Todo list

Solving missing survival times with entropy balancing weights

Universität Stuttgart

Ioan Scheffel

February 11, 2023

Contents

1 Balancing Weights

1.1 Introduction

Let X be a d-dimensional random variable from which we sample $N \in \mathbb{N}$ independent and identically distributed copies $X_1, \ldots, X_N$. Furthermore, let $T \in \{0, 1\}$ denote the indicator of treatment and let $T_1, \ldots, T_N$ be i.i.d. copies. Let $(Y(0), Y(1)) \in [-M, M]^2$ be a real-valued bounded random variables with $M \in \mathbb{R}$. We call X the covariate vector, T the indicator of treatment and $(Y(0), Y(1))$ the marginal potential outcomes under treatment. Before we begin the analysis we order the i.i.d. copies, such that the first n units received treatment and the remaining $N - n$ units are in the control group.

We want know more about the (potential) outcome under treatment in the population. But we observe the outcome Y depending on T, that is

$$Y_i \sim Y|T_i\,. \tag{1.1}$$

In the presence of confounders we have

$$Y|T = 1 \nsim Y(1)\,. \tag{1.2}$$

We want to create weights $w_1, \ldots, w_n$ for the treated, such that for large N we get

$$w_i \cdot Y_i \sim Y(1)\,. \tag{1.3}$$

We extend ideas from entropy balancing weights [WZ19].

To be precise in our statements, we introduce some notation.

Basis Functions

We adopt ideas from [GKKW02]. Another angle would be sieve estimates [New97] where the number of basis functions can grow slower than N. We need a different notion of Consistency as is given in [GKKW02, Definitien 1.1].

$$\mathbf{E}\left[\int_{\mathcal{X}} \left|\sum_{k=1}^{N} B_k(x) \cdot m(X_k) - m(x)\right|^2 \mathbf{P}_X(dx)\right] \to 0 \qquad \text{as } n \to \infty\,. \tag{1.4}$$

Nevertheless we leverage (weak) universal consistency of regression basis as in [GKKW02, Definitien 1.3].

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

$$\mathbf{P}\left[\left|\sum_{k=1}^{N} B_k(X) \cdot m(X_k) - m(X)\right| \geq \varepsilon\right] \to 0 \qquad \text{as } n \to \infty\,. \tag{1.5}$$

Proof. By Markov's inequality it holds

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

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

Next we partitioning and kernel estimates form [GKKW02, §4,§5].

Partitioning Estimates

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

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

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

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

Under mild conditions, the basis functions are universally consistent.

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

$$\max_{j\colon A_{N,j} \cap S \neq \emptyset} \operatorname{diam} A_{N,j} \ \to\ 0 \qquad \text{for } N \to \infty \tag{1.6}$$

and

$$\frac{\#\{j\colon A_{N,j}\cap S\neq\emptyset\}}{N}\;\to\;0\qquad \textit{for } N\to\infty \tag{1.7}$$

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

Proof. [GKKW02, Theorem 4.2.] □

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

$$\mathbf{P}\left[\left|\sum_{k=1}^{N}B_k(X)\cdot m(X_k)-m(X)\right|\geq\varepsilon\right]\to 0\qquad \textit{as } n\to\infty\,. \tag{1.8}$$

Kernel Estimates

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

$$B_k(x):=\frac{K\left(\frac{x-X_k}{h_n}\right)}{\sum_{i=1}^{N}K\left(\frac{x-X_i}{h_n}\right)}\,. \tag{1.9}$$

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

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

$$\mathbf{1}_{\{x\in S_{0,R}\}}\geq K(x)\geq b\cdot\mathbf{1}_{\{x\in S_{0,r}\}} \tag{1.10}$$

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

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

$$\mathbf{P}\left[\left|\sum_{k=1}^{N}B_k(X)\cdot m(X_k)-m(X)\right|\geq\varepsilon\right]\to 0\qquad \textit{as } n\to\infty\,. \tag{1.11}$$

Objective Function

We will consider the sum. We need f to be strictly convex and its convex conjugate to be 2nd order smooth and strictly non-decreasing.

Negative Entropy

We define the negative entropy to be

$$f\colon [0,\infty) \to \mathbb{R}, \quad w \mapsto \begin{cases} 0 & \text{if } w = 0, \\ w \log w & \text{else.} \end{cases} \tag{1.12}$$

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

$$(f')^{-1} = \lambda \mapsto e^{\lambda - 1} \tag{1.13}$$

Thus

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

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

Sample Variance

We define the sample variance to be

$$f\colon \mathbb{R} \to \mathbb{R}, \quad w \mapsto (w - 1/n)^2 \tag{1.14}$$

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

$$(f')^{-1} = \lambda \mapsto \frac{\lambda}{2} + \frac{1}{n} \tag{1.15}$$

Thus

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

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

Convex Optimization

We now consider the convex optimization problem for the weights. We assume the objective function f to be strictly convex, such that its convex conjugate is continuously differentiable. Furthermore we assume the bound on the box-constraints $\delta > 0$ to converge to 0 in probability.

Problem 1.1.

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

To obtain consistency for the weights we will analyse the dual.

1.2 Dual Formulation

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

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

where

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

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

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

Plan of proof

We bring Problem 1.1 in the form of ts. Then we apply the results of the ts chapter. We wait to conclude on the weights. Then we eliminate the non-negativity constraints on the dual variables leveraging convexity and optimality.

Proof. **Form**

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

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

the constraints vectors

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

the matrix of the basis functions of the treated

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

and the constraint matrices

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

By Example 2.1 the convex conjugate of the objective function of Problem 1.1 is

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

Before we apply Theorem 2.4 we eliminate the non-negativity constraints. To this end, we consider the objective function G of the dual problem and update it until we reach its final form. We write

$$\lambda_d =: \begin{bmatrix} \rho \\ \lambda^+ \\ \lambda^- \end{bmatrix} \tag{1.16}$$

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

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

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

Non-negativity constraints

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

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

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

$$\tilde{\lambda} \;:=\; \Big[\, \lambda_1^+, \dots,\; \lambda_i^+ - (\lambda_i^+ \wedge \lambda_i^-)\,,\; \dots, \lambda_N^+,\; \lambda_1^-, \dots, \lambda_i^- - (\lambda_i^+ \wedge \lambda_i^-)\,,\; \dots, \lambda_N^-, \lambda_0 \Big]^\top . \tag{1.17}$$

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

$$G(\tilde{\lambda}) - G(\lambda) \;=\; 2N \cdot \delta_i \cdot (\lambda_i^+ \wedge \lambda_i^-) \;>\; 0\,, \tag{1.18}$$

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

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

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

We apply Theorem 2.4 to finish the proof. $\square$

1.3 Consistency

Consistency of the Dual

Let λ^* denote the vector with coordinates

$$\lambda_i^* := f'(1/\pi_i) - \lambda_0^\dagger \tag{1.19}$$

Theorem 1.4. *For all $\varepsilon > 0$ it holds*

$$\mathbf{P}\left[\left\|\lambda^\dagger - \lambda^*\right\| \geq \varepsilon\right] \;\to\; 0 \qquad \textit{for } N \to \infty\,. \tag{1.20}$$

The following Lemma allows us to leverage the convexity of the objective function of the dual to get

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

Lemma 1.1. *Let $m \in \mathbb{N}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ be convex. Then for all $y \in \mathbb{R}^m$ and $\varepsilon > 0$*

$$\inf_{\|\Delta\|=\varepsilon} g(y + \Delta) - g(y) \geq 0 \tag{1.21}$$

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

Proof. Since $y + \varepsilon B$ is convex, it contains a local minimum of g. Suppose towards a contradiction that $y^* \in y + \varepsilon B$ is a local minimum, but not a global one, and (1.21) is true. Then it holds

$$g(x) < g(y^*) \quad \text{for some } x \in \mathbb{R}^m \setminus (y + \varepsilon B)\,. \tag{1.22}$$

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

$$\theta x + (1-\theta)y^* = y + \Delta_x\,. \tag{1.23}$$

It follows

$$\begin{aligned} g(y^*) \leq g(y) \leq g(y + \Delta_x) &= g(\theta x + (1-\theta)y^*) \\ &\leq \theta g(x) + (1-\theta) g(y^*) < g(y^*), \end{aligned} \tag{1.24}$$

which is a contradiction. The first inequality is due to y^* being a local minimum of g in $y + \varepsilon B$, the second inequality is due to (1.21) being true, the equality is due to (1.23), the third inequality is due to the convexity of g and the strict inequality is due to (1.22). Thus every local minimum of g in $y + \varepsilon B$ is also a global minimum. □

Proof. *(Theorem 1.4)* We separate the differentiable part in G to get

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

Analysis of S

By the triangle inequality we get

$$S \leq \frac{1}{N}\sum_{i=1}^{N}|1 - T_i/\pi_i| + \max_{i=1,\ldots,n}\left|(f')^{-1}\left(\langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger\right) \ - \ 1/\pi_i\right| \tag{1.25}$$

$$=: S_1 + M \tag{1.26}$$

Analysis of S_1

Since X_i and T_i are i.i.d. and

$$\mathbf{E}[T_i/\pi_i] = \mathbf{E}[\mathbf{E}[T_i|X_i]/\pi_i] = 1 \tag{1.27}$$

it holds by the weak law of large numbers

$$\mathbf{P}[S_1 \leq \tilde{\varepsilon}/(4\varepsilon)] \to 1 \quad \text{for } n \to \infty \tag{1.28}$$

Analysis of M

Since

$$\langle B(X_i), 1_{\mathrm{N}} \rangle = 1 \tag{1.29}$$

and

$$\lambda_i^* = f'(1/\pi_i) - \lambda_0^\dagger \tag{1.30}$$

it holds

$$\langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger = \sum_{k=1}^{N} f'(1/\pi_i) \cdot B_k(X_i) \tag{1.31}$$

By the universal consistency of B this converges to $f'(1/\pi_i)$ in probability. By the continuity of $(f')^{-1}$ it follows

$$\mathbf{P}[M \leq \tilde{\varepsilon}/(4\varepsilon)] \to 1 \quad \text{for } n \to \infty \tag{1.32}$$

Conclusion

It follows

$$\mathbf{P}[S \leq \tilde{\varepsilon}/(2\varepsilon)] \to 1 \quad \text{for } n \to \infty \tag{1.33}$$

We get for all $\tilde{\varepsilon} > 0$

$$G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq -\tilde{\varepsilon} \tag{1.34}$$

with probability going to 1 for $n \to \infty$. Thus it also holds

$$\mathbf{P}\left[G(\lambda^* + \Delta, \lambda_0^\dagger + \Delta_0) - G(\lambda^*, \lambda_0^\dagger) \geq 0\right] \to 0 \qquad \text{for } n \to \infty\,. \tag{1.35}$$

Applying Lemma 1.1 finishes the proof.

□

Consistency of the Primal

Theorem 1.5. *For all $i \in \{1, \dots, n\}$ and all $\varepsilon > 0$ it holds*

$$\mathbf{P}\left[|1/\pi_i - w_i| \geq \varepsilon\right] \;\to\; 0 \qquad \textit{for } N \to \infty\,. \tag{1.36}$$

Proof. It holds

$$\begin{aligned} |1/\pi_i - w_i| \;&=\; \left|1/\pi_i - (f')^{-1}\left(\langle B(X_i), \lambda^\dagger \rangle + \lambda_0^\dagger\right)\right| \\ &\leq\; \left|1/\pi_i - (f')^{-1}\left(\langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger\right)\right| \\ &\quad+\; \left|(f')^{-1}\left(\langle B(X_i), \lambda^\dagger \rangle + \lambda_0^\dagger\right) - (f')^{-1}\left(\langle B(X_i), \lambda^* \rangle + \lambda_0^\dagger\right)\right| \end{aligned}$$

By the universal consistency of B and the continuity of $(f')^{-1}$ the first term converges to 0 in probability. By the continuity of $(f')^{-1}$, the uniform boundedness of $\|B\|$ and the dual consistency the second term converges to 0 in probability.

□

Consistency of the Weighted Mean

Theorem 1.6. *For all* $\varepsilon > 0$ *it holds*

$$\mathbf{P}\left[\left|\frac{1}{N}\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \geq \varepsilon\right] \;\to\; 0 \qquad \text{for } N \to \infty\,, \tag{1.37}$$

that is, the weighted mean is a consistent estimator.

Proof.

$$\begin{aligned}
\left|\frac{1}{N}\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \;\leq\; & \left|\frac{1}{N}\left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i)\right)^{\top} \mathbf{Y}(1)\right| \\
& + \left|\frac{1}{N}\sum_{i=1}^{n} w_i \left(\mathbf{E}[Y(1)|X_i] - \langle B(X_i)), \mathbf{Y}(1)\rangle\right)\right| \\
& + \left|\frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{E}[Y(1)|X_i] - \langle B(X_i)), \mathbf{Y}(1)\rangle\right)\right| \\
& + \left|\frac{1}{N}\sum_{i=1}^{n} (w_i - 1/\pi_i)\left(Y_i - \mathbf{E}[Y(1)|X_i]\right)\right| \\
& + \left|\frac{1}{N}\sum_{i=1}^{N} T_i/\pi_i \left(Y_i - \mathbf{E}[Y(1)|X_i]\right) + \left(\mathbf{E}[Y(1)|X_i] - \mathbf{E}[Y(1)]\right)\right| \\
=:\; & R_1 + R_2 + R_3 + R_4 + R_5
\end{aligned}$$

Analysis of R_1

By the Cauchy-Schwarz inequality it holds

$$R_1 \;=\; \left|\frac{1}{N}\left(\sum_{i=1}^{n} w_i B(X_i) - \sum_{i=1}^{N} B(X_i)\right)^{\top} \mathbf{Y}(1)\right| \leq \|\delta\|\,\|\mathbf{Y}(1)\| \tag{1.38}$$

Analysis of R_2 and R_3

By the universal consistency of B and $\frac{1}{N}\sum_{i=1}^{n} w_i = 1$ the terms R_2 and R_3 converge to 0 in probability.

Analysis of R_4

By the boundedness of Y and the primal consistency R_4 converges to 0 in probability.

Analysis of R_5

Since the expectation is 0, the term R_5 converges to 0 in probability by the weak law of large numbers.

$\square$

2 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.

2.1 A Convex Analysis Primer

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

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

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

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

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

$$\mathrm{ri}(\Omega) := \{x \in \Omega \colon \text{ there exists } \varepsilon > 0 \text{ such that } (x + \varepsilon B) \cap \mathrm{aff}(\Omega) \subset \Omega\}\,. \tag{2.1}$$

Next we collect some useful properties of relative interiors.

Proposition 2.1. *Let C be a non-empty convex set in $\mathbb{R}^n$. The following holds:*

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

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

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

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

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

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

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

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

$$z_i + t(z_i - x_i) \in C_i \qquad \text{for all } i \in \{1, 2\}\,. \tag{2.2}$$

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

$$z_i + t_i(z_i - x_i) \in C_i \qquad \text{for all } i \in \{1, 2\}\,. \tag{2.3}$$

If $t_1 = t_2$ we recover (2.2) from (2.3). By (iii) it holds $(z_1, z_2) \in \mathrm{ri}(C_1 \times C_2)$. If $t_1 < t_2$ we define $\theta := \frac{t_1}{t_2} \in (0, 1)$. Consider (2.3) with $i = 2$, together with $z_2 \in C_2$ and the convexity of C_2. It follows

$$z_2 + t_1(z_2 - x_2) = \theta \cdot (z_2 + t_2(z_2 - x_2)) + (1 - \theta) \cdot z_2 \in C_2\,. \tag{2.4}$$

Now we consider (2.4) and (2.3) with $i = 1$. This gives (2.2) with $t = t_1$. As before, it follows $(z_1, z_2) \in \mathrm{ri}(C_1 \times C_2)$. If $t_1 > t_2$ similar arguments lead to the same result. We have proven $\mathrm{ri}(C_1 \times C_2) \supseteq \mathrm{ri}\, C_1 \times \mathrm{ri}\, C_2$ and equality. $\square$

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

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

Theorem 2.1. (Convex separation in finite dimension) *Let C_1 and C_2 be two non-empty convex sets in $\mathbb{R}^n$. Then C_1 and C_2 can be properly separated if and only if* $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$.

Proof. [Roc70, Theorem 11.3] □

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

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

Definition 2.1. *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 2.1. *Let C_1 and C_2 be two non-empty convex sets in $\mathbb{R}^n$. For any $x^* \in \operatorname{dom} \sigma_{C_1 \cap C_2}$ the sets*

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

can by properly separated.

Proof. We fix $x^* \in \operatorname{dom} \sigma_{C_1 \cap C_2}$ and write $\alpha := \sigma_{C_1 \cap C_2}(x^*)$. In order to apply convex separation in finite dimension (Theorem 2.1) 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 2.1 (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 2.1 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*

$$(\sigma_{C_1 \cap C_2})(x^*) = (\sigma_{C_1} \Box \sigma_{C_2})(x^*) \qquad \text{for all } x^* \in \mathbb{R}^n. \tag{2.5}$$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = \sigma_{C_1}(x_1^*) + \sigma_{C_2}(x_2^*). \tag{2.6}$$

Proof. Using Lemma 2.1 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.

2.2 Conjugate Calculus

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

Definition 2.2. (Convex conjugate) *Given a function* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *, the* ***convex conjugate*** $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ *of* f *is defined as*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \tag{2.7}$$

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

Proposition 2.2. *Let* $f : \mathbb{R}^n \to (-\infty, \infty]$ *be a proper function. Then its convex conjugate* $f^* : \mathbb{R}^n \to (-\infty, \infty]$ *is convex.*

Proof. [MMN22, Proposition 4.2] □

Theorem 2.2. *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***

$$(f+g)^*(x^*) = (f^* \Box g^*)(x^*) \tag{2.8}$$

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

$$(f+g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{2.9}$$

Proof. [MMN22, Theorem 4.27(c)] □

Theorem 2.3. *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***

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{2.10}$$

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

Proof. [MMN22, Theorem 4.28(c)] □

Example 2.1. 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

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

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

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

$$p_i : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto x_i\,. \tag{2.11}$$

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

$$p_i^* : \mathbb{R} \to \mathbb{R}^n, \quad x \mapsto e_i \cdot x\,. \tag{2.12}$$

For the inverse image of the adjoint of p_i it holds

$$(p_i^*)^{-1}\{(x^*_1, \dots, x^*_n)\} \;=\; \begin{cases} \{x^*_i\}\,, & \text{if } x^*_j = 0 \text{ for all } j \neq i\,, \\ \emptyset & \text{else.} \end{cases} \tag{2.13}$$

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

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

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

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

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

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

$$\left(S_{f,n} \circ p_{\{1,\dots,n\}}\right)^* (x_1^*, \dots, x_{n+1}^*) = \begin{cases} S_{f,n}^*(x_1^*, \dots, x_n^*) & \text{if } x_{n+1}^* = 0\,, \\ \infty & \text{else,} \end{cases} \tag{2.14}$$

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

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

and

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

By the conjugate sum rule it follows

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

$\diamondsuit$

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

2.3 Duality of Optimal Solutions

We consider a general convex optimization problem with matrix equality and inequality constraints. For this problem there exists a related problem, which we call its dual. With ideas from [TB91] we establish a functional relationship between the optimal solution of the original problem and optimal solutions of the dual. The main assumption is that in the original problem we have a strictly convex objective function with continuously differentiable convex conjugate(cf. Definition 2.2).

Theorem 2.4. *Consider the optimization problem*

$$\begin{aligned} \underset{w \in \mathbb{R}^n}{\text{minimize}} \quad & f(w) \\ \text{subject to} \quad & \mathbf{U}w \;\geq\; d\,. \\ & \mathbf{A}w \;=\; a\,, \end{aligned} \tag{2.15}$$

and its dual problem

$$\begin{aligned} \underset{\lambda_d \in \mathbb{R}^r, \lambda_a \in \mathbb{R}^s}{\text{maximize}} \quad & \langle \lambda_d, d \rangle \;+\; \langle \lambda_a, a \rangle \;-\; f^*\Big(\mathbf{U}^\top \lambda_d + \mathbf{A}^\top \lambda_a\Big) \\ \text{subject to} \quad & \lambda_d \;\geq\; 0\,. \end{aligned} \tag{2.16}$$

Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ *be an optimal solution to* (2.16)*. If the objective function* f *of* (2.15) *is strictly convex and its convex conjugate* f^* *is continuously differentiable, then the unique optimal solution to* (2.15) *is given by*

$$w^\dagger = \nabla f^*\Big(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger\Big)\;. \tag{2.17}$$

Plan of Proof

We show that $w^\dagger$ and $(\lambda_d^\dagger, \lambda_a^\dagger)$ meet the Karush-Kuhn-Tucker conditions for 2.15, that is, **complementary slackness**

$$\langle \lambda_d^\dagger, d - \mathbf{U}w^\dagger \rangle \;=\; 0\,, \tag{2.18}$$

primal and **dual feasibility**

$$\begin{aligned} \mathbf{U}w^\dagger \;&\geq\; d\,, \qquad (2.19) \\ \mathbf{A}w^\dagger \;&=\; a\,, \\ \lambda_d^\dagger \;&\geq\; 0\,, \qquad (2.20) \end{aligned}$$

and **stationarity**

$$0_n \ \in \ [\partial f(w^\dagger) \ + \ \partial\,(w \mapsto d - \mathbf{U}w)\,(w^\dagger) \cdot \lambda_d^\dagger \ + \ \partial\,(w \mapsto a - \mathbf{A}w)\,(w^\dagger) \cdot \lambda_a^\dagger\,]\,. \qquad (2.21)$$

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

Proposition 2.3. [Roc70, Theorem 23.5(a)-(b)]. *For any proper convex function g and any vector w, it holds $t \in \partial f(w)$ if and only if $x \mapsto \langle x, t\rangle - f(x)$ achieves its supremum at w.*

Proof. Let $(\lambda_d^\dagger, \lambda_a^\dagger)$ be an optimal solution to (2.16).

Complementary Slackness

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

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

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

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

Let $\lambda_{d,i}^\dagger$ be the i-th coordinate of $\lambda_d^\dagger$ and $\nabla G_i(\lambda_d^\dagger)$ be the i-th coordinate of $\nabla G(\lambda_d^\dagger)$. To establish (2.18) 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

$$g(x) - g(y) \ge \nabla g(x)^\top (x - y) \qquad \text{for all } x, y\,. \qquad (2.22)$$

But G is concave by the convexity of f^* (cf. Proposition 2.2).

First, we show

$$\nabla G_i(\lambda_d^\dagger) \ \le \ 0 \qquad \text{for all } i \in \{1, \ldots, s\}\,. \qquad (2.23)$$

Assume towards a contradiction that $\nabla G_i(\lambda_d^\dagger) > 0$ for some $i \in \{1, \ldots, s\}$. By the continuity of ∇G there exists $\varepsilon > 0$ such that $\nabla G_i(\lambda_d^\dagger + e_i \cdot \varepsilon) > 0$. It follows from (2.22)

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

which contradicts the optimality of $\lambda_d^\dagger$ for (2.16). It follows (2.23).

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

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

which contradicts the optimality of $\lambda_d^\dagger$. It follows (2.18), that is, we proved complementary slackness.

Primal Feasibility

Since f^* is continuously differentiable it holds

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

Thus, by (2.23), $w^\dagger$ satisfies the inequality constraints in (2.15). To prove this for the equality constraints, we view G from a different angel. Let for fixed $\lambda_d^\dagger$

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

The function g inherits convexity and differentiability from f^*. From the optimality of $\lambda_a^\dagger$ we know that G takes its maximum there. But then by Proposition 2.3 and the differentiability of g it holds

$$a \;\in\; \partial g(\lambda_a^\dagger) \;=\; \left\{\mathbf{A} \cdot \nabla f^* \left(\mathbf{U}^\top \lambda_d^\dagger + \mathbf{A}^\top \lambda_a^\dagger\right)\right\} \;=\; \left\{\mathbf{A} w^\dagger\right\}\,. \tag{2.24}$$

Thus $a = \mathbf{A} w^\dagger$. But then $w^\dagger$ satisfies also the equality constraints. We proved (2.19).

Stationarity

First we show

$$\mathbf{U}^\top \lambda_d^\dagger \;+\; \mathbf{A}^\top \lambda_a^\dagger \;\in\; \partial f(w^\dagger)\,. \tag{2.25}$$

By Proposition 2.3 it suffices to show that

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

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

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

It follows (2.25). Next we show

$$-\mathbf{U}^\top \in \ \partial\,(w \mapsto d - \mathbf{U}w)\,(w^\dagger) \qquad \text{and} \qquad -\mathbf{A}^\top \in \ \partial\,(w \mapsto d - \mathbf{A}w)\,(w^\dagger)\,. \tag{2.26}$$

To this end, note that

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

Thus $-\mathbf{U}^\top \in \partial\,(w \mapsto d - \mathbf{U}w)\,(w^\dagger)$. In the same way it follows $-\mathbf{A}^\top \in \partial\,(w \mapsto d - \mathbf{A}w)\,(w^\dagger)$. From (2.25) and (2.26) we conclude

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

We have proved (2.21), that is, stationarity.

Dual Feasibility and Conclusion

Dual feasibility (2.20) follows immediately from the optimality of $\lambda_d^\dagger$ for (2.16). Thus, $(\lambda_d^\dagger, \lambda_a^\dagger)$ and $w^\dagger$ satisfy the Karush-Kuhn-Tucker conditions for (2.15). Applying [Roc70, Theorem 28.3] finishes the proof. $\square$

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

3 Simple yet useful Calculations

Theorem 3.1. (Multivariate Taylor Theorem) *Let* $f \in C^2(\mathbb{R}^n, \mathbb{R})$ *. Then for all* $x, \Delta \in \mathbb{R}^n$ *there exists* $\xi \in [0,1]$ *such that it holds*

$$\begin{aligned} f(x+\Delta) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\Delta_i + \sum_{\substack{i,j=1 \\ i\neq j}} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i \partial x_j}\Delta_i\Delta_j \\ + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i^2}\Delta_i^2 \end{aligned} \tag{3.1}$$

Corollary 3.1.1. *Let* $f \in C^2(\mathbb{R})$ *. Then for all* $a, x, \Delta \in \mathbb{R}^n$ *there exist* $\xi \in [0,1]$ *such that it holds*

$$f(a^T(x+\Delta)) - f(a^T x) = f^{'}(a^T x)\, \Delta^T a + \frac{1}{2} f^{''}(a^T(x+\xi\Delta))\, \Delta^T A\ \Delta, \tag{3.2}$$

where $A := aa^T \in \mathbb{R}^{n\times n}$ *.*

Proof. By the chain rule we have for all $a, x, \Delta \in \mathbb{R}^n$ and $\xi \in [0,1]$

$$\frac{\partial^2 f(a^T(x+\xi\Delta))}{\partial x_i \partial x_j} = f^{''}(a^T(x+\xi\Delta))\, a_i a_j. \tag{3.3}$$

Since $A := aa^T$ is symmetric we have

$$\Delta^T A\ \Delta = 2\sum_{\substack{i,j=1 \\ i\neq j}}^{n} a_i a_j \Delta_i \Delta_j + \sum_{i=1}^{n} a_i^2 \Delta_i^2. \tag{3.4}$$

Plugging (3.3) and (3.4) into (3.1) yields (3.2). □

Proposition 3.1. *For all* $x, y \in \mathbb{R}$ *it holds*

$$|x+y| - |x| \geq -|y| \tag{3.5}$$

Proof. Checking all 6 combinations of $x+y, x, y$ being nonnegative or negative yields the result. □

Notation Index

$\#A$ cardinality of the set A

$\mathbf{E}[X|Y]$ conditional expectation of the random variable X with respect to $\sigma(Y)$

$\mathbf{E}[X]$ expectation of the random variable X

$\mathbf{Var}[X]$ variance of the random variable X

$\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ extension of the real numbers

$\xrightarrow{\mathcal{D}}$ convergence of distributions

$\mathbf{P}$ generic probability measure

$\mathbf{P}_X = \mathbf{P} \circ X^{-1}$ distribution of the random variable X

$\mathbb{R}$ set of real numbers

$x \vee y, x \wedge y, x^+, x^-$ maximum, minimum, positive part, negative part of real numbers

$X \sim \mu$ the random variable has distribution μ

Bibliography

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

[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.

[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.

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