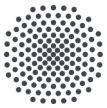
Todo list

Robust Weighting and Matching Techniques for Causal Inference in Observational Studies with Continuous Treatment

Universität Stuttgart



Universität Stuttgart

Ioan Scheffel

January 27, 2023

Contents

1 Balancing Weights

Introduction

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. Next we define m_n by

$$m_n(Y|x) \;:=\; \frac{\sum_{k=1}^n Y_k \cdot \mathbf{1}_{\{X_k \in A_n(x)\}}}{\sum_{j=1}^n \mathbf{1}_{\{X_j \in A_n(x)\}}}\,. \tag{1.1}$$

In the terminology of [GKKW02, §4] m_n is called a partitioning estimate. We want to control the sumands. To this end we define a set of basis functions by

$$B_k(x) \;:=\; \frac{\mathbf{1}_{\{X_k \in A_n(x)\}}}{\sum_{j=1}^n \mathbf{1}_{\{X_j \in A_n(x)\}}} \qquad \text{for } k \in \{1, \ldots, n\}\,. \tag{1.2}$$

This yields

$$m_n(Y|x) = \sum_{k=1}^n Y_k \cdot B_k(x)\,. \tag{1.3}$$

We consider the objective function

$$f \;:\; [0,\infty) \;\to\; \mathbb{R}, \quad x \;\mapsto\; x \log x\,, \tag{1.4}$$

together with

Problem 1.1.

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

Dual Problem

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

$$\underset{\lambda\in\mathbb{R}^n}{\text{minimize}} \qquad \frac{1}{n}\sum_{i=1}^{n} T_i \cdot f^*(m_n(\lambda|X_i)) - m_n(\lambda|X_i) \;+\; \langle \delta, |\lambda| \rangle \,,$$

where

$$f^* : \mathbb{R} \;\to\; \mathbb{R}\,, \qquad t \;\mapsto\; t\,(f')^{-1}(t) \;-\; f\left((f')^{-1}(t)\right)$$

is the Legendre transformation of f, $B(X_i) = [B_1(X_i), \ldots, B_K(X_i)]^\top$ denotes the K basis functions of the covariates of unit $i \in \{1, \ldots, n\}$ and $|\lambda| = [|\lambda_1|, \ldots, |\lambda_K|]^\top$, where $|\cdot|$ is the absolute value of a real-valued scalar. Moreover, if $\lambda^\dagger$ is an optimal solution then

$$w_i^* \;=\; (f')^{-1}\left(m_n(\lambda^\dagger|X_i)\right) \quad \textit{for all } i \textit{ with } T_i = 1$$

are uniquely part of any optimal solution to (P) .

Proof. We prove the following Lemma at the end of the section.

Lemma 1.1. *The dual of the optimization problem is*

$$\underset{\lambda\in\mathbb{R}^{2K}}{\text{minimize}} \qquad \frac{1}{n}\sum_{i=1}^{n} T_i \cdot f^*(\langle Q_{\bullet i}, \lambda\rangle) - \langle Q_{\bullet i}, \lambda\rangle \;+\; \langle d, \lambda\rangle$$

subject to

$$\lambda_k \geq 0 \quad \textit{for all } k \in \{1, \ldots, K\}\,,$$

where

$$\mathbf{Q} := \Big[\pm\,\mathbf{B}(\mathbf{X})\Big]\,, \qquad \mathbf{B}(\mathbf{X}) := \Big[B(X_1), \ldots, B(X_n)\Big]\,, \qquad \textit{and} \qquad d := \begin{bmatrix}\delta\\ \delta\end{bmatrix}.$$

Proof. First we disentangle the box constraints. To this end, we get

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

$$\pm\sum_{i=1}^{n} w_i T_i B_k(X_i) \;\geq\; -n\cdot\delta_k \;\pm\; \sum_{i=1}^{n} B_k(X_i)\,, \qquad k = 1, \ldots, K\,.$$

The corresponding matrix notation is

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

$$\mathbf{Q}w \;\geq\; d\,,$$

where

$$\begin{aligned} \mathbf{Q} &:= \left[\pm T\mathbf{B}(\mathbf{X})\right], \\ T\mathbf{B}(\mathbf{X}) &:= \left[T_1 B(X_1), \ldots, T_n B(X_n)\right], \\ d &:= \left[-n \cdot \delta \pm \textstyle\sum_{i=1}^{n} B_k(X_i)\right]. \end{aligned}$$

The convex conjugate is

$$\sum_{T_i=1} T_i f^*(\lambda_i) + \sum_{T_i=0} \delta_{\{0\}}(\lambda_i),$$

where

$$\delta_{\{0\}}(t) = \begin{cases} 0, & \text{if } t = 0, \\ \infty, & \text{else}. \end{cases}$$

Note that

$$\sum_{T_i=0} \delta_{\{0\}}(T_i \cdot \langle B(X_i), \lambda \rangle) = 0.$$

The corresponding dual problem in [TB91] is then

$$\underset{\lambda_1,\ldots,\lambda_K \geq 0}{\text{minimize}} \quad \sum_{i=1}^{n} T_i \cdot f^*(T_i \cdot \langle B(X_i), \lambda \rangle) - \langle B(X_i), \lambda \rangle \; + \; n\langle \lambda, d \rangle .$$

If we keep only the outer T_i and divide by $1/n$ the problem remains the same. This concludes the proof. □

Continue with paraphrase of [WZ19, page 20]

Consistency of the dual variables

Consistency of the weights

Consistency of the weighted mean

□

1.1 Consistency

We settle for partitioning estimates, although other (bounded) universal consistent basis functions would work as well.

We leverage (weak) universal consistency of partitioning estimates [GKKW02]. To this end, we choose the basis functions as

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

The euclidian norm of $B(x)$ is bounded by one.

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

The oracle parameters are $\left[f^{'}\left(\frac{1}{\pi_i}\right)\right]_{i=1,\ldots,n}$ and $[Y_i(1)]_{i=1,\ldots,n}$. It even holds

$$\lambda_i^{\dagger} \to f^{'}\left(\frac{1}{\pi_i}\right)$$

in probability.

$$\begin{aligned} G(\lambda) \;&:=\; \frac{1}{n}\sum_{j=1}^n T_i \cdot f^*\left(\langle B(X_i), \lambda\rangle\right) - \langle B(X_i), \lambda\rangle + \langle |\lambda|, \delta\rangle \\ &=:\; L(\lambda) + \langle |\lambda|, \delta\rangle \end{aligned}$$

The consistency of $\lambda^{\dagger}$ relates to the difference being nonnegative. For all $\varepsilon > 0$ it holds

$$\mathbf{P}\left[\left\|\lambda^{\dagger} - f^{'}\left(1/\pi\right)\right\| \leq \varepsilon\right] = \mathbf{P}\left[\inf_{\|\Delta\|=\varepsilon} G(f^{'}\left(1/\pi\right) + \Delta) - G(f^{'}\left(1/\pi\right)) \geq 0\right]\,.$$

The term on the right-hand side.

$$\begin{aligned} & G(f^{'}\left(1/\pi\right) + \Delta) \;-\; G(f^{'}\left(1/\pi\right)) \\ &\quad\geq\; \langle \nabla L(f^{'}\left(1/\pi\right)), \Delta\rangle \;+\; \langle |f^{'}\left(1/\pi\right) + \Delta| \;-\; |f^{'}\left(1/\pi\right)|, \delta\rangle \\ &\quad\geq\; -\|\Delta\| \left(\|\delta\| \;+\; \|B(X_i)\| \cdot \frac{1}{n}\sum_{i=1}^n \left| 1 \;-\; T_i \cdot (f^{'})^{-1}\langle B(X_i), f^{'}\left(1/\pi\right)\rangle \right| \right) \\ &\quad\geq\; -\varepsilon \left(\|\delta\| \;+\; \max_{i=1,\ldots,n} \left| 1/\pi_i \;-\; (f^{'})^{-1}\langle B(X_i), f^{'}\left(1/\pi\right)\rangle \right| + \frac{1}{n}\sum_{i=1}^n |\,1 \;-\; T_i/\pi_i| \right) \\ &\quad\geq\; -\varepsilon \left(\varepsilon_{\delta} \;+\; \omega((f^{'})^{-1}, \varepsilon_m) \;+\; \varepsilon_{\mathrm{WLLN}} \right) \\ &\quad\geq\; -\varepsilon_G\,, \end{aligned}$$

with probability tending to 1. It follows for all $\varepsilon, \varepsilon_G > 0$

$$\mathbf{P}\left[\inf_{\|\Delta\|=\varepsilon} G(f^{'}\left(1/\pi\right) + \Delta) - G(f^{'}\left(1/\pi\right)) \geq -\varepsilon_G\right] \;\to\; 1$$

for $n \to \infty$. Thus, for all $\varepsilon > 0$ it holds

$$\mathbf{P}\left[\left\|\lambda^\dagger - f^{'}\left(1/\pi\right)\right\| \leq \varepsilon\right] = \mathbf{P}\left[\inf_{\|\Delta\|=\varepsilon} G(f^{'}\left(1/\pi\right)+\Delta) - G(f^{'}\left(1/\pi\right)) \geq 0\right] \;\to\; 1\,,$$

for $n \to \infty$. Furthermore,

$$\begin{aligned}
T_i \cdot |w_i - 1/\pi_i| \;&=\; T_i \cdot \left|T_i (f^{'})^{-1}\langle T_i B(X_i), \lambda^\dagger\rangle - (f^{'})^{-1}\left(f^{'}\left(1/\pi_i\right)\right)\right| \\
&\leq\; \left|(f^{'})^{-1}\langle B(X_i), \lambda^\dagger\rangle - (f^{'})^{-1}\langle B(X_i), f^{'}\left(1/\pi\right)\rangle\right| \\
&\quad +\; \left|(f^{'})^{-1}\langle B(X_i), f^{'}\left(1/\pi\right)\rangle - (f^{'})^{-1}\left(f^{'}\left(1/\pi_i\right)\right)\right| \\
&\leq\; \omega\left((f^{'})^{-1}, \left\|\lambda^\dagger - f^{'}(1/\pi)\right\|\right) \;+\; \omega\left((f^{'})^{-1}, \varepsilon_m\right) \\
&\leq\; \omega\left((f^{'})^{-1}, \varepsilon_\dagger\right) \;+\; \omega\left((f^{'})^{-1}, \varepsilon_m\right) \\
&\leq\; \varepsilon
\end{aligned}$$

with probability tending to 1.

Next we consider the estimate.

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n w_i T_i Y_i - \mathbf{E}[Y(1)]\right| \;&\leq\; \left|\frac{1}{n}\sum_{i=1}^n (w_i T_i - 1)\langle B(X_i)), \mathbf{Y}(1)\rangle\right| \\
&\quad +\; \left|\frac{1}{n}\sum_{i=1}^n (w_i T_i - 1)\left(\mathbf{E}[Y(1)|X_i] - \langle B(X_i)), \mathbf{Y}(1)\rangle\right)\right| \\
&\quad +\; \left|\frac{1}{n}\sum_{i=1}^n T_i \cdot (w_i - 1/\pi_i)\left(Y_i - \mathbf{E}[Y(1)|X_i]\right)\right| \\
&\quad +\; \left|\frac{1}{n}\sum_{i=1}^n T_i/\pi_i \left(Y_i - \mathbf{E}[Y(1)|X_i]\right) + \left(\mathbf{E}[Y(1)|X_i] - \mathbf{E}[Y(1)]\right)\right| \\
&\leq\; C_Y \varepsilon_\delta + C_w \varepsilon_m + C_{\mathbf{E}Y}\varepsilon_w + \varepsilon_{\mathrm{WLLN}} \\
&\leq\; \varepsilon
\end{aligned}$$

with probability tending to 1.

Theorem 1.2. *Let Y be bounded and $\mathbf{E}\left[f^{'}\left(1/\pi(X)\right)^2\right] < \infty$. Then*

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

is a consistent estimator for $\mathbf{E}[Y(1)]$, that is, for all $\varepsilon > 0$ it holds

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^n w_i T_i Y_i - \mathbf{E}[Y(1)]\right| \geq \varepsilon\right] \;\to\; 0$$

for $n \to \infty$.

Remark. The entropy

$$f\colon (0,\infty) \to \mathbb{R}\,, \quad x \mapsto x \log x$$

is a prevailing choice. Then the requirement on π is

$$\mathbf{E}[(\log \pi(X))^2] < \infty\,.$$

This is a very weak requirement. Likewise, the requirement for the variance

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

is

$$\mathbf{E}[1/\pi(X)^2] < \infty\,.$$

♢

Remark. Both partitioning estimates are covered by [GKKW02, §4], because we only estimate quantities depending on X, that is $f^{-1}(1/\pi(X))$ and $\mathbf{E}[Y(1)|X]$. The corresponding coefficients in the partitioning estimate are then $f^{-1}(1/\pi_i)$ and $Y_i(1)$, which is the potential outcome of unit i under treatment. Note, that both quantities are generally unknown to us, but we can nevertheless leverage there existence in the proof. ♢

Takeaways To ensure double robustness, we leverage primal and dual optimization problem. The dual problem solves propensity score estimation and the primal problem the bias of the final estimate. Partitioning estimates as in [GKKW02].

1.2 Double Robustness

By double robustness we mean the property of an estimator that it is consistent if either one of treatment or outcome model is well specified. The augmented weighting estimator was designed to have exactly this property. Surprisingly, the weighted mean estimator in the balancing weights approach of [WZ19] retains this feature despite its simplicity [ZP17]. In the following we explore double robustness in the weighted mean estimator from the perspective of learning rates. We will adopt a similar notion of learning rates as in [SC08]. To the best of our knowledge this approach is new. We investigate how learning rates change in different scenarios.

1.2.1 Learning Rates of the weighted mean

What is the speed of convergence in the weak law of large numbers? The next statement gives a clear-cut answer: The arithmetic mean of independent, identically distributed, square-integrable random variables learns with rate $n^{-1/2}$. Furthermore, using Bienaymé's formula and Chebyshev's inequality, the statement is easy to prove (cf. [Kle20, Theorem 5.14]).

Theorem. *Let $X_1, X_2, \dots$ be i.i.d, square-integrable random variables with $V := \mathbf{Var}[X_1] < \infty$. Then, for any $\tau \in (0,1]$ and all $n \in \mathbb{N}$, we have*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}[X_i])\right| \le \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \ge 1-\tau\,. \tag{1.5}$$

Reflection. Bernsteins's inequality yields better confidence. ♠

Theorem 1.3. (Bernstein's inequality) *Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, $B > 0$ and $\sigma > 0$ be real numbers, and $n \ge 1$ be an integer. Furthermore, let $X_1, \dots, X_n : \Omega \to \mathbb{R}$ be independent random variables satisfying $\mathbf{E}[X_i] = 0$, $\|X_i\|_\infty \le B$ and $\mathbf{E}[X_i^2] \le \sigma^2$ for all $i = 1, \dots, n$. Then we have*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right| \le \sqrt{\frac{2\,\sigma^2\log(e/\tau)}{n}} + \frac{2B\log(e/\tau)}{3n}\right] \ge 1-\tau\,, \qquad \tau > 0\,.$$

Proof. Confer [SC08, Theorem 6.12] for the one-sided version. The two-sided version, as stated in the above theorem, is an easy consequence. We omit the details. □

$\sqrt{2}(\sigma \vee B)\frac{1}{\sqrt{n}}\log(e/\tau)$

Improve quality of presentation in order to merge this part with more cultivated sections of the thesis.

Deriving learning rates in this way, we call for *observed outcomes* of the treated matching *marginal potential outcomes* under treatment, that is,

$$Y(1)\,|\,T = 1 \;\sim\; \mathbf{P}_{Y(1)}\,. \tag{1.6}$$

In practice, virtually every scenario violates this assumption. Compare the health of an otherwise thriving smoker with that of an asthmatic nonsmoker. Or the dropout rates of study programs with and without compulsory entrance examinations. Indeed, any unbalanced external influence on both treatment and outcome ruins the above assessment. We speak of a **confounded** scenario, and the unbalanced external influence is called a **confounder**. When regarding the health risks of smoking, indicators of

asthma, such as frequent cough or shortness of breath, are confounders. We emphasize that they are so, not merely by being signs of bad health, but because asthmatics are also less likely to smoke.

Let us take a moment to understand the word "to confound".

Theorem 1.4. (Work in progress) *The verb to confound has nuances of meaning [zot]*

to throw (a person) into confusion

in this case the Statistician or whoever wishes to inform his decisions on the basis of data.

refute

A change of effect when stratifying is a strong refutation.

to put to shame

To embarrass some on who used invalid analysis as an argument in a public debate.

damn

to fail to discern differences between : mix up

Maybe in the sense of conflate: see causal effects where there are non.

to increase the confusion of

after a change of effect. what should be trusted?

baffle, frustrate

yield to this and stop asking causal questions.

archaic : to bring to ruin

ruin the analysis, assesment.

obsolete : consume, waste

waste the effort of conducting the analysis.

Short, confounding is nasty.

Encouraging practicioners to address confounding in their investigations, the statistical literature coins the term *Simpson's paradox*. It relates to surprising changes of effect in subclasses of the data (cf. [Wag82] for a brief discussion of Simpson's paradox and some real world examples). Use non-probabilistic frameworks, such as causal framework. Despite widespread acceptance, their use in empirical data analysis is still subject to debate [Pea09, §6].

What solutions were offerd to deal with confounding? Start from propensity scores.

Statisticians have wrestled with this issue for nearly a century.

In experimental studies we usually specify treatment assignment as opposed to merely

observing a unit receiving treatment.

The next statement makes use of the propensity score.

Theorem. *Consider the weighted mean estimator with weights*

$$w_i \;=\; \frac{1}{n}\frac{T_i}{\pi(X_i)}\,. \tag{1.7}$$

Denote $V := \mathbf{E}[(Y(1))^2\,/\,\pi(X)] - \mathbf{E}[Y(1)]^2$ *. Assume that weak unconfoundedness holds. Then, for any* $\tau \in (0,1]$ *and all* $n \in \mathbb{N}$*, we have*

$$\mathbf{P}\left[\left|\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \;\leq\; \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \;\geq\; 1-\tau\,. \tag{1.8}$$

Proof. We want to reinforce coherent use of the weak law of large numbers. To this end, we verify

$$\begin{aligned} n\,\mathbf{E}[w(T,X)\,Y(T)] \;&=\; \mathbf{E}[Y(1)]\,,\\ n^2\,\mathbf{Var}[w(T,X)\,Y(T)] \;&=\; \mathbf{E}[(Y(1))^2\,/\,\pi(X)] - \mathbf{E}[Y(1)]^2\,. \end{aligned}$$

Essentially, the random weight $w(T,X)$ acts on $Y(T)$ through $T\,/\,\pi(X)$. It does so by inducing independence of observed outcome $Y(T)$ and treatment T. This requires that weak unconfoundedness holds, i.e.,

$$(Y(0),Y(1)) \perp\!\!\!\perp T \,|\, X\,. \tag{1.9}$$

To showcase the details we added an n and n^2 factor in the above display. The calculations go as follows.

$$\begin{aligned} n\,\mathbf{E}[w(T,X)\,Y(T)] \;&=\; \mathbf{E}\left[Y(T)\cdot(T\,/\,\pi(X))\right]\\ &=\; \mathbf{E}\left[Y(1)\,/\,\pi(X)\,|\,T=1\right]\cdot\mathbf{P}[T=1]\\ &=\; \int_{\mathcal{X}} \mathbf{E}\left[Y(1)\,|\,X=x,T=1\right]\cdot(\mathbf{P}[T=1]\,/\,\pi(x))\,\mathbf{P}_{X|T}(dx\,|\,1)\\ &=\; \int_{\mathcal{X}} [Y(1)|X=x]\,\mathbf{P}_X(dx) = \mathbf{E}[Y(1)]. \end{aligned} \tag{1.10}$$

The first equality holds because of the definition of the weights. The second, third and last equality stem from $T \in \{0,1\}$, and the law of total expectation, applied with T and X. The fourth equality is justified by the assumption of weak unconfoundedness. The density transformation is due to Bayes's Theorem. With slight modifications in the above argument, it follows

$$\mathbf{E}\left[\Big(Y(T)\cdot(T\,/\,\pi(X))\Big)^2\right] \;=\; \mathbf{E}\Big[(Y(1))^2\,/\,\pi(X)\Big]\,. \tag{1.11}$$

We omit the details. Invoking the weak law of large numbers finishes the proof. □

Formulate improved version with Bernstein.

Formulate better transition to weights with estimated propensity score.

We started by asking an easy question, so it is time for a more challenging one: How do we proceed in deriving learning rates if the propensity score is unknown. How do we generally procede? [what has been done in the past. Why are some methods obsolete] A naive answer would be: We hope to select a proper model and try to estimate the propensity score. Stunningly, a lot of practicioners still settle for obsolete methods when it comes to propensity score analysis.

Next, we consider the event that we have a consistent estimator of the propensity score and the distribution of the covariate vector X, along with that of the outcome Y, has compact support. The next assumptions reduce the technical task to a minimum.

Assumptions. *Let the following hold.*

(i) *There exists* $C_Y \geq 1$ *such that* $|Y(1)| \leq C_Y$ *almost surely.*

(ii) *There exists* $C_\pi > 0$ *such that* $C_\pi < \pi(X)$ *almost surely.*

(iii) *There exists a function class* $\mathcal{F}$ *with unit ball* $B_\mathcal{F} := \{f \in \mathcal{F} \colon \|f\|_\infty \leq 1\}$ *such that* $\log N_{[]}(\varepsilon, B_\mathcal{F}, L_2(\mathbf{P})) \leq C_\mathcal{F}(1/\varepsilon)^{1/k}$ *for some* $k > 1/2$ *and some constant* $C_\mathcal{F} \geq 1$.

(iv) *The random function* f_w *defined by* $f_w(T, X, Y) := \left(n\, w(X) - \frac{1}{\pi(X)}\right) T\, Y$ *satisfies* $f_w \in \mathcal{F}$ *almost surely .*

(v) *There exist a learning rate* (r_n), *confidence constants* (γ_τ) *and uniform constant* $C_w \geq 1$ *such that for all* $\tau \in (0, 1]$ *and all* $n \in \mathbb{N}$ *it holds* $\mathbf{P}\left[\left\|w(X) - \frac{1}{\pi(X)}\right\|_\infty \leq C_w c_\tau \varepsilon_n\right] \geq 1 - \tau$

(vi) *There exists* $\alpha > 1$ *such that* $\varepsilon_n \cdot c_{n^{-\alpha}} \to 0$ *as* $n \to \infty$ *and* $\varepsilon_n \cdot c_{n^{-\alpha}} \leq 1$ *for all* $n \in \mathbb{N}$.

Elaborate on assumptions. For example (iii) [vdvW13, §2.7.1] [vdV00, §19.9]

Theorem. *Let the assumptions. Then the weighted mean learns with rate* (ε_n) *defined by*

$$\varepsilon_n := \inf\left\{ t \in (0, 1] : r_n \cdot \gamma_{t/n} \leq t \right\} \wedge 1 \qquad \text{for all } n \in \mathbb{N}.$$

Furthermore, it has confidence constants (c_τ) *given by*

$$c_\tau = \gamma_\tau / \tau \tag{1.12}$$

and uniform constant

$$C_{\mathbf{P}} = \max\left\{\sqrt{C_{\mathcal{F}}}C_w, C_Y C_w, \frac{C_Y}{C_\pi}\right\} \tag{1.13}$$

Proof. We consider the following error decomposition.

$$\begin{aligned}
&\sum_{i=1}^{n} w_i T_i\, Y_i \;-\; \mathbf{E}[Y(1)] \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)}\,(Y_i - \mathbf{E}[Y(1)]\,) \;+\; \sum_{i=1}^{n} T_i\left(w_i - \frac{1}{n\,\pi(X_i)}\right) Y_i \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)}\,(Y_i - \mathbf{E}[Y(1)]\,) \\
&\qquad\qquad + \;\frac{1}{\sqrt{n}}\,\mathbb{G}_n f_w \\
&\qquad\qquad\quad - \;\mathbf{E}[f_w(T,X,Y)\,].
\end{aligned}$$

We already bounded the first term. To bound the remaining terms we will use the learning rates of w. To this end, we employ maximal inequalities for empirical processes to bound the second term. We bound the third term by the law of total expectation and balancing learning rates and confidence.

2nd term

Denote $\mathcal{F}_{n,\tau} := (C_Y C_w \gamma_\tau r_n)\cdot B_{\mathcal{F}} =: \delta_{n,\tau}\cdot B_{\mathcal{F}}$. It holds by maximal inequalities

$$\begin{aligned}
\mathbf{E}^*\left[\|\mathbb{G}_n\|_{\mathcal{F}_{n,\tau}}\right] \;&\le\; \int_0^{\delta_{n,\tau}} \sqrt{\log N_{[\,]}\left(\varepsilon/\delta_{n,\tau}, B_{\mathcal{F}}, L_2(\mathbf{P})\right)}\,d\varepsilon \\
&\le\; \int_0^{\delta_{n,\tau}} \left(\frac{\delta_{n,\tau}}{\varepsilon}\right)^{1/(2k)} d\varepsilon \;=\; \delta_{n,\tau}\,.
\end{aligned}$$

For $t>0$, Markov's inequality gives

$$\mathbf{P}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}} \ge t\right] \;\le\; \frac{1}{t}\,\mathbf{E}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}}\right] \;\le\; \frac{1}{t}\,\mathbf{E}^*\left[\|\mathbb{G}_n\|_{\mathcal{F}_{n,\tau}}\right] \le \frac{\delta_{n,\tau}}{t}\,, \tag{1.14}$$

and consequently

$$\mathbf{P}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}} \le \frac{1}{\tau} C_Y C_w \gamma_\tau r_n\right] \;\ge\; 1-\tau \tag{1.15}$$

Next, note that

$$\|f_w\|_\infty \le C_Y \left\| nw - \frac{1}{\pi(X)}\right\| \le C_Y C_w \gamma_\tau r_n \tag{1.16}$$

with probability greater than $1-\tau$. Thus $f_w \in \mathcal{F}_{n,\tau}$ with probability greater than $1-\tau$. It follows

$$\mathbf{P}\left[\mathbb{G}_n f_w \leq \frac{1}{\tau} C_Y C_w \gamma_\tau r_n\right] \geq 1 - 2\tau\,. \tag{1.17}$$

Streamline analysis of second term.

3rd term

We localize with regards to $f_w \in \mathcal{F}_{n,\tau}$. We require the weights to be smalle that 1, such that we always have $\left\|nw - \frac{1}{\pi}\right\| \leq n + \frac{1}{C_\pi}$.

$$\mathbf{E}[f_w] \leq C_Y C_w \gamma_\tau r_n (1-\tau) + C_Y (n + \frac{1}{C_\pi})\tau \tag{1.18}$$

$$\leq C_Y \left(C_w \gamma_\tau r_n + \left(n + \frac{1}{C_\pi}\right)\tau\right)\,. \tag{1.19}$$

If we choose $\tau = \varepsilon_n$ we get

$$\mathbf{E}[f_w] \leq C_Y \left(C_w + 1 + \frac{1}{C_\pi}\right)\varepsilon_n\,. \tag{1.20}$$

Selecting the worst instance of learning rate, confidence and uniform constant concludes the proof.

Streamline analysis of third term.

□

Reflection. Why did we use empirical process theory instead of conventional concentration inequalities? The weights w are random, so f_w is random as well.

Best is $\gamma_\tau = 1$, when we recover the learning rate of the estimator, that is, (r_n). The confidence γ_τ/τ is substandard. Improvements may involve Bernstein like concentration for empirical processes (cf. [vdvW13, Section 2.14.2]) ♠

Continue section with rates for right outcome model and then both models right. Do the rates improve?

Next we need to assume that the weights and outcome are independent given treatment and covariates, that is,

$$Y(1) \perp\!\!\!\perp w \mid X, T\,. \tag{1.21}$$

Also the bias of the outcome regression has to be bounded by the weights. We get the following error decomposition

$$\frac{1}{n}\sum_{i=1}^{n} n w_i T_i (Y_i - \mathbf{E}[Y(1)|X_i]) + (\mathbf{E}[Y(1)|X_i] - \mathbf{E}[Y(1)]) \tag{1.22}$$

$$+ \sum_{i=1}^{n} (T_i w_i - 1/n)(\mathbf{E}[Y(1)|X_i] - B(X_i)\lambda) \tag{1.23}$$

$$+ \sum_{i=1}^{n} (w_i - 1/n) B(X_i)\lambda \tag{1.24}$$

The expectation of summand in the first term is zero. Hence the convergence by wlln or more refined methods. The second term goes to 0 by the consistency of the outcome regression. The third term is bounded by the weights.

$$\mathbf{E}[Tw \cdot (Y(T) - \mathbf{E}[Y(1)\,|\,X]\,)] \;=\; \mathbf{E}[T \cdot \mathbf{E}[w\,|\,T,X] \cdot \mathbf{E}[Y(T) - Y(1)\,|\,T,X]\,] \;=\; 0$$

The first equality stems from the conditional independence assumptions. The resulting term vanishes because the difference $Y(T) - Y(1)$ does so after conditioning on $T = 1$.

$$\left|\sum_{i=1}^{n} (T_i w_i - 1/n)(\mathbf{E}[Y(1)|X_i] - B(X_i)\lambda)\right| \leq 2 \max_{i=1,\ldots,n} |\mathbf{E}[Y(1)|X_i] - B(X_i)\lambda| \tag{1.25}$$

$$\leq 2\,\text{lerning rate triple of regression} \tag{1.26}$$

with probability greater than $1 - \tau$. Note, that the weights sum to 1. Assume $|\sum_{i=1}^{n}(T_i w_i - 1/n) B_k(X_i)| \leq \delta_k$ for all $k = 1, \ldots, K$ and the deltas go to zero with some rate. The by Cauchy-Schwarz it follows

$$\left|\sum_{i=1}^{n} (T_i w_i - 1/n) B(X_i) \cdot \lambda\right| \leq |\langle(\delta_k), \lambda\rangle| \leq \text{rate of deltas} \cdot \|\lambda\| \tag{1.27}$$

So we assume $\lambda \in \Theta$, where the parameter space Θ is compact or grows moderately with K.

Introduce concept of semiparametric efficiency. For the semiparametric efficiency bound of propensity score weighting, [Hah98] is a good reference. General introduction to semiparametric models see [vdV00, §25].

Takeaways Each error decomposition furnishes information about the asymptotic properties of the weighted mean. We always get consistency, but learning rates may differ. Obtaining good confidence depends on employing more refined concentration inequalities like Bernstein's inequality. If both treatment and outcome model are well specified we reach the semi-parametric efficiency bound of weighting with the true inverse propensity score.

1.3 Error Decompositions

In this section we extend the weighting scheme to estimating the marginal distribution of the outcome. We extend the error decomposition in [WZ19, page 27]

The following decomposition is flexible in Φ. We get different causal estimands $\mathbf{E}[\Phi(Y(1))]$, for example, the population average of $Y(1)$ for $\Phi(Y) = Y$, that is, $\mathbf{E}[Y(1)]$. Likewise we get the distribution function of $Y(1)$ at t for $\Phi(Y) = \mathbf{1}_{(-\infty,t]}(Y)$, that is, $\mathbf{P}[Y(1) \leq t]$.

$$\sum_{i=1}^{n} w_i T_i \Phi(Y_i) - \mathbf{E}[\Phi(Y(1))] = \frac{1}{n}\sum_{i=1}^{n} S_i + R_0 + R_1 + R_2, \tag{1.28}$$

where

$$\begin{aligned}
S_i &:= \frac{T_i}{\pi_i}\left(\Phi(Y_i) - \mathbf{E}[\Phi(Y_i(1))|X_i]\right) + \left(\mathbf{E}[\Phi(Y_i(1))|X_i] - \mathbf{E}[\Phi(Y(1))]\right) \quad \text{for } i \in \{1,\dots,n,\},\\
R_0 &:= \sum_{i=1}^{n} T_i \left(w_i - \frac{1}{n\pi_i}\right)\left(\Phi(Y_i) - \mathbf{E}[\Phi(Y_i(1))|X_i]\right),\\
R_1 &:= \sum_{i=1}^{n} \left(T_i w_i - \frac{1}{n}\right)\left(\mathbf{E}[\Phi(Y_i(1))|X_i] - B(X_i)^\top \lambda\right),\\
R_2 &:= \sum_{i=1}^{n} \left(T_i w_i - \frac{1}{n}\right) B(X_i)^\top \lambda \quad \text{for } \lambda \in \mathbb{R}^K.
\end{aligned}$$

We can even view $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i$ as an empirical process $\mathbb{G}_n f$ indexed over

$$f_\Phi(T, X, Y) = \frac{T}{\pi(X)}\left(\Phi(Y) - \mathbf{E}[\Phi(Y)|X]\right) + \mathbf{E}[\Phi(Y)|X]. \tag{1.29}$$

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

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 .

1.4 Application of Matrix Concentration Inequalities

We extend the analysis to covariates with unbounded support. To the best of our knowledge this has not been done in this setting. Indeed, [WZ19, Assumption 1] only allows for covariates with bounded support. Our idea is to apply Matrix Moment Inequalities instead of Bernstein's inequality. We learned from this possibility from [Tro15, (6.1.6)]

Analysis of $\mathbf{E}[\max_{i\leq n} \|A_i\|^2]$

We start from the premise that the fourth moment of the random quantities $B_k(X_i)$ and $1/\pi_i$ is uniformly bounded in k and i.

Assumptions. *(i) There exists a constant $C_B \geq 1$ such that*

$$\mathbf{E}\left[B_k(X_i)^4\right] \;\leq\; C_B \quad \textit{for all } (k,i) \;\in\; \{1,\dots,K\}\times\{1,\dots,n\}.$$

(ii) There exists a constant $C_\pi \geq 1$ such that

$$\mathbf{E}\left[\left(\frac{1}{\pi_i}\right)^4\right] \;\leq\; C_\pi \quad \textit{for all } i \;\in\; \{1,\dots,n\}.$$

Note, that these assumptions allow for covariate distributions with unbounded support. The coming example ought to reinforce this observation.

Example. Let us assume a logistic regression model for the propensity score. Then there exist coefficients $\vartheta \in \mathbb{R}^N$ and $\vartheta_0 \in \mathbb{R}$ (N is the number of covariates and ϑ_0 is the intercept of the model) such that

$$1\,/\,\pi(X) \;=\; 1 \;+\; \exp(\,\vartheta_0 \,+\, \langle \vartheta, X\rangle\,)\,, \tag{1.30}$$

$$\mathbf{E}\left[\left(\frac{1}{\pi(X)}\right)^4\right] \;=\; \sum_{j=1}^{4} \binom{4}{j} e^{j\vartheta_0}\, M_X(\,j\,\vartheta\,), \tag{1.31}$$

where M_X is the momement-generating function of the random vector X (we assume it exists.). While the quantity in (1.30) may be unbounded when X has unbounded

support, that in (1.31) remains bounded for moderate decay rates of the underlying distribution.

The multivariate normal distribution with location parameter $\mu \in \mathbb{R}^N$ and covariance matrix $\Sigma \in \mathbb{M}_N$ has momement-generating function

$$M_{\mathcal{N}(\mu,\,\Sigma)}(t) \;=\; \exp\left(\langle\, t, \mu \,\rangle + \frac{\langle\, t, \Sigma\, t \,\rangle}{2} \right) \quad \text{for all } \; t \;\in\; \mathbb{R}^N .$$

In particular, the expression in (1.31) is finite if X follows a multivariate normal distribution.

Likewise we give a negative example. Let $N = 1$ and $X \sim \mathrm{Exp}(\lambda)$, that is, only one exponentially distributed covariate is taken into account. The momement-generating function is then confined to take arguments $t < \lambda$, and thus (1.31) becomes pointless if $4\,\vartheta \geq \lambda$. ♢

Next, we recall the entity we want to examin.

$$A_i \;=\; \frac{1}{n} \left(1 \;-\; \frac{T_i}{\pi_i} \right) B(X_i) \qquad \text{for } i \in \{1, \ldots, n\} .$$

For all $i \in \{1, \ldots, n\}$ we get the bound

$$\left| 1 \;-\; \frac{T_i}{\pi_i} \right| \;\leq\; \left(1 \;\vee\; \frac{1 - \pi_i}{\pi_i} \right) \;\leq\; 1 \;+\; \frac{1 - \pi_i}{\pi_i} \;=\; \frac{1}{\pi_i} . \tag{1.32}$$

Let $i^* \in \{1, \ldots, n\}$ be the index where $\|A_i\|$ attains its maximum.

$$\begin{aligned} \mathbf{E}\left[\max_{i \leq n} \|A_i\|^2 \right] \;&=\; \mathbf{E}\left[\|A_{i^*}\|^2 \right] \;\leq\; \mathbf{E}\left[\left(\frac{\|B(X_{i^*})\|}{\pi_{i^*}} \right)^2 \right] \;/\; n^2 \\ &\leq\; \mathbf{E}\left[\left(\frac{1}{\pi_{i^*}} \right)^4 \right]^{1/2} \cdot\; \mathbf{E}\left[\|B(X_{i^*})\|^4 \right]^{1/2} \;/\; n^2 \\ &\leq\; K \,/\, n^2 \;\cdot\; \sqrt{C_\pi C_B} . \end{aligned} \tag{1.33}$$

The first inequality comes from the bound (1.32). The Cauchy-Schwarz inequality provides the second inequality. In the last step we use the assumptions made at the start of the section. Paying the price of an extra n factor, the maximal inequality (1.33) yields a bound of the sum, that is,

$$\sum_{i=1}^{n} \mathbf{E}\left[\|A_i\|^2 \right] \;\leq\; \frac{K}{n} \sqrt{C_\pi C_B}$$

Assumption 1.1. .

Assumption 1.2. .

Remark. With Assumption we also get a bound on the fourth moment of $\|B(X_i)\|$. Indeed, by the convexity of $x \mapsto x^2$, the monotonicity and linearity of the expectation it holds

$$\begin{aligned}\mathbf{E}[\|B(X_i)\|^4] &= \mathbf{E}\left[\left(\sum_{k=1}^K B_k^2(X_i)\right)^2\right] = K^2\mathbf{E}\left[\left(\sum_{k=1}^K \frac{1}{K}B_k^2(X_i)\right)^2\right] \leq K^2\mathbf{E}\left[\sum_{k=1}^K \frac{1}{K}B_k^4(X_i)\right] \\ &= K\sum_{k=1}^K \mathbf{E}\left[B_k^4(X_i)\right] \leq K^2 C_B\end{aligned}$$

♢

Analysis of $v(\mathbf{S})$

We use the fact that $\|A\|_2 \leq \|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ It holds

$$\sum_{i=1}^n \mathbf{E}[A_i A_i^\top] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B(X_i)B(X_i)^\top\right] = \frac{1}{n^2}\left(\sum_{i=1}^n \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B_k(X_i)B_l(X_i)\right]\right)_{1\leq k,l\leq K}.$$

Thus

$$\begin{aligned}&\left\|\sum_{i=1}^n \mathbf{E}[A_i A_i^\top]\right\|_2^2 \\ &\leq \left\|\sum_{i=1}^n \mathbf{E}[A_i A_i^\top]\right\|_F^2 = \frac{1}{n^4}\sum_{k,l=1}^K \left(\sum_{i=1}^n \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B_k(X_i)B_l(X_i)\right]\right)^2 \\ &\leq \frac{1}{n^4}\sum_{k,l=1}^K \left(\sum_{i=1}^n \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right]^{\frac{1}{2}} \mathbf{E}[B_k(X_i)^4]^{\frac{1}{4}}\mathbf{E}[B_l(X_i)^4]^{\frac{1}{4}}\right)^2 \leq \left(\frac{K}{n}\right)^2 C_\pi C_B\end{aligned}$$

On the other hand

$$\begin{aligned}\left\|\sum_{i=1}^n \mathbf{E}[A_i^\top A_i]\right\|_2 &= \sum_{i=1}^n \mathbf{E}[A_i^\top A_i] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 \|B(X_i)\|_2^2\right] \\ &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right]^{\frac{1}{2}} \|B(X_i)\|_2^4]^{\frac{1}{2}} \leq \frac{K}{n}\sqrt{C_\pi C_B}\end{aligned}$$

It follows

$$v(\mathbf{S}) \leq \frac{K}{n}\sqrt{C_\pi C_B}$$

Thus we can apply Theorem 3.3 to get

$$\mathbf{E}[\|\mathbf{S}\|_2] \leq \sqrt{2e\frac{K}{n}\sqrt{C_\pi C_B}\log(K+1)} + 4e\frac{\sqrt{K}}{n}\sqrt[4]{C_\pi C_B}\log(K+1) \leq 14 C_\pi C_B\sqrt{\frac{K\log(K+1)}{n}}$$

Give explanation why simple estimates are insufficient. See notes.

1.5 Continuous Treatment

We introduce the measure of proximity

$$d_n(t,s) := \frac{\mathbf{1}_{s\in N_n(t)}}{\lambda[N_n(t)]} \tag{1.34}$$

where $N_n(t)$ is a neighborhood of t with $\lambda[N_n(t)] \to 0$ for $n \to \infty$. If we can apply the dominated convergence theorem we get

$$\mathbf{E}\left[\frac{d_n(t,T_i)}{h_{T|X}(t,X_i)}\right] = \mathbf{E}\left[\frac{\mathbf{P}[T_i \in N_n(t)|X_i]}{\lambda[N_n(t)]} \cdot \frac{1}{h_{T|X}(t,X_i)}\right] \to \mathbf{E}\left[\frac{h_{T|X}(t,X_i)}{h_{T|X}(t,X_i)}\right] = 1\,. \tag{1.35}$$

Furthermore, if $Z_i \sim Y(t)|T_i$ we have

$$\mathbf{P}\left[d_n(t,T_i)\,|Y(T_i) - Z_i| \geq \varepsilon\right] = \mathbf{P}[T_i \in N_n(t)] \cdot \mathbf{P}\left[|Y(T_i) - Z_i| \geq \varepsilon \cdot \lambda(N_n(t))\right]\,. \tag{1.36}$$

If T is continuously distributed, $\mathbf{P}[T_i \in N_n(t)] \to 0$ which implies $d_n(t,T_i)\,(Y(T_i) - Z_i) \to 0$ in probability. We shall employ concentration inequalities to also derive learning rates. The optimization problem is for fixed $t \in \mathcal{T}$

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

subject to the constraints

$$\left|\frac{1}{n}\sum_{i=1}^{n}(w_i \cdot d_n(t,T_i) - 1) \cdot B_k(X_i)\right| \;\leq\; \delta_k, \qquad k = 1,\ldots,K$$

We shall derive

$$d_n(t,T_i)\cdot\left|w_i-\frac{1}{h_{T|X}(t,X_i)}\right|\to 0 \tag{1.37}$$

in probability for all . From this we control the error

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}d_n(t,T_i)w_iY_i-\mathbf{E}[Y(t)]\right|\\
&\leq\ \left|\frac{1}{n}\sum_{i=1}^{n}(w_id_n(t,T_i)-1)\langle B(X_i)),\mathbf{Y}(t)\rangle\right|\\
&\quad+\ \left|\frac{1}{n}\sum_{i=1}^{n}(w_id_n(t,T_i)-1)\left(\mathbf{E}[Y(t)|X_i]-\langle B(X_i)),\mathbf{Y}(t)\rangle\right)\right|\\
&\quad+\ \left|\frac{1}{n}\sum_{i=1}^{n}d_n(t,T_i)\cdot(w_i-1/h_{T|X}(t,X_i))\left(Z_i-\mathbf{E}[Y(t)|X_i]\right)\right|\\
&\quad+\ \left|\frac{1}{n}\sum_{i=1}^{n}h_T(t)/h_{T|X}(t,X_i)\left(Z_i-\mathbf{E}[Y(t)|X_i]\right)\right|\\
&\quad+\ \left|\frac{1}{n}\sum_{i=1}^{n}\left(h_T(t)-d_n(t,T_i)\right)/h_{T|X}(t,X_i)\left(Z_i-\mathbf{E}[Y(t)|X_i]\right)\right|\\
&\quad+\ \left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{E}[Y(t)|X_i]-\mathbf{E}[Y(t)]\right)\right|\\
&\quad+\ \left|\frac{1}{n}\sum_{i=1}^{n}w_id_n(t,T_i)\left(Y_i-Z_i\right)\right|
\end{aligned}$$

$$\begin{aligned}
&d_n(t,T_i)\cdot\left|w_i-1/h_{T|X}(t,X_i)\right|\\
&=\ d_n(t,T_i)\cdot\left|(f^{'})^{-1}\langle B(X_i),\lambda^{\dagger}\rangle-(f^{'})^{-1}\left(f^{'}\left(1/h_{T|X}(t,X_i)\right)\right)\right|\\
&\leq\ d_n(t,T_i)\left|(f^{'})^{-1}\langle B(X_i),\lambda^{\dagger}\rangle-(f^{'})^{-1}\langle B(X_i),f^{'}\left(1/h_{T|X}(t,X)\right)\rangle\right|\\
&\quad+\ d_n(t,T_i)\left|(f^{'})^{-1}\langle B(X_i),f^{'}\left(1/h_{T|X}(t,X)\right)\rangle-(f^{'})^{-1}\left(f^{'}\left(1/h_{T|X}(t,X_i)\right)\right)\right|\\
&\leq\ \omega\left((f^{'})^{-1},\left\|\lambda^{\dagger}-f^{'}\left(1/h_{T|X}(t,X)\right)\right\|\right)\ +\ \omega\left((f^{'})^{-1},\varepsilon_m\right)\\
&\leq\ \omega\left((f^{'})^{-1},\varepsilon_{\dagger}\right)\ +\ \omega\left((f^{'})^{-1},\varepsilon_m\right)\\
&\leq\ \varepsilon
\end{aligned}$$

with probability tending to 1.

$$\begin{aligned}
&G(f^{'}\left(1/h_{T|X}(t,X)\right)+\Delta)\ -\ G(f^{'}\left(1/\pi f^{'}\left(1/h_{T|X}(t,X)\right)\right))\\
&\geq\ -\|\Delta\|\left(\|\delta\|\ +\ \|B(X_i)\|\cdot\frac{1}{n}\sum_{i=1}^{n}\left|1\ -\ d_n(t,T_i)\cdot(f^{'})^{-1}\langle B(X_i),f^{'}\left(1/h_{T|X}(t,X)\right)\rangle\right|\right)
\end{aligned}$$

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 conjugate sum 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 well known Fenche-Rockafellar Duality theorem is a corollary of conjugate sum and chain rule. It gives general conditions under which dual and primal values coincide. The material we present is very well known, so we claim no originality. We paraphrase the approach of [MMN22] to Duality. As an introduction, we recommend this recently published book together with the 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.

Given (not necessary convex) sets $\Omega, \Omega_1, \Omega_2 \subseteq \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, define the **set addition** and **multiplication** by a real scalar as $\Omega_1 + \Omega_2 := \{x_1 + x_2 \colon x_1 \in \Omega_1, x_2 \in \Omega_2\}$ and $\lambda\Omega := \{\lambda x \colon x \in \Omega\}$. For convex sets the addition and multiplication by a real scalar are convex.

Throughout this section, we shall denote by $B := \left\{x = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n \colon (\sum_{i=1}^n x_i^2)^{1/2} \leq 1\right\}$

Solve editorial issue with ball.

the **Euclidian unit ball** in $\mathbb{R}^n$. This is a closed convex set. For any $a \in \mathbb{R}^n$, the **ball with radius** $\varepsilon > 0$ **and center** a is given by $\left\{a + x \in \mathbb{R}^n \colon (\sum_{i=1}^n x_i^2)^{1/2} \leq \varepsilon\right\} = a + \varepsilon B$. For any set Ω in $\mathbb{R}^n$, the set of points x whose distance from Ω does not exceed ε is $\Omega + \varepsilon B$. The **closure** $\operatorname{cl}(\Omega)$ and **interior** $\operatorname{int}(\Omega)$ of Ω can therefore be expressed by $\operatorname{cl}(\Omega) = \bigcap_{\varepsilon > 0} \Omega + \varepsilon B$ and $\operatorname{int}(\Omega) = \{x \in \Omega \colon \text{there exists } \varepsilon > 0 \text{ such that } x + \varepsilon B \subseteq \Omega\}$.

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. This concept is motivated by the fact that a line segment embedded in $\mathbb{R}^2$ does have a natural interior in $\mathbb{R}$ which is not a true interior in $\mathbb{R}^2$. The relative interior of C is defined as the interior which results when C is regarded as a subset of its affine hull.

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

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

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

(ii) The sets $\operatorname{cl} C$ *and* $\operatorname{ri} C$ *are convex*

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

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

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

(vi) 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* $\operatorname{ri} L^{-1}(C) = L^{-1}(\operatorname{ri} C)$.

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

(viii) $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$ *if and only if* $0 \notin \operatorname{ri}(C_1 - C_2)$.

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

To prove (vii) we use (iv). 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 (iv) 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 (iv), 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 (iv) 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. [MMN22, Theorem 2.92] □

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

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* $i = 1, \dots, m$, *we define the* ***infimal convolution*** *of these functions to be*

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

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

Lemma 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*

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

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*

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

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 and Fenchel-Rockafellar Theorem

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*

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

Add comment on nomenclature. What is Legendre transformation in this context?

Note that f in Definition 2.2 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. (Conjugate Chain Rule) *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*

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{2.8}$$

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, Proposition 4.28] □

Theorem 2.3. *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*

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*) \tag{2.9}$$

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

$$(f + g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{2.10}$$

Include lemma on convex conjugates of indicator functions. This should be straightforward.

Streamline example. Provide explanaition in the end. Confer [Roc70, bottom p.337]

Example. 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

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

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

$$(p_i^*)^{-1} \{(x_1^*, \ldots, x_n^*)\} \;=\; \begin{cases} \{x_i^*\}\,, & \text{if } x_j^* = 0 \text{ for all } j \neq i\,, \\ \emptyset & \text{else.} \end{cases} \tag{2.13}$$

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

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

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 $\inf \emptyset = \infty$.

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

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

By the conjugate sum rule it follows

$$(S_{f,n+1})^*(x_1^*,\dots,x_{n+1}^*) = (S_{f,n}+f_{n+1})^*(x_1^*,\dots,x_{n+1}^*) = ((S_{f,n})^* \square f_{n+1}^*)(x_1^*,\dots,x_{n+1}^*)$$
$$(f_1^* \square \cdots \square f_{n+1}^*)(x_1^*,\dots,x_{n+1}^*) = \sum_{i=1}^{n+1} f_i^*(x_i^*) = S_{f,n+1}^*(x_1^*,\dots,x_{n+1}^*) .$$

♢

Find right moment to introduce nomenclature for optimization problem. See also end of Tseng Bertsekas chapter.

Given proper convex functions $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a matrix $A \in \mathbb{R}^{n\times n}$, we define the primal minimization problem as follows:

Problem 2.1. *(Primal)* Given proper convex functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ and a matrix $A \in \mathbb{R}^{m\times n}$ we define the **primal optimization problem** to be

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \qquad f(x) + g(Ax)$$

Remark. Problem 2.1 appears in the unconstrained form. We can impose constraints by controling for the domains of f and g. To incorporate linear constraints $Ax \leq 0$ or more general constraints $x \in \Omega$, where Ω is a convex set, we can choose

$$g(x) = \delta_\Omega(x) := \tag{2.14}$$

where $x \notin \Omega$ leads to $f(x) + g(x) = \infty$ and the optimization problem (if feasible) will exclude x from the solutions. ♢

Problem 2.2. *(Dual)* Consider the same setting as in Problem 2.1. Using the convex conjugates of f, g and the transpose of A we define the **dual problem** of Problem 2.1 to be

$$\underset{y^*\in\mathbb{R}^m}{\text{maximize}} \qquad -f^*(A^\top y^*) - g^*(y^*).$$

Theorem 2.4. *Let f and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper convex functions and*

$$\operatorname{ri}(\operatorname{dom} g) \cap \operatorname{ri}(A \operatorname{dom} f) \neq \emptyset .$$

Then the optimal values of (2.1) *and* (2.2) *are equal, that is,*

$$\inf_{x\in\mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y\in\mathbb{R}^n} \left\{-f^*\left(A^T y\right) - g^*(-y)\right\} .$$

Proof. [MMN22, Theorem 4.63] □

Insert lemma in chapter 1.

Lemma 2.2. *Let* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be convex. Then for all* $y \in \mathbb{R}^n$ *and* $\varepsilon > 0$

$$\inf_{\|\Delta\|=\varepsilon} f(y+\Delta) - f(y) \geq 0 \tag{2.15}$$

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

Proof. Since $y + \varepsilon B$ is convex, it contains a local minimum of f . Suppose towards a contradiction that $y^* \in y + \varepsilon B$ is a local minimum, but not a global one, and (2.15) is true. Then it holds

$$f(x) < f(y^*) \quad \text{for some } x \in \mathbb{R}^n \setminus (y + \varepsilon B) \ . \tag{2.16}$$

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

$$\theta x + (1-\theta)y^* = y + \Delta_x \, . \tag{2.17}$$

It follows

$$\begin{aligned} f(y^*) \leq f(y) \leq f(y+\Delta_x) &= f(\theta x + (1-\theta)y^*) \\ &\leq \theta f(x) + (1-\theta) f(y^*) < f(y^*), \end{aligned} \tag{2.18}$$

which is a contradiction. The first inequality is due to y^* being a local minimum of f in $y + \varepsilon B$, the second inequality is due to (2.15) being true, the equality is due to (2.17), the third inequality is due to the convexity of f and the strict inequality is due to (2.16). Thus every local minimum of f in $y + \varepsilon B$ is also a global minimum. □

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 Tseng Bertsekas

We present the relevant parts of the paper [BT03].

Problem 2.3.

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & \mathbf{A}x \;\geq\; b\,. \end{array}$$

Assumptions. *Assume that the map* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *has the following properties.*

(i) f *is strictly convex.*

(ii) f *is lower-semicontinuous and continuous on* $\operatorname{dom} f$.

(iii) The convex conjugate f^* *of* f *is finite.*

The dual optimization problem of Problem 2.3 is

Problem 2.4.

$$\begin{array}{ll} \underset{y\in\mathbb{R}^m}{\text{maximize}} & \langle y, b\rangle \;-\; f^*(A^\top y) \\ \text{subject to} & y \;\geq\; 0\,. \end{array}$$

Let q denote the objective function of Problem 2.4, that is,

$$q \,:\, \mathbb{R}^m \to\; \mathbb{R}\,, \qquad y \;\mapsto\; \langle y, b\rangle \;-\; f^*(A^\top y)\,. \tag{2.19}$$

q is concave. The dual problem (D) is a concave program with nonnegativity constraints on the dual variable y_a of the inequality constraints in (P). Furthermore, strong duality holds for (P) and (D), that is, they have the same optimal value.

Since f^* is real-valued and f is strictly convex, f^* and q are continuously differentiable (cf. [Roc70, Theorem 26.3]).

Theorem. *A closed proper convex function is strictly convex if and only if its conjugate is continuously differentiable.*

What does closed mean and does f meet this condition?

We will denote the gradient of q at p by $d(p)$ and its i th coordinate by $d_i(p)$. Since q is continuously differentiable, $d_i(p)$ is continuous, and since q is concave, $d_i(p)$ as nonincreasing in p_i.

By differentiating and by using the chain rule, we obtain the dual cost gradient

$$d(p) = b - \mathbf{A}x, \quad \text{where} \quad x := \nabla f^*(\mathbf{A}^\top p) = \operatorname{argsup}_{\xi\in\mathbb{R}^m} \langle p, \mathbf{A}\xi\rangle - f(\xi). \tag{2.20}$$

Read first paper of tseng Bertsekas for equality constraints.

The last equality follows from Danskin's Theorem and [Roc70, Theorem 23.5]

Read and understand proof (p.80)

Proposition 2.3. (Danskin's Theorem [BT03, page 649]) *Let $Z \subseteq \mathbb{R}^m$ be a non-empty set, and let $\phi : \mathbb{R}^n \times Z \to \mathbb{R}$ be a continuous function such that $\phi(\cdot, z) : \mathbb{R}^n \to \mathbb{R}$, viewed as a function of its first argument, is convex for each $z \in Z$. Then the function*

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad x \mapsto \sup_{z \in Z} \phi(x, z) \tag{2.21}$$

is convex and has directional derivative given by

$$f'(x; y) = \sup_{z \in Z(x)} \phi'(x, z; y), \tag{2.22}$$

where $\phi'(x, z; y)$ is the directional derivative of the function $\phi(\cdot, z)$ at x in the direction y, and

$$Z(x) := \left\{ \overline{z} \in \mathbb{R}^m \colon \phi(x, \overline{z}) = \sup_{z \in Z} \phi(x, z) \right\}. \tag{2.23}$$

In particular, if $Z(x)$ consists of a unique point $\overline{z}$ and $\phi(\cdot, \overline{z})$ is differentiable at x, and $\nabla f(x) = \nabla_x \phi(x, \overline{z})$, where $\nabla_x \phi(x, \overline{z})$ is the vector with coordinates $(\partial \phi / \partial x_i)(x, \overline{z})$

Note that x is the unique vector satisfying

$$\mathbf{A}p \in \partial f(x). \tag{2.24}$$

From the optimality conditions for (D) it follows that a dual vector is an optimal solution of (D) if and only if

$$p = [p + d(p)]^+, \tag{2.25}$$

where $[\,\cdot\,]^+$ is the projection onto the positive orthant, i.e., $[y]^+ = [0 \vee y_1, \ldots 0 \vee y_n,]^\top$.

Provide details. See notes.

Complementary Slackness

We show that for all $i \in \{1, \ldots, n\}$ it holds

$$\begin{aligned} \text{either} \quad & p_i = 0 \quad \text{and} \quad d_i(p) \leq 0 \\ \text{or} \quad & p_i > 0 \quad \text{and} \quad d_i(p) = 0\,. \end{aligned}$$

Given an optimal dual solution p, we may obtain an optimal primal solution from the equation $x = \nabla f^*(\mathbf{A}^\top p)$. To see this, note that

$$\mathbf{A}x \geq b \quad \text{and} \quad p_i = 0 \quad \text{for all } i \text{ such that} \quad \sum_{j=1}^{m} a_{ij}x_j > b_i. \tag{2.26}$$

We can show that p and x satisfy the KKT conditions and thus x is an optimal solution to (P).

n

Definition 2.3. [Roc70, §28] *By an **ordinary convex program** (P) we mean an optimization problem of the following form*

$$\underset{x \in C}{\text{minimize}} \qquad f_0(x)$$

subject to the constraints

$$f_1(x) \leq 0, \ldots, f_r(x) \leq 0, \qquad f_{r+1}(x) = 0, \ldots, f_m(x) = 0, \tag{2.27}$$

where $C \subseteq \mathbb{R}^n$ is a non-empty convex set, f_i is a finite convex function on C for $i \in \{1, \ldots, r\}$ and f_i is an affine function on C for $i \in \{r+1, \ldots, m\}$.

Definition 2.4. *We define $[\lambda_1, \ldots, \lambda_m] \in \mathbb{R}^m$ to be a **Karush-Kuhn-Tucker (KKT) vector** for (P), if*

(i) $\lambda_i \geq 0$ for all $i \in \{1, \ldots, r\}$.

(ii) The infimum of the proper convex function $f_0 + \sum_{i=1}^{m} \lambda_1 f_i$ is finite and equal to the optimal value in (P).

Theorem 2.5. (Karush-Kuhn-Tucker conditions) *Let (P) be an ordinary convex program, $\overline{\alpha} \in \mathbb{R}^m$, and $\overline{z} \in \mathbb{R}^n$. Then $\overline{\alpha}$ is a KKT vector for (P) and $\overline{z}$ is an optimal solution to (P) if and only if $\overline{z}$ and the components α_i of $\overline{\alpha}$ satisfy the following conditions.*

(i) $\alpha_i \geq 0$, $f_i(\overline{z}) \leq 0$, and $\alpha_i f_i(\overline{z}) = 0$ for all $i \in \{1, \ldots, r\}$.

(ii) $f_i(\overline{z}) = 0$ for $i \in \{r+1, \ldots, m\}$.

(iii) $0_n \in [\partial f_0(\overline{z}) + \sum_{\alpha_i \neq 0} \alpha_i \partial f_i(\overline{z})]$.

Proof. [Roc70, Theorem 28.3] □

Takeaways Employing the Karush-Kuhn-Tucker conditions, we derive a dual relationship between optimal solutions for strictly convex functions.

3 Random Matrix Inequalities

In our application we want to bound moments of vector-valued random variables. For this we choose the theory of random matrix inequalities which lately received a lot of attention. In particular an approach via the method of exchangable pairs [MJC+14] has been fruitful in simplifying the proofs of long standing results such as the matrix Khintchin inequality. The paper offers a comprehensive introduction to this method.

We will cite the matrix Khintchin inequality and inequalities for moments of matrices that follow from it [CGT12]. As a novelty, we will apply intrinsic dimension results and Hermition Dilitation from [Tro15] to matrix moments inequalities. Even though it is straightforward, to the best of our knowledge the calculations have not been carried out in any publication so far.

3.1 A Matrix Analysis Primer

The **trace** of a square matrix, denoted by tr, is the sum of its diagonal entries, i.e. $\operatorname{tr}(\mathbf{B}) = \sum_{j=1}^{d} b_{jj}$ for $\mathbf{B} \in \mathbb{M}_d$. The trace is unitarily invariant, i.e. $\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\mathbf{Q}\mathbf{B}\mathbf{Q}^*)$ for all $\mathbf{B} \in \mathbb{M}_d$ for all unitary $\mathbf{Q} \in \mathbb{M}_d$. In particular, the existence of an eigenvalue value decomposition shows that the trace of a Hermitian matrix equals the sum of its eigenvalues. Let $f : I \to \mathbb{R}$ where $I \subseteq \mathbb{R}$ is an interval. Consider a matrix $\mathbf{A} \in \mathbb{H}_d$ whose eigenvalues are contained in I. We define the matrix $f(\mathbf{A}) \in \mathbb{H}_d$ using an eigenvalue decomposition of $\mathbf{A}$:

$$f(\mathbf{A}) = \mathbf{Q} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} \mathbf{Q}^* \qquad \text{where} \qquad \mathbf{A} = \mathbf{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \mathbf{Q}^* = \sum_{i=1}^{d} \lambda_i \mathbf{Q}_{\bullet i} \mathbf{Q}_{\bullet i}^*. \tag{3.1}$$

The definition of $f(\mathbf{A})$ does not depend on which eigenvalue decomposition we choose. Any matrix function that arises in this fashion is called a **standard matrix function**.

For each $p \geq 1$ the **Schatten p-norm** is defined as $\|\mathbf{B}\|_p := (\operatorname{tr}(|\mathbf{B}|^p))^{1/p}$ for $\mathbf{B} \in \mathbb{M}_d$. In this setting, $|\mathbf{B}| := (\mathbf{B}^*\mathbf{B})^{1/2}$. The **spectral norm** of an Hermitian matrix $\mathbf{A}$ is defined by the relation $\|\mathbf{A}\| := \lambda_{\max}(\mathbf{A}) \vee (-\lambda_{\min}(\mathbf{B}))$. For a general matrix $\mathbf{B}$, the

spectral norm is defined to be the largest singular value: $\|\mathbf{B}\| := \sigma_1(\mathbf{B})$. The Schatten p-norm dominates the spectral norm for all $p \geq 1$.

3.2 Matrix Khintchin Inequality and Applications

In this section we state the matrix Khintchin inequality and matrix moment inequalities as an applications. We provide the proof of auxiliary theorems which are cited without proof in [MJC+14]. They are needed to prove the matrix Khintchin inequality.

Proposition 3.1. (Generalized Klein inequality) *Let $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ be real-valued functions on an interval I of the real line. Suppose*

$$\sum_{k=1}^{n} u_k(a) v_k(b) \geq 0 \qquad \text{for all } a, b \in I. \tag{3.2}$$

Then

$$\overline{\mathrm{tr}} \left(\sum_{k=1}^{n} u_k(\mathbf{A}) v_k(\mathbf{B}) \right) \geq 0 \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}_d(I). \tag{3.3}$$

Proof. *[Pet94, Proposition 3]* Let $\mathbf{A} = \sum_{i=1}^{d} \lambda_i \mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^*$ and $\mathbf{B} = \sum_{j=1}^{d} \mu_j \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^*$ be the orthonormal decompositions of $\mathbf{A}$ and $\mathbf{B}$. Then

$$\overline{\mathrm{tr}} \left(\sum_{k=1}^{n} u_k(\mathbf{A}) v_k(\mathbf{B}) \right) = \sum_{k=1}^{n} \sum_{i,j=1}^{d} \overline{\mathrm{tr}} \left(u_k(\lambda_i) \mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* v_k(\mu_j) \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right) \tag{3.4}$$

$$= \sum_{i,j=1}^{d} \overline{\mathrm{tr}} \left(\mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right) \sum_{k=1}^{n} u_k(\lambda_i) v_k(\mu_j) \geq 0 \tag{3.5}$$

by the hypothesis. To see that $\mathrm{tr}\left(\mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right)$ is non-negative for all $i, j \in \{1, \ldots, d\}$, we apply a well known extension of von Neumann's trace inequality [Ruh70, Lemma 1], namely

$$\mathrm{tr}(\mathbf{PQ}) \geq \sum_{i=1}^{d} p_i q_{d-i+1} \geq 0 \qquad \text{for all } \mathbf{P}, \mathbf{Q} \in \mathbb{H}_d([0, \infty)), \tag{3.6}$$

where the eigenvalues $p_1 \geq \ldots \geq p_d$ and $q_1 \geq \ldots \geq q_d$ are sorted decreasingly. □

Lemma 3.1. (Mean value trace inequality) *Let I be an interval of the real line. Suppose that $g : I \to \mathbb{R}$ is a weakly increasing function and that $h : I \to \mathbb{R}$ is a function whose derivative h' is convex. Then for all matrices $\mathbf{A}, \mathbf{B} \in \mathbb{H}_d(I)$ it holds*

$$\overline{\mathrm{tr}}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (h(\mathbf{A}) - h(\mathbf{B}))] \leq \frac{1}{2} \overline{\mathrm{tr}}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))]. \tag{3.7}$$

When $h^{'}$ is concave, the inequality is reversed. The same result holds for the standard trace.

Proof. *[MJC+ 14, Lemma 3.4]* Fix $a, b \in I$. Since g is weakly increasing, $(g(a) - g(b)) \cdot (a - b) \geq 0$. The fundamental theorem of calculus and the convexity of $h^{'}$ yield the estimate

$$(g(a) - g(b)) \cdot (h(a) - h(b)) = (g(a) - g(b)) \cdot (a - b) \int_0^1 h^{'}(\tau a + (1 - \tau)b) \mathrm{d}\tau \tag{3.8}$$

$$\leq (g(a) - g(b)) \cdot (a - b) \int_0^1 [\tau h^{'}(a) + (1 - \tau) h^{'}(b)] \mathrm{d}\tau \tag{3.9}$$

$$= \frac{1}{2} [(g(a) - g(b)) \cdot (a - b) \cdot (h^{'}(a) + h^{'}(b))]. \tag{3.10}$$

The inequality is reversed, if $h^{'}$ is concave. To apply the Kleins inequality we expand the terms. The RHS is

$$\begin{aligned} &(g(a) - g(b)) \cdot (a - b) \cdot (h^{'}(a) + h^{'}(b)) \\ &\quad = [g(a) \cdot a \cdot h^{'}(a)] + [g(a) \cdot a] \cdot h^{'}(b) - b \cdot [h^{'}(a) \cdot g(a)] - [b \cdot h^{'}(b)] \cdot g(a) \\ &\qquad + [\text{ the same as above with } a \text{ and } b \text{ reversed }](a \rightleftarrows b) \end{aligned} \tag{3.11}$$

Taking the trace yields

$$\begin{aligned} &\mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{A} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] - \mathrm{tr}[\mathbf{B} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B})) \cdot g(\mathbf{A})] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\ &\quad = \mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{A} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] - \mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{B} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\ &\quad = \mathrm{tr}[g(\mathbf{A}) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\ &\quad = \mathrm{tr}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))]. \end{aligned} \tag{3.12}$$

On the LHS we have only products of two factors which commute under the trace operation. Thus we may use the same expression as in the scalar case without further calculations. The result follows immediately from the Klein inequality. □

Theorem 3.1. (Matrix Khintchin) *Suppose that $p = 1$ or $p \geq 3/2$. Consider a finite sequence $(\mathbf{Y}_k)_{k \geq 1}$ of independent, random, Hermitian matrices and a deterministic sequence $(\mathbf{A}_k)_{k \geq 1}$ for which*

$$\mathbf{E}[\mathbf{Y}_k] = 0 \quad \textit{and} \quad \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2 \qquad \textit{almost surely for all } k \geq 1. \tag{3.13}$$

Then

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{p-\frac{1}{2}}\left\|\left(\sum_{k\geq 1}(\mathbf{A}_k^2+\mathbf{E}[\mathbf{Y}_k^2])\right)^{1/2}\right\|_{2p}. \tag{3.14}$$

In particular, when $(\xi_k)_{k\geq 1}$ *is an independent sequence of Rademacher random variables,*

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\xi_k\mathbf{A}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{2p-1}\left\|\left(\sum_{k\geq 1}\mathbf{A}_k^2\right)^{1/2}\right\|_{2p}. \tag{3.15}$$

Proof. [MJC+14, Corollary 7.3] □

Theorem 3.2. *Assume* $n \geq 3$

(i) Suppose that $p \geq 1$, *and fix* $r \geq p \vee 2\log(n)$. *Consider a finite sequence* $(\mathbf{S}_k)_{k\geq 1}$ *of independent, random, positive-semidefinite matrices with dimension* $n \times n$. *Then*

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{S}_k\right\|^p\right]^{1/p} \leq \left[\left\|\sum_{k\geq 1}\mathbf{E}[\mathbf{S}_k]\right\|^{1/2} + 2\sqrt{er}\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/(2p)}\right]^2. \tag{3.16}$$

(ii) Suppose that $p \geq 2$, *and fix* $r \geq p \vee 2\log(n)$. *Consider a finite sequence* $(\mathbf{Y}_k)_{k\geq 1}$ *of independent, symmetric, random, self-adjoint matrices with dimension* $n \times n$. *Then*

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|^p\right]^{1/p} \leq \sqrt{er}\left\|\left(\sum_{k\geq 1}\mathbf{E}[\mathbf{Y}_k^2]\right)^{1/2}\right\| + 2er\mathbf{E}[\max_{k\geq 1}\|\mathbf{Y}_k\|^p]^{1/p}. \tag{3.17}$$

Proof. [CGT12, Theorem A.1] In the proof, they shift to the Schatten-norm in order to apply matrix Khintchin. Then they shift back to the spectral norm, loosing a dimensional factor. This factor can be improved using the intrinsic dimension lemma. □

3.3 Generalzed Inequalities by Hermitian Dilation

For an introduction to Hermitian Dilation see [Tro15, §2.1.16]

Definition 3.1. (Hermitian Dilation) *The Hermitian dilation*

$$\mathfrak{H} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{H}_{d_1 \times d_2}$$

is a map from a general matrix to an Hermitian matrix defined by

$$\mathfrak{H}(B) := \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} \tag{3.18}$$

Properties:

$\|\mathbf{B}\| = \|\mathfrak{H}\mathbf{B}\|$ Linearity confer [Tro15, §7.7.3] for combination with intrinsic dimension argument.

Theorem 3.3. (Matrix Rosenthal-Pinelis) *Let* $\mathbf{A}_1, \ldots, \mathbf{A}_n$ *be independent, random matrices with dimension* $d_1 \times d_2$ *. Introduce the random matrix*

$$\mathbf{S} := \sum_{k=1}^{n} \mathbf{A}_k.$$

Let $v(\mathbf{S})$ *be the matrix variance statistic of the sum:*

$$v(\mathbf{S}) := \left\| \mathbf{E}[\mathbf{S}\mathbf{S}^\top] \right\| \vee \left\| \mathbf{E}[\mathbf{S}^\top \mathbf{S}] \right\| = \left\| \sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k \mathbf{A}_k^\top] \right\| \vee \left\| \sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k^T \mathbf{A}_k] \right\|. \tag{3.19}$$

Then

$$\left(\mathbf{E}\left[\|\mathbf{S}\|^2 \right] \right)^{\frac{1}{2}} \leq \sqrt{2ev(\mathbf{S}) \log(d_1 + d_2)} + 4e \left(\mathbf{E}[\max_{k \leq n} \|\mathbf{A}_k\|^2] \right)^{\frac{1}{2}} \log(d_1 + d_2). \tag{3.20}$$

Remark. Since $\mathbf{E}[\|S\|] \leq \mathbf{E}[\|S\|^2]^{\frac{1}{2}}$ by the Cauchy-Schwarz inequality, Theorem 3.3 also holds with $\mathbf{E}[\|S\|]$ on the left-hand side of (3.20). To obtain a tail bound we can employ the Markov inequality and Theorem 3.3:

$$\begin{aligned} &\mathbf{P}[\, \|S\| \geq t] \\ &\quad \leq \frac{\mathbf{E}[\|S\|]}{t} \leq \frac{1}{t} \left(\sqrt{2ev(\mathbf{S}) \log(d_1 + d_2)} + 4e \left(\mathbf{E}[\max_{k \leq n} \|\mathbf{A}_k\|^2] \right)^{\frac{1}{2}} \log(d_1 + d_2) \right) \quad \text{for } t > 0. \end{aligned} \tag{3.21}$$

It might be possible to improve the log term employing an intrinsic dimension argument. ♢

3.4 Intrinsic Dimension

Definition. *For a positive-semidefinite matrix* $\mathbf{S}$, *the* ***intrinic dimension*** *is the quantity*

$$\operatorname{intdim} \mathbf{A} \;:=\; \operatorname{tr} \mathbf{A} \,/\, \|\mathbf{A}\| \;.$$

Lemma. (Intrinsic dimenision) *Let* $\varphi : [0,\infty) \to \mathbb{R}$ *be a convex function with* $\varphi(0) = 0$. *For any positive-semidefinite matrix* $\mathbf{S}$ *it holds*

$$\operatorname{tr} \varphi(\mathbf{S}) \;\leq\; \operatorname{intdim} \mathbf{S} \;\cdot\; \varphi\|\mathbf{S}\| \;.$$

Proof. *[Tro15, Lemma 7.5.1]* Since φ is convex on any interval $[0, L]$ with $L > 0$, and $\varphi(0) = 0$, it holds

$$\varphi(a) \;\leq\; (1 - a/L) \cdot \varphi(0) \;+\; a/L \cdot \varphi(L) \;=\; a/L \cdot \varphi(L) \qquad \text{for all } a \in [0, L] \,.$$

Since $\mathbf{S}$ is positive-semidefinite, the eigenvalues of $\mathbf{S}$ fall in the interval $[0, L]$, where $L = \|\mathbf{S}\|$. It follows

$$\begin{aligned} \operatorname{tr} \varphi(\mathbf{S}) \;&=\; \sum_{i=1}^{d} \varphi(\lambda_i) \;\leq\; \sum_{i=1}^{d} \lambda_i / \|\mathbf{S}\| \;\cdot\; \varphi(\|\mathbf{S}\|) \\ &=\; \operatorname{tr}(\mathbf{S}) / \|\mathbf{S}\| \;\cdot\; \varphi(\|\mathbf{S}\|) \;=\; \operatorname{intdim} \mathbf{S} \;\cdot\; \varphi\|\mathbf{S}\| \;. \end{aligned}$$

□

The next example applies the preceding lemma to bound the p-Schatten-norm, when $p \geq 2$, by the spectral norm and the intrinsic dimension.

Example. Let $\mathbf{B} \in \mathbb{C}^{m \times n}$ be any rectangular matrix and let $p \geq 2$. Then $\varphi(x) := |x|$ defines a convex function with $\varphi(0) = 0$. The intrinsic dimension lemma yields

$$\|\mathbf{B}\|_p^p \;=\; \operatorname{tr} |\mathbf{B}^* \mathbf{B}|^{p/2} \;\leq\; \operatorname{intdim} \mathbf{B}^* \mathbf{B} \;\cdot\; \|\mathbf{B}^* \mathbf{B}\|^{p/2} \;=\; \operatorname{intdim} \mathbf{B}^* \mathbf{B} \;\cdot\; \|\mathbf{B}\|^p \;.$$

If, additionally, $\mathbf{B}$ is self-adjoint and positive-semidefinite then it holds

$$\operatorname{tr} \mathbf{B}^* \mathbf{B} \;=\; \operatorname{tr} \mathbf{B}^2 \;=\; \sum_{i=1}^{n} \lambda_i^2 \;\leq\; \left(\sum_{i=1}^{n} \lambda_i \right)^2 \;=\; (\operatorname{tr} \mathbf{B})^2 \;,$$

and consequently

$$\|\mathbf{B}\|_p^p \;\leq\; (\operatorname{intdim} \mathbf{B})^2 \;\cdot\; \|\mathbf{B}\|^p \;.$$

♢

Apply to estimate [CGT12, above (A.4)] to get intrinsic dimension version of Rosenthal-Pinelis inequality.

Takeaways The notion of intrinsic dimension is useful when bounding convex functions of a positive-semidefinite matrix by its spectral norm. We saw how to derive bounds on the p -Schatten-norm when $p \geq 2$.

4 Empirical Processes

Classical references are [vdV00] and [vdvW13]. For maximal inequalities see [vdV00, §19] For Functional Delta-Method see [vdV00, §20] For an introduction to empirical processes and outer expectation see the beginning of [vdvW13].

4.1 A Primer on Empirical Processes

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, $(\mathcal{X}, \Sigma)$ a measurable space, and $X_1, \dots, X_n : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{X}, \Sigma)$ a sample of independent and identically-distributed random variables with probability distribution $\mathbf{P}_X$. Throughout this section we consider the **empirical measure** of this sample, that is, the discrete random measure

$$\mathbf{P}_n : \Sigma \to [0,1], \quad C \mapsto \frac{1}{n} \# \{1 \leq i \leq n \colon X_i \in C\} \,. \tag{4.1}$$

A family $\mathcal{F}$ of measurable functions $f : (\mathcal{X}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induces a stochastic process by

$$f \mapsto \mathbf{P}_n f \,, \tag{4.2}$$

where for a measure Q on $(\mathcal{X}, \Sigma)$ we denote $Qf := \int_{\mathcal{X}} f \, Q(dx)$. In this way we define the $\mathcal{F}$-indexed **empirical process** $\mathbb{G}_n$ by

$$f \;\mapsto\; \mathbb{G}_n f \;:=\; \sqrt{n}(\mathbf{P}_n - \mathbf{P})f \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbf{P}f) \,. \tag{4.3}$$

The purpose of this notation is to abstract the behaviour of $\mathbb{G}_n$ ranging over $\mathcal{F}$. Conforming with this integral viewpoint, we define the (random) norm

$$\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |G_n f| \,. \tag{4.4}$$

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

$$\mathbf{E}^*[T] \;:=\; \inf \left\{ \mathbf{E}[U] \mid U \geq T,\; U : (\Omega, \mathcal{A}, \mathbf{P}) \to \left(\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}})\right) \text{measurable and } \mathbf{E}[U] < \infty \right\} . \tag{4.5}$$

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

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

5 Simple yet useful Calculations

Theorem 5.1. (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*

$$\begin{aligned} f(x+\Delta) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\Delta_i + \sum_{\substack{i,j=1 \\ i\neq j}} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i \partial x_j}\Delta_i\Delta_j \\ + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i^2}\Delta_i^2 \end{aligned} \tag{5.1}$$

Corollary 5.1.1. *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*

$$f(a^T(x+\Delta)) - f(a^T x) = f'(a^T x)\,\Delta^T a + \frac{1}{2} f''(a^T(x+\xi\Delta))\,\Delta^T A\ \Delta, \tag{5.2}$$

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 $\xi \in [0,1]$

$$\frac{\partial^2 f(a^T(x+\xi\Delta))}{\partial x_i \partial x_j} = f''(a^T(x+\xi\Delta))\, a_i a_j. \tag{5.3}$$

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

$$\Delta^T A\ \Delta = 2\sum_{\substack{i,j=1 \\ i\neq j}}^{n} a_i a_j \Delta_i \Delta_j + \sum_{i=1}^{n} a_i^2 \Delta_i^2. \tag{5.4}$$

Plugging (5.3) and (5.4) into (5.1) yields (5.2). □

Proposition 5.1. *For all* $x, y \in \mathbb{R}$ *it holds*

$$|x+y| - |x| \geq -|y| \tag{5.5}$$

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

[BT03] Dimitri P. Bertsekas and John N. Tsitsiklis. Parallel and Distributed Computation:Numerical Methods. November 2003.

[CGT12] Richard Y. Chen, Alex Gittens, and Joel A. Tropp. The Masked Sample Covariance Estimator: An Analysis via Matrix Concentration Inequalities, June 2012.

[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.

[Hah98] Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315, March 1998.

[Kle20] Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer International Publishing, Cham, 2020.

[MJC+14] Lester Mackey, Michael I. Jordan, Richard Y. Chen, Brendan Farrell, and Joel A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3), May 2014.

[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.

[Pea09] J. Pearl. *Causality*. Cambridge University Press, 2009.

[Pet94] Dénes Petz. A survey of certain trace inequalities. *Banach Center Publications*, 30(1):287–298, 1994.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Ruh70] Axel Ruhe. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10(3):343–354, September 1970.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, September 2008.

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

[Tro15] Joel A. Tropp. An Introduction to Matrix Concentration Inequalities, January 2015.

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

[vdvW13] Aad van der vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, March 2013.

[Wag82] Clifford H. Wagner. Simpson's Paradox in Real Life. *The American Statistician*, 36(1):46–48, 1982.

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

[zot] Definition of CONFOUND. https://www.merriam-webster.com/dictionary/confound.

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