned}\sup_{z\in\mathbb{R}} |R_1(z)| &= \sup_{z\in\mathbb{R}} \left|\sqrt{N}\sum_{k=1}^{N}\left[\frac{1}{N}\left(\sum_{i=1}^{n} w(X_i)B_k(X_i) - \sum_{i=1}^{N} B_k(X_i)\right) F_{Y(1)}(z|X_k)\right]\right| \\ &\leq \sqrt{N}\sum_{k=1}^{N}\left|\frac{1}{N}\left(\sum_{i=1}^{n} w(X_i)B_k(X_i) - \sum_{i=1}^{N} B_k(X_i)\right)\right| \sup_{z\in\mathbb{R}} F_{Y(1)}(z|X_k) \\ &\leq \sqrt{N}\,\|\delta\|_1\end{aligned} \tag{6.11}$$

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 1. Since we assume $\sqrt{N}\,\|\delta\|_1 \xrightarrow{\mathbf{P}} 0$ it holds $\sup_{z\in\mathbb{R}} |R_1(z)| \xrightarrow{\mathbf{P}} 0$. □

Remark. We want to comment on the box constraints of Problem 1, that is,

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

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

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

Now both sums go over $\{1,\ldots,N\}$ 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 (6.11) are

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

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

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

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

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

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 (**??**), where we estimate

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

♢

Lemma 6.9. *Let the conditions of Theorem* **??** *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*

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

where ω is the modulus of continuity. Then it holds $\sup_{z\in\mathbb{R}} |R_2(z)| \xrightarrow{\mathbf{P}} 0$.

Proof.

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

The equality is due to

$$1 \;=\; \frac{1}{N}\sum_{i=1}^{n} w_i^{\dagger} \;=\; \frac{1}{N}\sum_{i=1}^{n} w(X_i) \;=\; \frac{1}{N}\sum_{i=1}^{N} T_i w(X_i)\,, \tag{6.12}$$

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

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

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 $\sup_{z\in\mathbb{R}} |R_2(z)| \xrightarrow{\mathbf{P}} 0$. □

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. ♢

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

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

and the function classes

$$\begin{aligned} \mathcal{F} &:= \{ f_z \mid z \in \mathbb{R} \} \\ \mathcal{G} &:= \left\{ \frac{f_z}{\pi(\cdot)} + F_{Y(1)}(z|\cdot) - F_{Y(1)}(z) \;:\; z \in \mathbb{R} \right\}. \end{aligned} \tag{6.13}$$

Assumption 3. *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 6.10. *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 3 holds, then for all* $z \in \mathbb{R}$

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

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

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

The third equality is due to Assumption 3. □

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

Lemma 6.11. *The function class* $\mathcal{F}$ *and* $\mathcal{G}$ *defined in* (6.13) *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

$$\mathbf{P}\left[Y(1) \in [z_{l-1}, z_l]\,\right] \;\leq\; \varepsilon \qquad \text{for all } l \in \{1, \ldots, m\} \tag{6.14}$$

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

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

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

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

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

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

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

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

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

We used (6.14), $0 \leq T, \pi(X) \leq 1$ and Lemma 6.3. It follows

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

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

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

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

$$\begin{aligned}
&\left\| \frac{T}{\pi(X)} \left(\mathbf{1}\{Y(T) \in [z_{l-1}, z_l]\} + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right) \right\|_{L^2(\mathbf{P})} \\
&\lesssim \left(\mathbf{E}\left[\frac{1}{\pi(X)} \frac{T}{\pi(X)} \left(\mathbf{1}\{Y(T) \in [z_{l-1}, z_l]\} + \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right]\right) \right] \right)^{1/2} \\
&\lesssim \left(\mathbf{E}\left[\frac{1}{\pi(X)} \mathbf{P}\left[Y(1) \in [z_{l-1}, z_l] \,|\, X\right] \right] \right)^{1/2} \\
&\lesssim \left(\|1/\pi(X)\|_{L^2(\mathbf{P})} \sqrt{\varepsilon} \right)^{1/2} = \varepsilon^{1/4} \|1/\pi(X)\|_{L^2(\mathbf{P})}^{1/2}
\end{aligned}$$

and

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

Thus

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

As before, it follows

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

□

Lemma 6.12. *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 (6.3) 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 6.10 it holds

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

Thus, it follows from Lemma 3.10

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

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 6.3, it follows

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

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

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

The convergence of the first term follows from Lemma 6.4. The convergence of the second term follows from Theorem **??**.

□

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.

Lemma 6.13. *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*

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

Proof. By Lemma 6.10 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

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

By Lemma 6.11 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

$$
\begin{aligned}
&\mathbf{E}\left[\left(f_{1/\pi}^{z_1} + F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(f_{1/\pi}^{z_2} + F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[f_{1/\pi}^{z_1} \cdot f_{1/\pi}^{z_2}\right] \\
&\quad + \mathbf{E}\left[f_{1/\pi}^{z_1}\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] + \mathbf{E}\left[f_{1/\pi}^{z_2}\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\right] \\
&\quad + \mathbf{E}\left[\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&=: C_0 \quad + \quad C_1 + C_2 \quad + \quad C_3\,.
\end{aligned}
$$

It holds

$$
\begin{aligned}
C_0 &= \mathbf{E}\left[f_{1/\pi}^{z_1} \cdot f_{1/\pi}^{z_2}\right] \\
&= \mathbf{E}\left[\frac{1}{\pi(X)}\frac{T}{\pi(X)}\left(\mathbf{1}\{Y(T) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(\mathbf{1}\{Y(T) \le z_2\} - F_{Y(1)}(z_2|X)\right)\right] \\
&= \mathbf{E}\left[\frac{1}{\pi(X)}\left(\mathbf{1}\{Y(1) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(\mathbf{1}\{Y(1) \le z_2\} - F_{Y(1)}(z_2|X)\right)\right] \\
&= \mathbf{E}\left[\frac{1}{\pi(X)}\left(F_{Y(1)}(z_1 \wedge z_2|X) \; - \; F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right)\right]\,.
\end{aligned}
$$

$$
\begin{aligned}
C_1 &= \mathbf{E}\left[f_{1/\pi}^{z_1}\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[\frac{T}{\pi(X)}\left(\mathbf{1}\{Y(T) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[\left(\mathbf{1}\{Y(1) \le z_1\} - F_{Y(1)}(z_1|X)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= 0\,.
\end{aligned}
$$

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

$$
\begin{aligned}
C_3 &= \mathbf{E}\left[\left(F_{Y(1)}(z_1|X) - F_{Y(1)}(z_1)\right)\left(F_{Y(1)}(z_2|X) - F_{Y(1)}(z_2)\right)\right] \\
&= \mathbf{E}\left[F_{Y(1)}(z_1|X) \cdot F_{Y(1)}(z_2|X)\right] \; - \; F_{Y(1)}(z_1) \cdot F_{Y(1)}(z_2)\,.
\end{aligned}
$$

Adding up the results gives us (6.9). □

We have gathered all the results to prove Theorem 6.3.

Proof. *(Theorem 6.3)* We connect the statement of the theorem to the error decomposition by Lemma 6.7. By Lemma 6.8, Lemma 6.9, Lemma 6.12 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 6.13. □

6.1 Application to Plug In Estimators

A plethora of applications of the delta method to estimates of the distribution function are to be found in [vdV00] and [vdVW13]. This includes Quantile estimation [vdV00, §21] [vdVW13, §3.9.21/24], survival analysis via Nelson-Aalen and Kaplan-Meier estimator [vdVW13, §3.9.19/31], Wilcoxon Test [vdVW13, §3.9.4.1], and much more. Maybe Boostrapping from the weighted distribution is also sensible .

7 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.

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

$$\operatorname{ri}(\Omega) := \{x \in \Omega \mid \exists\, \varepsilon > 0 : (x + \varepsilon B_{\mathbb{R}^n}) \cap \operatorname{aff}(\Omega) \subset \Omega\}\,. \tag{7.1}$$

Theorem 7.1. (Convex separation in finite dimension) *Let* C_1 *and* C_2 *be two non-empty convex sets in* $\mathbb{R}^n$ *. Then* C_1 *and* C_2 *can be properly separated if and only if* $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$.

Proof. [Roc70, Theorem 11.3] □

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

Proposition 7.1. *Let* C *be a non-empty convex set in* $\mathbb{R}^n$*. The following holds:*

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

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

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

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

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

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

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

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

$$z_i + t(z_i - x_i) \in C_i \qquad \text{for all } i \in \{1, 2\}\,. \tag{7.2}$$

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

$$z_i + t_i(z_i - x_i) \in C_i \qquad \text{for all } i \in \{1, 2\}. \tag{7.3}$$

If $t_1 = t_2$ we recover (7.2) from (7.3). By (iii) it holds $(z_1, z_2) \in \operatorname{ri}(C_1 \times C_2)$. If $t_1 < t_2$ we define $\theta := \frac{t_1}{t_2} \in (0, 1)$. Consider (7.3) with $i = 2$, together with $z_2 \in C_2$ and the convexity of C_2. It follows

$$z_2 + t_1(z_2 - x_2) \;=\; \theta \cdot (z_2 + t_2(z_2 - x_2)) \;+\; (1 - \theta) \cdot z_2 \in C_2. \tag{7.4}$$

Now we consider (7.4) and (7.3) with $i = 1$. This gives (7.2) with $t = t_1$. As before, it follows $(z_1, z_2) \in \operatorname{ri}(C_1 \times C_2)$. If $t_1 > t_2$ similar arguments lead to the same result. We have proven $\operatorname{ri}(C_1 \times C_2) \supseteq \operatorname{ri} C_1 \times \operatorname{ri} C_2$ and equality. □

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

$$f(\theta x + (1 - \theta)y) \;\leq\; \theta f(x) + (1 - \theta) f(y) \qquad \text{for all } x, y \in \mathbb{R}^n \text{ and all } \theta \in [0, 1]. \tag{7.5}$$

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

$$f\left(\sum_{i=1}^m \theta_i x_i\right) \;\leq\; \sum_{i=1}^m \theta_i f(x_i) \qquad \text{for all } x_1, \ldots, x_m \in \mathbb{R}^n. \tag{7.6}$$

We call a function **strictly convex** if the inequality in (7.5) is strict.

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

$$\operatorname{dom} f \;:=\; \{x \in \mathbb{R}^n : f(x) < \infty\}. \tag{7.7}$$

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

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

$$\langle x^*, x - \overline{x} \rangle \leq f(x) - f(\overline{x}). \tag{7.8}$$

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

$$\langle \nabla f(\overline{x}), x - \overline{x} \rangle \leq f(x) - f(\overline{x}). \tag{7.9}$$

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

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

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

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

can by properly separated.

Proof. We fix $x^* \in \operatorname{dom} \sigma_{C_1\cap C_2}$ and write $\alpha := \sigma_{C_1\cap C_2}(x^*)$. In order to apply convex separation in finite dimension (Theorem 7.1) 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 7.1 (v) and the convexity of C_2 we get the convexity of $L^{-1}[\alpha, \infty)$ and with it that of $\Theta_2(x^*)$.

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

We start by calculating the relative interiors. It holds

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

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

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

Then it follows

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

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

Applying Theorem 7.1 finishes the proof. □

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

$$(\sigma_{C_1 \cap C_2})(x^*) = (\sigma_{C_1} \square\, \sigma_{C_2})(x^*) \qquad \text{for all } x^* \in \mathbb{R}^n. \tag{7.10}$$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = \sigma_{C_1}(x_1^*) + \sigma_{C_2}(x_2^*). \tag{7.11}$$

Proof. Using Lemma 7.1 the rest of the proof is as that of *[MMN22, Theorem 4.23(b)]*. □

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

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

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} \langle x^*, x\rangle - f(x). \tag{7.12}$$

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 7.2. *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***

$$(f+g)^*(x^*) = (f^* \square g^*)(x^*) \tag{7.13}$$

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

$$(f+g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{7.14}$$

Proof. [MMN22, Theorem 4.27(c)] □

Theorem 7.3. *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***

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{7.15}$$

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

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

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

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

$$p_i : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto x_i. \tag{7.16}$$

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

$$p_i^* : \mathbb{R} \to \mathbb{R}^n, \quad x \mapsto e_i \cdot x\,. \tag{7.17}$$

For the inverse image of the adjoint of p_i it holds

$$(p_i^*)^{-1}\{(x_1^*, \ldots, x_n^*)\} \;=\; \begin{cases} \{x_i^*\}\,, & \text{if } x_j^* = 0 \text{ for all } j \neq i\,, \\ \emptyset & \text{else.} \end{cases} \tag{7.18}$$

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

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

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

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

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

$$\left(S_{f,n} \circ p_{\{1,\ldots,n\}}\right)^*(x_1^*, \ldots, x_{n+1}^*) = \begin{cases} S_{f,n}^*(x_1^*, \ldots, x_n^*) & \text{if } x_{n+1}^* = 0\,, \\ \infty & \text{else,} \end{cases} \tag{7.19}$$

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

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

and

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

By the conjugate sum rule it follows

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

♢

7.2 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.

Assumption 4. *The objective function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is strictly convex and its convex conjugate f^* is continuously differentiable.*

Theorem 7.4. *Consider the optimization problem*

$$\begin{aligned}&\underset{w\in\mathbb{R}^n}{\text{minimize}} && f(w) && (7.20)\\ &\text{subject to} && \mathbf{U}w\ \geq\ d\,.\\ &&& \mathbf{A}w\ =\ a\,,\end{aligned}$$

and its dual problem

$$\begin{array}{ll} \underset{\lambda_d \in \mathbb{R}^r, \lambda_a \in \mathbb{R}^s}{\text{maximize}} & \langle \lambda_d, d \rangle \; + \; \langle \lambda_a, a \rangle \; - \; f^* \Big(\mathbf{U}^\top \lambda_d + \mathbf{A}^\top \lambda_a \Big) \\ \text{subject to} & \lambda_d \; \geq \; 0 \, . \end{array} \tag{7.21}$$

Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ *be an optimal solution to* (7.21)*. If the objective function* f *of* (7.20) *is strictly convex and its convex conjugate* f^* *is continuously differentiable, then the unique optimal solution to* (7.20) *is given by*

$$w^\dagger = \nabla f^* \Big(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \Big) \, . \tag{7.22}$$

Plan of Proof

We show that $w^\dagger$ and $(\lambda_d^\dagger, \lambda_a^\dagger)$ meet the Karush-Kuhn-Tucker conditions for 7.20, that is, **complementary slackness**

$$\langle \lambda_d^\dagger \, , d - \mathbf{U} w^\dagger \rangle \; = \; 0 \, , \tag{7.23}$$

primal and **dual feasibility**

$$\mathbf{U} w^\dagger \; \geq \; d \, , \tag{7.24}$$

$$\mathbf{A} w^\dagger \; = \; a \, ,$$

$$\lambda_d^\dagger \; \geq \; 0 \, , \tag{7.25}$$

and **stationarity**

$$0_n \; \in \; [\partial f(w^\dagger) \; + \; \partial \left(w \mapsto d - \mathbf{U} w \right) (w^\dagger) \cdot \lambda_d^\dagger \; + \; \partial \left(w \mapsto a - \mathbf{A} w \right) (w^\dagger) \cdot \lambda_a^\dagger \,] \, . \tag{7.26}$$

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 7.2. [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 (7.21).

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

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

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 (7.23) we will show for all coordinates

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

It is well know that a concave functions g satisfies

$$g(x) - g(y) \ge \nabla g(x)^\top (x-y) \qquad \text{for all } x, y\,. \tag{7.27}$$

But G is concave by the convexity of f^*.

First, we show

$$\nabla G_i(\lambda_d^\dagger) \;\le\; 0 \qquad \text{for all } i \in \{1,\dots,s\}\;. \tag{7.28}$$

Assume towards a contradiction that $\nabla G_i(\lambda_d^\dagger) > 0$ for some $i \in \{1,\dots,s\}$. By the continuity of ∇G there exists $\varepsilon > 0$ such that $\nabla G_i(\lambda_d^\dagger + e_i\cdot\varepsilon) > 0$. It follows from (7.27)

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

which contradicts the optimality of $\lambda_d^\dagger$ for (7.21). It follows (7.28).

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

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

which contradicts the optimality of $\lambda_d^\dagger$. It follows (7.23), 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 (7.28), $w^\dagger$ satisfies the inequality constraints in (7.20). To prove this for the equality constraints, we view G from a different angel. Let for fixed $\lambda_d^\dagger$

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

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 7.2 and the differentiability of g it holds

$$a \in \partial g(\lambda_a^\dagger) = \left\{\mathbf{A} \cdot \nabla f^*\left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger\right)\right\} = \left\{\mathbf{A} w^\dagger\right\}. \tag{7.29}$$

Thus $a = \mathbf{A}w^\dagger$. But then $w^\dagger$ satisfies also the equality constraints. We proved (7.24).

Stationarity

First we show

$$\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger \in \partial f(w^\dagger). \tag{7.30}$$

By Proposition 7.2 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 (7.30). Next we show

$$-\mathbf{U}^\top \in \partial\left(w \mapsto d - \mathbf{U}w\right)(w^\dagger) \qquad \text{and} \qquad -\mathbf{A}^\top \in \partial\left(w \mapsto d - \mathbf{A}w\right)(w^\dagger). \tag{7.31}$$

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 (7.30) and (7.31) we conclude

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

We have proved (7.26), that is, stationarity.

Dual Feasibility and Conclusion

Dual feasibility (7.25) follows immediately from the optimality of $\lambda_d^\dagger$ for (7.21). Thus, $(\lambda_d^\dagger, \lambda_a^\dagger)$ and $w^\dagger$ satisfy the Karush-Kuhn-Tucker conditions for (7.20). Applying [Roc70, Theorem 28.3] finishes the proof. □

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

References

[AB07] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, May 2007.

[CYZ16] Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally Efficient Non-Parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):673–700, June 2016.

[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.

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

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

[vdVW13] Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, March 2013.

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

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

[Zub15] José R. Zubizarreta. Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association*, 110(511):910–922, July 2015.

Index