# Todo list

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

Universität Stuttgart

Universität Stuttgart

Ioan Scheffel

December 27, 2022

# Contents

# 1 Causal Inference

In this chapter we want to give a introduction to causal inference. We particularly highlight the role of propensity score analysis and explain its importance in observational studies.

## 1.1 The Rubin Causal Model

from wiki: The Rubin causal model (RCM), also known as the Neyman–Rubin causal model,[1] is an approach to the statistical analysis of cause and effect based on the framework of potential outcomes, named after Donald Rubin. The name "Rubin causal model" was first coined by Paul W. Holland.[2] The potential outcomes framework was first proposed by Jerzy Neyman in his 1923 Master's thesis,[3] though he discussed it only in the context of completely randomized experiments.[4] Rubin extended it into a general framework for thinking about causation in both observational and experimental studies.[1]

## 1.2 Propensity Score Analysis

from wiki: In the statistical analysis of observational data, propensity score matching (PSM) is a statistical matching technique that attempts to estimate the effect of a treatment, policy, or other intervention by accounting for the covariates that predict receiving the treatment. PSM attempts to reduce the bias due to confounding variables that could be found in an estimate of the treatment effect obtained from simply comparing outcomes among units that received the treatment versus those that did not. Paul R. Rosenbaum and Donald Rubin introduced the technique in 1983.[1]

The possibility of bias arises because a difference in the treatment outcome (such as the average treatment effect) between treated and untreated groups may be caused by a factor that predicts treatment rather than the treatment itself. In randomized experiments, the randomization enables unbiased estimation of treatment effects; for each covariate, randomization implies that treatment-groups will be balanced on average, by the law of large numbers. Unfortunately, for observational studies, the assignment of treatments to research subjects is typically not random. Matching attempts to reduce the treatment assignment bias, and mimic randomization, by creating a sample of units that received the treatment that is comparable on all observed covariates to a sample of units that did not receive the treatment.

For example, one may be interested to know the consequences of smoking. An observational study is required since it is unethical to randomly assign people to the treatment 'smoking.' The treatment effect estimated by simply comparing those who smoked to those who did not smoke would be biased by any factors that predict smoking (e.g.: gender and age). PSM attempts to control for these biases by making the groups receiving treatment and not-treatment comparable with respect to the control variables.

from a paper: Propensity score weighting is one of the techniques used in controlling for selection biases in non- experimental studies. Propensity scores can be used as weights to account for selection assignment differences between treatment and comparison groups. One of the advantages of this approach is that all the individuals in the study can be used for the outcomes evaluation

## 1.3 Weighting beyond the PS

from [WZ19]: Conventionally, the weights are estimated by modeling the propensities of receiving treatment or exhibiting missingness and then inverting the predicted propensities. However, with this approach it can be difficult to properly adjust for or balance the observed covariates. The reason is that this approach only balances covariates in expectation, by the law of large numbers, but in any particular data set it can be difficult to balance covariates, especially if the data set is small or if the covariates are sparse (Zubizarreta et al., 2011). In addition, this approach can result in very unstable estimates when a few observations have very large weights (e.g., Kang and Schafer 2007). To address these problems, a number of methods have been proposed recently. Instead of explicitly modeling the propensities of treatment or missingness, these methods directly balance the covariates. Some of these methods also minimize a measure of dispersion of the weights.

Most of these weighting methods balance covariates exactly rather than approximately. This is a subtle but important difference because approximate balance can trade bias for variance whereas exact balance cannot. Also, exact balance may not admit a solution whereas approximate balance may do so. For a fixed sample size, approximate balance may balance more functions of the covariates than exact balance.

# 2 Balancing Weights

## 2.1 Introduction

We work in the Rubin Causal Model.

We assume a sample of $n$ units which is drawn from a population distribution.

In i.i.d. fashion.

We observe $(\mathbf{X}_i, T_i, Y_i)$, where $\mathbf{X}$ are covariates, $T$ is the indicator if treatment has been received and $Y$ is the observed outcome.

In the Rubin Causal Model we assume that for each unit the potential outcome exist, i.e. $(Y_i^0, Y_i^1)$ where $Y^1$ stands for the potential outcome had the unit received treatment and $Y^0$ for the potential outcome had the unit received **no** treatment.

It is clear that $Y_i = Y_i^{T_i}$ i.e. we can observe only one of the potential outcomes.

Thus there is a connection to missing data problems.

This is the dilemma of causal inference.

On the population level it is possible to estimate both.

Usually the means of the potential outcomes are compared against each other.

In randomized trials this is a valid approach to causal inference.

In observational studies however the treatment assignment is not known and direct comparison can lead to systematically wrong results.

This phenomenon is called **confounding**.

To address the issue of confounding many methods have been proposed.

An intuitive way to think about potential outcomes is to think of a stochastic process $Y(\cdot)$ indexed over $\{0, 1\}$. By observing $Y_i$ we in fact sample from this process at random index $T$, i.e. from $Y(T)$. We have

$$\mathbf{E}[Y(T)] = \mathbf{E}[Y(1)|T = 1]\mathbf{P}[T = 1] + \mathbf{E}[Y(0)|T = 0]\mathbf{P}[T = 0]. \tag{2.1}$$

Suppose we observe $T = 1$. Clearly we have

$$\mathbf{E}[Y(T)|T = 1] = \mathbf{E}[Y(1)|T = 1] \tag{2.2}$$

## 2.2 Double Robustness

### 2.2.1 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, the statement is easy to prove using Bienaymé's formula and Chebyshev's inequality (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*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}[X_i])\right| \leq \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \geq 1-\tau\,. \tag{2.3}$$

What about more refined concentration inequalities?

Deriving learning rates in this way, we make an implicit assumption. We assume that observed outcomes of the treated follow the same distribution as marginal potential outcomes under treatment. In other words, we require

$$Y(1)\,|\,T=1 \;\sim\; \mathbf{P}_{Y(1)}\,. \tag{2.4}$$

In practice, virtually every scenario violates this assumption. Indeed, any external influence on both $T$ and $Y(1)$ can ruin the above assessment simply by imposing

$$\mathbf{E}[Y(T)\,|\,T=1] \;\neq\; \mathbf{E}[Y(1)]\,. \tag{2.5}$$

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*

$$w_i \;=\; \frac{1}{n}\frac{T_i}{\pi(X_i)}\,. \tag{2.6}$$

*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*

$$\mathbf{P}\left[\left|\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \leq \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \geq 1-\tau\,. \tag{2.7}$$

***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.,

$$(Y(0), Y(1)) \perp\!\!\!\perp T \,|\, X\,. \tag{2.8}$$

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

$$\begin{aligned} n\,\mathbf{E}[w(T,X)\,Y(T)] &= \mathbf{E}\,[Y(T)\cdot(T\,/\,\pi(X))] \\ &= \mathbf{E}\,[Y(1)\,/\,\pi(X)\,|\,T=1]\cdot\mathbf{P}[T=1] \\ &= \int_{\mathcal{X}} \mathbf{E}\,[Y(1)\,|\,X=x, T=1]\cdot(\mathbf{P}[T=1]\,/\,\pi(x))\,\mathbf{P}_{X|T}(dx\,|\,1) \\ &= \int_{\mathcal{X}} [Y(1)|X=x]\,\mathbf{P}_X(dx) = \mathbf{E}[Y(1)]. \end{aligned} \tag{2.9}$$

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

$$n^2\mathbf{E}\left[\Big(Y(T)\cdot(T\,/\,\pi(X))\Big)^2\right] = \mathbf{E}\Big[(Y(1))^2\,/\,\pi(X)\Big]\,. \tag{2.10}$$

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

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 $X$ has compact support. In this case, there exists a constant $C_\pi \in (0, 1/2)$ such that

$$C_\pi \;\leq\; \pi(x) \;\leq\; 1 - C_\pi \qquad \text{for all } x \in \mathcal{X}\,. \tag{2.11}$$

Furthermore,

$$\begin{aligned}\sum_{i=1}^n w_i\, T_i\, Y_i \;-\; \mathbf{E}[Y(1)] \\ = \frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi(X_i)}\,(\,Y_i \;-\; \mathbf{E}[Y(1)]\,) \;+\; \sum_{i=1}^n \left(w_i \;-\; \frac{1}{n}\,\frac{T_i}{\pi(X_i)}\right) Y_i\,.\end{aligned} \tag{2.12}$$

Previously we bounded the first term, so let us seek a bound for the second term. To this end, we shall use maximal inequalities from the theory of empirical processes. Consider

$$f_0(T, X, Y) \;:= \tag{2.13}$$

The next assumptions reduce the technical task to a minimum.

**Assumptions.** *(i)*

## 2.3 Error Decompositions

The following decomposition is flexible in $\Phi$. We get different causal estimaands $\mathbf{E}[\Phi(Y(1))]$, e.g. the popolation average of $Y(1)$ for $\Phi(Y) = Y$, i.e. $\mathbf{E}[Y(1)]$, ot the distribution function of $Y(1)$ at $t$ for $\Phi(Y) = \mathbf{1}_{(-\infty,t]}(Y)$, i.e. $\mathbf{P}[Y(1) \leq t]$.

$$\sum_{i=1}^n w_i T_i \Phi(Y_i) - \mathbf{E}[\Phi(Y(1))] = \frac{1}{n}\sum_{i=1}^n S_i + R_0 + R_1 + R_2, \tag{2.14}$$

where

$$S_i := \frac{T_i}{\pi_i}\left(\Phi(Y_i) - \mathbf{E}[\Phi(Y_i(1))|X_i]\right) + \left(\mathbf{E}[\Phi(Y_i(1))|X_i] - \mathbf{E}[\Phi(Y(1))]\right) \quad \text{for } i \in \{1, \dots, n, \}\,,$$

$$R_0 := \sum_{i=1}^n T_i \left(w_i - \frac{1}{n\pi_i}\right)\left(\Phi(Y_i) - \mathbf{E}[\Phi(Y_i(1))|X_i]\right),$$

$$R_1 := \sum_{i=1}^n \left(T_i w_i - \frac{1}{n}\right)\left(\mathbf{E}[\Phi(Y_i(1))|X_i] - B(X_i)^\top \lambda\right),$$

$$R_2 := \sum_{i=1}^n \left(T_i w_i - \frac{1}{n}\right) B(X_i)^\top \lambda \quad \text{for } \lambda \in \mathbb{R}^K.$$

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

$$f_\Phi(T,X,Y) = \frac{T}{\pi(X)}\left(\Phi(Y) - \mathbf{E}[\Phi(Y)|X]\right) + \mathbf{E}[\Phi(Y)|X]. \tag{2.15}$$

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.

## 2.4 Estimating the Population Mean of Potential Outcomes

We want to estimate the populition mean of the outcome under treatment, i.e. $\mathbf{E}\left[Y^1\right]$.

Since $Y_i^1$ is only observed for the treated units, i.e. if $T_i = 1$ we will consider a weighted mean of the observed outcomes as an estimator, i.e. $\hat{Y}_w^1 = \sum_{i=1}^n w_i T_i Y_i$ where we use convex optimization to compute the weights.

We consider the following decomposition

$$\hat{Y}_w^1 - \mathbf{E}[Y^1] = \frac{1}{n}\sum_{i=1}^n S_i + R_0 + R_1 + R_2, \tag{2.16}$$

where

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

We want to prove asyptotic normality

**Theorem 2.1.** *Suppose that conditions hold. Then*

$$\sqrt{n}\left(\hat{Y}_{w^*}^1 - \mathbf{E}[Y^1]\right) \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,\sigma_*^2).$$

To accomplish this we need

**Theorem 2.2.** *If $T_i = 1$ then $w^*(X_i)$ is a consistent estimator of $\frac{1}{n\pi(X_i)}$.*

We study the following problem:

**Problem 2.1.**

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

*subject to the constraints*

$$w_i T_i \geq 0, \qquad i = 1,\dots,n,$$

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

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

We aim to prove that the solution to Problem (**??**) is asymptotical consistent with the propensity score, i.e.

**Theorem 2.3.** *Under some (non-optimal) Assumptions, there exist constants $c_1, c_2 > 0$ and decreasing sequences $(\varepsilon_n^1), (\varepsilon_n^2) \subset (0,1]$ that converge to 0 such that for all $\tau \in (0,1]$ there exists a constant $c_\tau \in [0,\infty)$ only depending on $\tau$ such that for all $n \geq 1$ and $\tau \in (0,1]$ it holds*

$$\begin{aligned} \mathbf{P}\left(\left\|w_i^* - \frac{1}{n\pi(X_i)}\right\|_\infty \leq c_1 c_\tau \varepsilon_n^1\right) \geq 1-\tau, \\ \left\|w_i^* - \frac{1}{n\pi(X_i)}\right\|_{\mathbf{P},2} \leq c_2 \varepsilon_n^2, \end{aligned} \tag{2.17}$$

*where $w^*$ is the solution to Problem (**??**).*

**Assumption 2.1.** *Assume, the following conditions hold:*

**2.1.1.** *The minimizer $\lambda_0 = \arg\min_{\lambda\in\Theta} \mathbf{E}\left[-Tn\rho\left(B(X)^T\lambda\right) + B(X)^T\lambda\right]$ is unique, where $\Theta \subseteq \mathbb{R}^n$ is the parameter space for $\lambda$.*

**2.1.2.** *The parameter space $\Theta \subseteq \mathbb{R}^n$ is compact.*

**2.1.3.** *$\lambda_0 \in int(\Theta)$, where $int(\cdot)$ stands for the interior of a set.*

**2.1.4.** *There exists $\lambda_1^* \in \Theta$ such that $\left\|m^*(\cdot) - B(\cdot)^T \lambda_1^*\right\|_\infty \leq \varphi_{m^*}$, where $m^*(\cdot) := \left(\rho^{'}\right)^{-1}\left(\frac{1}{n\pi(\cdot)}\right)$.*

**2.1.5.** *There exists a constant $\varphi_\pi \in \left(0, \frac{1}{2}\right)$ such that $\pi(x) \in (\varphi_\pi, 1-\varphi_\pi)$ for all $x \in \mathcal{X}$*

**2.1.6.** *There exists $\varphi_{\rho^{''}} > 0$ such that $-\rho^{''} \geq \varphi_{\rho^{''}} > 0$*

**2.1.7.** *There exists $\varphi_{B(x)B(x)^T} > 0$ such that $B(x)B(x)^T \succcurlyeq \varphi_{B(x)B(x)^T} I$*

**2.1.8.** *There exists $\varphi_{\|B\|} > 0$ such that $\sup_{x\in\mathcal{X}} \|B(x)\|_2 \leq \varphi_{\|B\|}$.*

**2.1.9.** *The number of basis functions satisfies $K = o(n)$.*

## Plan of Proof

It is easier to study the dual of Problem (**??**). Thus we employ results from convex analysis [MMN22] to establish

**Proposition 2.1.** *The dual of Problem (**??**) is equivalent to the unconstrained optimization problem*

$$\underset{\lambda\in\mathbb{R}^K}{\text{minimize}} \quad \frac{1}{n}\sum_{j=1}^{n}\left[-T_j n\rho\left(B(X_j)^T\lambda\right) + B(X_j)^T\lambda\right] + |\lambda|^T\delta, \tag{2.18}$$

*where $B(X_j) = (B_k(X_j))_{1\leq k\leq K}$ denotes the $K$ basis functions of the covariates, $\rho(t) := \frac{t}{n} - t(h^{'})^{-1}(t) + h((h^{'})^{-1}(t))$ with $h(x) := f\left(\frac{1}{n} - x\right)$ and $|\lambda| := (|\lambda_k|)_{1\leq k\leq K}$. Moreover, the primal solution $w_j^*$ satisfies*

$$w_j^* = \rho^{'}\left(B(X_j)^T\lambda^\dagger\right) \tag{2.19}$$

*for $j = 1, \ldots, n$, where $\lambda^\dagger$ is the solution to the dual optimization problem.*

The core of the subsequent analysis is based on Assumption 2.1.4, i.e. the existence of an oracle parameter $\lambda_1^*$ in a sieve estimate of the true propensity score (or a transformation). It is then natural to enquire about the convergence of the dual solution $\lambda^\dagger$ to $\lambda_1^*$. Making certain assumptions and employing matrix concentration inequalitys [Tro15] we can establish

**Proposition 2.2.** *Under some (non-optimal) Assumptions, there exists a constant* $c_3 > 0$ *and a decreasing sequence* $(\varepsilon_n^3) \subset (0,1]$ *that converges to 0 such that for all* $\tau \in (0,1]$ *there exists a constant* $\tilde{c_\tau} \in [0,\infty)$ *only depending on* $\tau$ *such that for all* $n \geq 1$ *and* $\tau \in (0,1]$ *it holds*

$$\mathbf{P}\left(\left\|\lambda^\dagger - \lambda_1^*\right\|_2 \leq c^3 \tilde{c_\tau}(\varepsilon_n^3)\right) \geq 1 - \tau. \tag{2.20}$$

It is then straightforward to prove a more general result than Theorem 2.3.

**Theorem 2.4.** *Under some (non-optimal) Assumptions, there exist constants* $c_1, c_2 > 0$ *and decreasing sequences* $(\varepsilon_n^1), (\varepsilon_n^2) \subset (0,1]$ *that converge to 0 such that for all* $\tau \in (0,1]$ *there exists a constant* $c_\tau \in [0,\infty)$ *only depending on* $\tau$ *such that for all* $n \geq 1$ *and* $\tau \in (0,1]$ *it holds*

$$\mathbf{P}\left(\left\|w^*(\cdot) - \frac{1}{n\pi(\cdot)}\right\|_\infty \leq c_1 c_\tau \varepsilon_n^1\right) \geq 1 - \tau,$$
$$\left\|w^*(X) - \frac{1}{n\pi(X)}\right\|_{\mathbf{P},2} \leq c_2 \varepsilon_n^2,$$

*where* $w^*(X)$ *is as in* (2.19) *without the index.*

## Proof of theorem 2.2

***Proof.*** Motivated by Proposition 3.2 we consider

$$G(\lambda) := \frac{1}{n}\sum_{j=1}^{n}\left[-T_j n\rho\left(B(X_j)^T\lambda\right) + B(X_j)^T\lambda\right] + |\lambda|^T\delta. \tag{2.21}$$

Since $\rho \in C^2(\mathbb{R})$ we can employ (2.21), Corollary 6.1.1 and Proposition 6.1 to get

$$\begin{aligned}
&G(\lambda_1^* + \Delta) - G(\lambda_1^*) \\
&\quad \geq \frac{1}{n}\sum_{j=1}^{n}\left[-T_j n \rho^{'}\left(B(X_j)^T\lambda_1^*\right) + 1\right]\Delta^T B(X_j) \\
&\qquad + \frac{1}{2}\sum_{j=1}^{n} -T_j\rho^{''}\left(B(X_j)^T(\lambda_1^* + \xi\Delta)\right)\Delta^T\left(B(X_j)B(X_j)^T\right)\Delta \\
&\qquad - |\Delta|^T\delta \\
&\quad \geq -\left\|\Delta\right\|_2\left(\left\|\frac{1}{n}\sum_{j=1}^{n}\left[-T_j n\rho^{'}\left(B(X_j)^T\lambda_1^*\right) + 1\right]B(X_j)\right\|_2 + \left\|\delta\right\|_2\right) \\
&\qquad + n\left\|\Delta\right\|_2^2 \varphi_{\rho''}\underline{\varphi_{BB^T}} \\
&\quad := -\left\|\Delta\right\|_2\left(I_1 + \left\|\delta\right\|_2\right) + \left\|\Delta\right\|_2^2 I_2.
\end{aligned} \tag{2.22}$$

The second inequality is due to the Cauchy-Schwarz-Inequality and Assumptions 2.1.6 and 2.1.7 . We want to establish probabilistic upper bounds of the factor associated with $-\left\|\Delta\right\|_2$. This will be done with appropriate assumptions on $\left\|\delta\right\|_2$ and a thorough analysis of $I_1$. If we then restrict lower bounds of $I_2$ to appropriately slow convergence to 0, e.g. by assumptions on $\varphi_{\rho''}$ and $\underline{\varphi_{BB^T}}$, we can choose $\left\|\Delta\right\|_2$ large enough, such that (2.22) yields $G(\lambda_1^* + \Delta) - G(\lambda_1^*) > 0$ with arbitrarily large probabitity for $n$ large enough. With Proposition 3.2 it follows then immediately Proposition 2.2 .

## Analysis of $I_1$

We want to use Assumption 2.1.3. Thus we perform the following split:

$$\begin{aligned}
I_1 \leq &\left\|\sum_{j=1}^{n} T_j\left[\rho^{'}\left(B(X_j)^T\lambda_1^*\right) - \frac{1}{n\pi(X_j)}\right]B(X_j)\right\|_2 \\
&+ \left\|\frac{1}{n}\sum_{j=1}^{n}\left[\frac{T_j}{\pi(X_j)} - 1\right]B(X_j)\right\|_2 \\
&=: J_1 + J_2
\end{aligned} \tag{2.23}$$

### Analysis of $J_1$

By the Lipschitz-continuity of $\rho^{'}$, Assumption 2.1.8 and Assumption 2.1.4, $T \in \{0,1\}$ and the triangle inequality we have

$$J_1 \leq nL_{\rho'}\varphi_{\|B(x)\|}\varphi_{m^*} \tag{2.24}$$

### Analysis of $J_2$

We want to employ Theorem **??**. To this end we define the independent random matrices

$$\begin{aligned} A_j &:= \frac{1}{n}\left[\frac{T_j}{\pi(X_j)} - 1\right] B(X_j)\,, \quad j = 1, \ldots, n, \\ S &:= \sum_{j=1}^{n} A_j \end{aligned} \tag{2.25}$$

and check conditions (**??**) and (4.31). Note that $\|S\|_2 = J_2$ . By the properties of conditional expectation it holds

$$\mathbf{E}\left[\frac{T_j}{\pi(X_j)}B(X_j)\right] = \mathbf{E}\left[\mathbf{E}\left[T_j \mid X_j\right]\frac{1}{\pi(X_j)}B(X_j)\right] = \mathbf{E}[B(X_j)]. \tag{2.26}$$

Taking the expectation in (2.25) and using (2.26) we get $\mathbf{E}[A_j] = 0$ for all $j = 1, \ldots, n$ . Since

$$\left|\frac{T_j}{\pi(X_j)} - 1\right| \leq 1 + \frac{1-\varphi_\pi}{\varphi_\pi} = \frac{1}{\varphi_\pi} \tag{2.27}$$

by Assumption 2.1.5, we can employ Assumption 2.1.8 together with (2.27) and (2.25) to get

$$\|A_j\|_2 \leq \frac{\varphi_{\|B\|}}{n\varphi_\pi} =: L. \tag{2.28}$$

Thus, condition (**??**) is satisfied. Next we turn to the matrix variance statistic $v(S)$ (4.31). By (2.25) and (2.27) we have

$$\mathbf{E}\left[A_j A_j^T\right] \leq \left(\frac{1}{n\varphi_\pi}\right)^2 \mathbf{E}\left[B(X)B(X)^T\right] \tag{2.29}$$

and by (2.28)

$$\mathbf{E}\left[A_j^T A_j\right] \leq L^2. \tag{2.30}$$

Since $\max\{a, b\} \leq |a| + |b|$ we can use (2.29) and (2.30) to get

$$v(S) \leq \frac{1}{n} \frac{\lambda_{\max}}{\varphi_\pi^2} + nL^2, \tag{2.31}$$

where $\lambda_{\max}$ is the maximal eigenvalue of the symmetric (non-random) matrix $\mathbf{E}\left[B(X)B(X)^T\right]$. Having dealt with (**??**) and (4.31) we can establish the expectation bound (**??**) of Theorem **??**. Together with (2.28) and (2.31) we get

$$\begin{aligned} &\mathbf{E}[J_2] \\ &\leq \sqrt{\frac{2\log(K+1)\left(\lambda_{\max} + \varphi_{\|B\|}^2\right)}{n\varphi_\pi^2}} + \frac{\log(K+1)\varphi_{\|B\|}}{3n\varphi_\pi} \\ &\leq \frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(\sqrt{2} + \frac{1}{3}\sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{2\lambda_{\max}}\right]. \end{aligned} \tag{2.32}$$

Since $K = o(n)$ by Assumption 2.1.9 we can discuss the other influences on the quality of the bound (2.32). On a high-level it is readily clear that appropriate bounds on $\varphi_\pi, \varphi_{\|B\|}$ and $\lambda_{\max}$ will shrink $\mathbf{E}[J_2]$ to 0 and will assist in establishing learning rates.

We could also have invoked the probabitity bound (**??**) of Theorem **??**. But for the sake of simplicity we prefer the combination of the expectation bound (2.32) and the Markov inequality. With the latter we get

$$J_2 \leq \frac{1}{\tau}\frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n}}\left[\varphi_{\|B\|}\left(\sqrt{2} + \frac{1}{3}\sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{2\lambda_{\max}}\right] \tag{2.33}$$

with probabitity $\geq 1 - \tau$.

If we choose $\|\Delta\|_2$ to be

$$\begin{aligned} &\Bigg(\sqrt{2}\frac{1}{\tau}\frac{1}{\varphi_\pi}\sqrt{\frac{\log(K+1)}{n^3}}\left[\varphi_{\|B\|}\left(1 + \sqrt{\frac{\log(K+1)}{n}}\right) + \sqrt{\lambda_{\max}}\right] \\ &\quad + L_{\rho'}\varphi_{\|B\|}\varphi_{m^*} + \frac{\|\delta\|_2}{n}\Bigg)\frac{1}{\varphi_{\rho''}\underline{\varphi_{BB^T}}} \end{aligned} \tag{2.34}$$

we get by (2.22), (2.23), (2.24), (2.33) and Proposition 3.2

$$\begin{aligned} \mathbf{P}\left(\left\|\lambda^\dagger - \lambda_1^*\right\|_2 \leq C\right) &= \mathbf{P}\left(\inf_{\|\Delta\|_2 = C} G(\lambda_1^* + \Delta) - G(\lambda_1^*) > 0\right) \\ &\geq 1 - \tau, \end{aligned} \tag{2.35}$$

where $C$ is as in (2.34). With appropriate Assumptions (as discussed before) we can then establish Proposition 2.2.

We can invoke (2.35) to derive bounds as in Theorem 2.17:

$$\begin{aligned}\left\|w^*(X) - \frac{1}{n\pi(X)}\right\|_{\mathbf{P},2} &\leq L_{\rho'}\left[\left\|B(X)^T\left(\lambda^\dagger - \lambda_1^*\right)\right\|_{\mathbf{P},2}\right. \\ &\qquad \left. + \left\|m^*(X) - B(X)^T\lambda_1^*\right\|_{\mathbf{P},2}\right] \\ &\leq L_{\rho'}\left(\varphi_{\|B\|}\sqrt{C^2(1-\tau) + \mathrm{diam}(\Theta)^2\tau} + \varphi_{m^*}\right)\end{aligned}$$

$$\begin{aligned}\left\|w^*(\cdot) - \frac{1}{n\pi(\cdot)}\right\|_{\infty} &\leq L_{\rho'}\left[\left\|B(\cdot)^T\left(\lambda^\dagger - \lambda_1^*\right)\right\|_{\infty}\right. \\ &\qquad \left. + \left\|m^*(\cdot) - B(\cdot)^T\lambda_1^*\right\|_{\infty}\right] \\ &\leq L_{\rho'}\left(\varphi_{\|B\|}C + \varphi_{m^*}\right)\end{aligned}$$

with probabitity greater than $1-\tau$. □

**Remark 2.1.** *By Corollary* **??** *we can get rid of the* $\log(K)$ *term in* (2.34). ♢

**Remark 2.2.** *By the matrix Rosenthal-Pinelis Inequality [CGT12][Thm.A.1] we can weaken Assumption 2.1.5 to a lower bound on the expectition of* $\pi(X)$ ♢

The next step consists of strenghtening the Assumptions to get concrete learning rates. This can be done in a series of examples.

## 2.5 Application of Convex Optimization

**Assumption 2.2.** *Assume that the map* $f : \mathbb{R} \to \overline{\mathbb{R}}$ *has the following properties.*

*(i)* *$f$ is strictly convex.*

*(ii)* *$f$ is lower-semicontinuous and continuously differentiable on* $\mathrm{int}(\mathrm{dom}(f))$.

*(iii)* *The derivative of $f$ on* $\mathrm{int}(\mathrm{dom}(f))$ *is a diffeomorphism.*

*(iv)* *The Legendre transformation $f^*$ of $f$ is finite.*

*(v)* *The function $x \mapsto xt - f(x)$ takes its supremum on* $\mathrm{int}(\mathrm{dom}(f))$ *for all* $t \in \mathbb{R}$.

We consider the following optimization problem.

**Problem 2.2.**

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

*subject to the constraints*

$$\begin{aligned} w_i T_i \geq 0, \qquad i = 1, \ldots, n, \\ \sum_{i=1}^{n} w_i T_i = 1 \\ \left| \sum_{i=1}^{n} w_i T_i B_k(X_i) - \frac{1}{n} \sum_{i=1}^{n} B_k(X_i) \right| \leq \delta_k, \qquad k = 1, \ldots, K \end{aligned}$$

**Theorem 2.5.** *Under Assumption, the dual of the above Problem is the unconstrained optimization problem*

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

*where* $t \mapsto f^*(t) = t\,(f^{'})^{-1}(t) - f\left((f^{'})^{-1}(t)\right)$ *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*

$$w_i^* = (f^{'})^{-1}(\langle B(X_i), \lambda^\dagger\rangle), \quad i \in \{1, \ldots, n\} \tag{2.36}$$

*are the unique optimal solutions to (P) .*

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

**Lemma 2.1.** *The dual of the optimization problem is*

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

*subject to*

$$\lambda_k \geq 0 \quad \textit{for all } k \in \{1, \ldots, K\}, \tag{2.37}$$

*where*

$$\mathbf{Q} := \begin{bmatrix} \mathbf{I}_n \\ \mathbf{B}(\mathbf{X}) \\ -\mathbf{B}(\mathbf{X}) \end{bmatrix}, \qquad \mathbf{B}(\mathbf{X}) := \Big[B(X_1), \ldots, B(X_n)\Big], \qquad \textit{and} \qquad d := \begin{bmatrix} 0_n \\ \delta \\ \delta \end{bmatrix}. \tag{2.38}$$

***Proof.*** *Lemma* We write the optimization problem in the form of [TB91].

$$\underset{w_1,\dots,w_n\in\mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^{n} T_i f(w_i) \tag{2.39}$$

$$w_i T_i \geq 0, \qquad i = 1,\dots,n, \tag{2.40}$$

$$\sum_{i=1}^{n} w_i T_i = 1 \tag{2.41}$$

$$\sum_{i=1}^{n} w_i T_i B_k(X_i) \geq -\delta_k + \frac{1}{n}\sum_{i=1}^{n} B_k(X_i) \quad k = 1,\dots,K \tag{2.42}$$

$$-\sum_{i=1}^{n} w_i T_i B_k(X_i) \geq -\delta_k - \frac{1}{n}\sum_{i=1}^{n} B_k(X_i) \quad k = 1,\dots,K \tag{2.43}$$

Next we write

$$\mathbf{Q} := \begin{bmatrix} \mathbf{diag}[T_1,\dots,T_n] \\ \mathbf{B}(\mathbf{X}) \\ -\mathbf{B}(\mathbf{X}) \end{bmatrix}, \qquad \mathbf{B}(\mathbf{X}) := \Big[T_1 B(X_1),\dots,T_n B(X_n)\Big], \qquad \text{and} \qquad d := \begin{bmatrix} 0_n \\ -\delta + \overline{B(\mathbf{X})} \\ -\delta - \overline{B(\mathbf{X})} \end{bmatrix}, \tag{2.44}$$

where $\overline{B(\mathbf{X})} := [\overline{B_1(\mathbf{X})},\dots,\overline{B_K(\mathbf{X})}]^\top$ and $\overline{B_k(\mathbf{X})} := \frac{1}{n}\sum_{i=1}^{n} B_k(X_i)$. We get

$$\underset{w_1,\dots,w_n\in\mathbb{R}}{\text{minimize}} \quad \sum_{i=1}^{n} T_i f(w_i) \tag{2.45}$$

$$1_n \cdot w = 0 \tag{2.46}$$

$$\mathbf{Q}w \geq d \tag{2.47}$$

The convex conjugate is

$$\sum_{T_i=1} T_i f^*(\lambda_i) + \sum_{T_i=0} \delta_{\{0\}}(\lambda_i). \tag{2.48}$$

Since $\mathbf{Q}_{\bullet i} = 0_n$ if $T_i = 0$ we get the desired representation. Note that the equality constraint vanishes automatically. □

□

## 2.6 Application of Matrix Concentration Inequalities

**Analysis of $\mathbf{E}[\max_{i\le r} \|\mathbf{A}_i\|^2]$**

We have

$$\mathbf{A}_i := \frac{1}{r}\left(\frac{1-\pi_i}{\pi_i}\right)\mathbf{B}(X_i) \qquad \text{for } i \in \{1,\ldots,r\}. \tag{2.49}$$

Since we take the maximum over a finite set it is attained for some $i^* \in \{1,\ldots,r\}$:

$$\begin{aligned}\mathbf{E}[\max_{i\le r}\|\mathbf{A}_i\|^2] &= \mathbf{E}[\|\mathbf{A}_{i^*}\|^2] \\ &= \frac{1}{r^2}\mathbf{E}\left[\left(\frac{1-\pi_{i^*}}{\pi_{i^*}}\right)^2\|\mathbf{B}(X_{i^*})\|^2\right] \le \frac{1}{r^2}\mathbf{E}\left[\left(\frac{1-\pi_{i^*}}{\pi_{i^*}}\right)^4\right]^{\frac{1}{2}}\mathbf{E}[\|\mathbf{B}(X_{i^*})\|^4]^{\frac{1}{2}} \\ &\le \frac{K}{r^2}\sqrt{C_\pi C_{\mathbf{B}}}\end{aligned} \tag{2.50}$$

In the last two steps we applied the Cauchy-Schwarz inequality and Assumption. Note that

$$\sum_{i=1}^{r}\mathbf{E}[\|\mathbf{A}_i\|^2] \le \frac{K}{r}\sqrt{C_\pi C_{\mathbf{B}}} \tag{2.51}$$

**Assumption 2.3.** *There exists $C_\pi \ge 1$ such that $\mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right] \le C_\pi$ for all $i \in \{1,\ldots,r\}$ .*

**Remark 2.3.** *If we assume a logistic regression model for the propensity score it holds for some $\theta \in \mathbb{R}^N$ ($N$ is the number of covariates)*

$$\frac{1-\pi(X)}{\pi(X)} = \exp(-\theta X) \qquad \text{and} \qquad \mathbf{E}\left[\left(\frac{1-\pi(X)}{\pi(X)}\right)^4\right] = \mathbf{E}[\exp(-4\theta X)] = M_X(-4\theta), \tag{2.52}$$

*where $M_X$ is the momement-generating function of $X$. While the first quantity in (2.52) may be unbounded when $X$ has unbounded support, the latter quantity in (2.52) is still bounded for reasonable choices of $X$.* ♢

**Assumption 2.4.** *There exists $C_{\mathbf{B}} \ge 1$ such that $\mathbf{E}[\mathbf{B}_k(X_i)^4] \le C_{\mathbf{B}}$ for all $(k,i) \in \{1,\ldots,K\} \times \{1,\ldots,r\}$ .*

**Remark 2.4.** *With Assumption we also get a bound on the fourth moment of $\|\mathbf{B}(X_i)\|$. Indeed, by the convexity of $x \mapsto x^2$, the monotonicity and linearity of the expectation it holds*

$$\begin{aligned}\mathbf{E}[\|\mathbf{B}(X_i)\|^4] = \mathbf{E}\left[\left(\sum_{k=1}^K \mathbf{B}_k^2(X_i)\right)^2\right] = K^2\mathbf{E}\left[\left(\sum_{k=1}^K \frac{1}{K}\mathbf{B}_k^2(X_i)\right)^2\right] \le K^2\mathbf{E}\left[\sum_{k=1}^K \frac{1}{K}\mathbf{B}_k^4(X_i)\right] \\ = K\sum_{k=1}^K \mathbf{E}\left[\mathbf{B}_k^4(X_i)\right] \le K^2 C_{\mathbf{B}}\end{aligned} \tag{2.53}$$

$\diamondsuit$

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

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

$$\sum_{i=1}^r \mathbf{E}[\mathbf{A}_i\mathbf{A}_i^\top] = \frac{1}{r^2}\sum_{i=1}^r \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 \mathbf{B}(X_i)\mathbf{B}(X_i)^\top\right] = \frac{1}{r^2}\left(\sum_{i=1}^r \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B_k(X_i)B_l(X_i)\right]\right)_{1\le k,l\le K}. \tag{2.54}$$

Thus

$$\begin{aligned}&\left\|\sum_{i=1}^r \mathbf{E}[\mathbf{A}_i\mathbf{A}_i^\top]\right\|_2^2 \\ &\le \left\|\sum_{i=1}^r \mathbf{E}[\mathbf{A}_i\mathbf{A}_i^\top]\right\|_F^2 = \frac{1}{r^4}\sum_{k,l=1}^K\left(\sum_{i=1}^r \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2 B_k(X_i)B_l(X_i)\right]\right)^2 \\ &\le \frac{1}{r^4}\sum_{k,l=1}^K\left(\sum_{i=1}^r \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right]^{\frac{1}{2}} \mathbf{E}[B_k(X_i)^4]^{\frac{1}{4}}\mathbf{E}[B_l(X_i)^4]^{\frac{1}{4}}\right)^2 \le \left(\frac{K}{r}\right)^2 C_\pi C_B\end{aligned} \tag{2.55}$$

On the other hand

$$\begin{aligned}\left\|\sum_{i=1}^r \mathbf{E}[\mathbf{A}_i^\top\mathbf{A}_i]\right\|_2 &= \sum_{i=1}^r \mathbf{E}[\mathbf{A}_i^\top\mathbf{A}_i] = \frac{1}{r^2}\sum_{i=1}^r \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^2\|\mathbf{B}(X_i)\|_2^2\right] \\ &\le \frac{1}{r^2}\sum_{i=1}^r \mathbf{E}\left[\left(\frac{1-\pi_i}{\pi_i}\right)^4\right]^{\frac{1}{2}} \mathbf{E}[\|\mathbf{B}(X_i)\|_2^4]^{\frac{1}{2}} \le \frac{K}{r}\sqrt{C_\pi C_B}\end{aligned} \tag{2.56}$$

It follows

$$v(\mathbf{S}) \leq \frac{K}{r}\sqrt{C_\pi C_B} \tag{2.57}$$

Thus we can apply Theorem 4.4 to get

$$\mathbf{E}[\|\mathbf{S}\|_2] \leq \sqrt{2e\frac{K}{r}\sqrt{C_\pi C_B}\log(K+1)} + 4e\frac{\sqrt{K}}{r}\sqrt[4]{C_\pi C_B}\log(K+1) \leq 14C_\pi C_B\sqrt{\frac{K\log(K+1)}{r}} \tag{2.58}$$

# 3 Convex Analysis

In our application we want to analyze 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, which we will prove in full detail employing a theorem on convex separation is finite dimensions. To this end, we will have a closer look in relative interiors and support functions. As an application we may prove the conjugate chain and sum rule, which are vital to application of duality. As a simple corollary we will obtain the classical Fenche-Rockafellar Duality theorem which gives general conditions for dual und primal optimal values to be equal.

The material we present is very well known, so we claim no originality. We orient our exposition closly by [Roc70, MMN22] .

We finish the chapter with an exposition of [TB91], where for strictly convex functions we get a dual relationship in terms of the optimal solutions.

## 3.1 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 $B := \left\{x = [x_1, \dots, x_n]^\top \in \mathbb{R}^n \colon (\sum_{i=1}^n x_i^2)^{1/2} \leq 1\right\}$

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

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** $\mathrm{aff}(\Omega)$ of a set $\Omega \subseteq \mathbb{R}^n$ is the smallest affine set that includes $\Omega$. 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 3.1.** *Let $\Omega \subseteq \mathbb{R}^n$. We define the **relative interior** of $\Omega$ by*

$$\mathrm{ri}(\Omega) := \{x \in \Omega\colon \textit{there exists } \varepsilon > 0 \textit{ such that } (x + \varepsilon B) \cap \mathrm{aff}(\Omega) \subset \Omega\}. \tag{3.1}$$

Next we collect some useful properties of relative interiors.

**Proposition 3.1.** *Let $C$ be a non-empty convex set in $\mathbb{R}^n$. Then we get the representation*

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

*(ii)* $\mathrm{ri}(C) \neq \emptyset$ *if* $C \neq \emptyset$.

*(iii)* $\mathrm{cl}(C)$ *and* $\mathrm{ri}(C)$ *are convex sets.*

*(iv)* $\mathrm{cl}(\mathrm{ri}(C)) = \mathrm{cl}(C)$ *and* $\mathrm{ri}(\mathrm{cl}(C)) = \mathrm{ri}(C)$.

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

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

*(vii)* $\mathrm{ri}(C_1 \times C_2) = \mathrm{ri}(C_1) \times \mathrm{ri}(C_2)$.

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

Order results to give pretty proof.

***Proof.*** (i) [Roc70, Theorem 6.4]

(ii) [Roc70, Theorem 6.2]

(iii) [Roc70, Theorem 6.2]

(iv) [Roc70, Theorem 6.3]

(v) [Roc70, Theorem 6.5]

(vi) [Roc70, Theorem 6.6-6.7]

(vii) 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

$$z_i + t(z_i - x_i) \in C_i \qquad \text{for } i \in \{1, 2\}\,. \tag{3.2}$$

This proves $\subseteq$ . Suppose $z_1 \in \mathrm{ri}(C_1)$ and $z_2 \in \mathrm{ri}(C_2)$. Let $(x_1, x_2) \in C_1 \times C_2$ with If $t_1 = t_2$ everything is clear. W.l.o.g. assume $t_1 < t_2$. Define $\theta := \frac{t_1}{t_2} \in (0, 1)$. By the convexity of $C_2$ it follows

$$z_2 + t_1(z_2 - x_2) = \theta(z_2 + t_2(z_2 - x_2)) + (1 - \theta)z_2 \in C_2. \tag{3.3}$$

Thus $(z_1, z_2) \in \mathrm{ri}(C_1 \times C_2)$. This proves $\supseteq$ and equality.

(viii) [MMN22, Theorem 2.92]

□

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

**Definition 3.2.** *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 3.1.** *Let $C_1$ and $C_2$ be two non-empty convex sets in $\mathbb{R}^n$. There exists a hyperplane separating $C_1$ and $C_2$ properly if and only if there exists a vector $b \in \mathbb{R}^n$ such that*

$$\sup_{x \in C_2} \langle x, b\rangle \le \inf_{x \in C_1} \langle x, b\rangle \quad \text{and} \quad \inf_{x \in C_2} \langle x, b\rangle < \sup_{x \in C_1} \langle x, b\rangle. \tag{3.4}$$

***Proof.*** [Roc70, Theorem 11.1] □

**Theorem 3.2.** (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 3.3.** *Given a nonempty subset $\Omega \subseteq \mathbb{R}^n$ the **support function** $\sigma_\Omega : \mathbb{R}^n \to \overline{\mathbb{R}}$ of $\Omega$ is defined by*

$$\sigma_\Omega(x^*) := \sup_{x \in \Omega} \langle x^*, x \rangle \qquad \textit{for } x^* \in \mathbb{R}^n. \tag{3.5}$$

**Definition 3.4.** *Given functions $f_i : \mathbb{R}^n \to (-\infty, \infty]$ for $i = 1, \ldots, n$ the **infimal convolution** of these functions is defined as*

$$(f_1 \Box \ldots \Box f_m)(x) := \inf_{\substack{x_i \in \mathbb{R}^n \\ \sum_{i=1}^m x_i = x}} \sum_{i=1}^m f_i(x_i) \tag{3.6}$$

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.

**Theorem 3.3.** *Let $C_1$ and $C_2$ be two non-empty convex sets in $\mathbb{R}^n$ with* $\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) \neq \emptyset$*. Then the support function of the intersection $C_1 \cap C_2$ is represented as*

$$(\sigma_{C_1 \cap C_2})(x^*) = (\sigma_{C_1} \Box \sigma_{C_2})(x^*) \qquad \textit{for all } x^* \in \mathbb{R}^n. \tag{3.7}$$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = \sigma_{C_1}(x_1^*) + \sigma_{C_2}(x_2^*). \tag{3.8}$$

***Proof.*** *[MMN22, Theorem 4.23]*

Read begining of proof in reference (p.266).

We want to use results on convex separation. To make the geometric property of convex separation fruitful to our purpose we consider two special sets. We will verify that these sets meet the requirements for convex separation, i.e., that they are convex and the intresection of their relative interiors is empty. Then a simple calculation will yield $\geq$. To this end, consider the sets

$$\Theta_1 \;:=\; C_1 \times [\,0,\infty) \quad \text{and} \quad \Theta_2 \;:=\; \{(x,\lambda) \in \mathbb{R}^n \;:\; x \in C_2 \text{ and } \lambda \;\leq\; \langle x^*, x\rangle - \alpha\} \;. \tag{3.9}$$

Simplify proof with properties of relative interiors.

Clearly, $\Theta_1$ is convex by the convexity of $C_1$. To see that $\Theta_2$ is convex consider the affine function $\varphi : \mathbb{R}^n\times\mathbb{R} \to \mathbb{R}$, $(x,\lambda) \mapsto \alpha - \langle x^*, x\rangle - \lambda$ . From the definitions of $\varphi$ and $\Theta_2$ we get the identity

$$\Theta_2 \;=\; (C_2\times\mathbb{R}) \;\cap\; \varphi^{-1}(-\infty, 0\,]\,.$$

Thus, by the convexity of the sets $C_2$ and $\varphi^{-1}(-\infty, 0]$ it follows the convexity of $\Theta_2$. Next we show that the relative interiors of $\Theta_1$ and $\Theta_2$ do not intersect, i.e., $\operatorname{ri}\Theta_1 \cap \operatorname{ri}\Theta_2 = \emptyset$. First note that

$$\operatorname{ri}(\Theta_1) = \operatorname{ri}(C_1) \times \operatorname{ri}([0,\infty)) \subseteq \operatorname{ri}(C_1) \times (0,\infty). \tag{3.10}$$

Indeed, if $0 \in \operatorname{ri}([0,\infty))$ then there exists $t > 0$ such that $-tx \geq 0$ for some $x > 0$. A contradiction. Furthermore

$$\operatorname{ri}(\Theta_2) \subseteq \{(x,\lambda) \in \mathbb{R}^n \colon x \in \operatorname{ri}(C_2) \text{ and } \lambda < \langle x^*, x\rangle - \alpha\}\,. \tag{3.11}$$

To see this, assume there is $(x,\lambda) \in \operatorname{ri}(\Theta_2)$ with $\lambda = \langle x^*, x\rangle - \alpha$. Then for some $(y,\mu) \in \Theta_2$ with $\mu < \langle x^*, y\rangle - \alpha$ there exists $t > 0$ such that $(x,\lambda) + t((x,\lambda) - (y,\mu)) \in \Theta_2$. It follows

$$0 \leq (1+t)(\langle x^*, x\rangle - \alpha - \lambda) + t(\mu - \langle x^*, y\rangle + \alpha) < 0, \tag{3.12}$$

a contradiction. The first inequality is due to $(x,\lambda) + t((x,\lambda) - (y,\mu)) \in \Theta_2$ and the second inequality due to $\mu < \langle x^*, y\rangle - \alpha$ and $\lambda = \langle x^*, x\rangle - \alpha$. But then $\operatorname{ri}(\Theta_1) \cap \operatorname{ri}(\Theta_2) = \emptyset$. Indeed, suppose that there exists $(x,\lambda) \in \operatorname{ri}(\Theta_1) \cap \operatorname{ri}(\Theta_2)$. Then it holds $\langle x^*, x\rangle - \alpha \leq 0$ and $\lambda > 0$ since $x \in \operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) \subseteq C_1 \cap C_2$. On the other hand

$$0 < \lambda < \langle x^*, x\rangle - \alpha \leq 0, \tag{3.13}$$

a contradiction.

Finish proof.

$\square$

**Takeaways** This primer is somewhat confusing.
Add more meaning.

## 3.2 Conjugate Calculus and Fenchel-Rockafellar Theorem

The goal of this section is to establish the tools to calculate convex conjugates. We prove the conjugate sum and chain rule. After some examples, we will derive the Fenchel-Rockafellar Theorem.

**Definition 3.5.** (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*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \tag{3.14}$$

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

Note that $f$ in Definition 3.5 does not have to be convex. On the other hand, the convex conjugate is always convex:

**Proposition 3.2.** *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]

Give proof Mordukhovich2022 p.256

□

**Lemma 3.1.** *For any proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ we have*

$$f^*(x^*) = \sigma_{\mathrm{epi}(f)}(x^*, -1) \qquad \text{for } x^* \in \mathbb{R}^n. \tag{3.15}$$

***Proof.*** Let $x^* \in \mathbb{R}^n$ and $(x, \lambda) \in \mathrm{epi}(f)$. Then $x \in \mathrm{dom}(f)$ and $f(x) \leq \lambda$. Thus

$$\langle x^*, x\rangle - f(x) \geq \langle x^*, x\rangle - \lambda \qquad \text{for all } (x, \lambda) \in \mathrm{epi}(f). \tag{3.16}$$

On the other hand $(x, f(x)) \in \mathrm{epi}(f)$ for all $x \in \mathrm{dom}(f)$. It follows

$$\langle x^*, x\rangle - f(x) \leq \sup_{(x,\lambda)\in \mathrm{epi}(f)} \langle x^*, x\rangle - \lambda \qquad \text{for all } x \in \mathrm{dom}(f). \tag{3.17}$$

Taking the supremum in the last two displays yields

$$f^*(x^*) = \sup_{x \in \mathrm{dom}(f)} \langle x^*, x\rangle - f(x) = \sup_{(x,\lambda)\in \mathrm{epi}(f)} \langle x^*, x\rangle - \lambda \tag{3.18}$$

$$= \sup_{(x,\lambda)\in \mathrm{epi}(f)} \langle (x^*, -1), (x, \lambda)\rangle = \sigma_{\mathrm{epi}(f)}(x^*, -1). \tag{3.19}$$

□

**Theorem 3.4.** (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*

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{3.20}$$

*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]

Provide proof. Mordukhovich2022 p.270

□

**Theorem 3.5.** *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*

$$(f + g)^*(x^*) = (f^* \Box g^*)(x^*) \tag{3.21}$$

*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*

$$(f + g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{3.22}$$

***Proof.*** Let $x^* \in \mathbb{R}^n$ and fix $x_1^*, x_2^* \in \mathbb{R}^n$ such that $x^* = x_1^* + x_2^*$. We get

$$\begin{aligned} f^*(x_1^*) + g^*(x_2^*) &= \sup_{x\in\mathbb{R}^n} \langle x_1^*, x\rangle - f(x) + \sup_{x\in\mathbb{R}^n} \langle x_2^*, x\rangle - g(x) \\ &\geq \sup_{x\in\mathbb{R}^n} \langle x_1^*, x\rangle - f(x) + \langle x_2^*, x\rangle - g(x) = \sup_{x\in\mathbb{R}^n} \langle x_1^* + x_2^*, x\rangle - (f(x) + g(x)) \\ &= \sup_{x\in\mathbb{R}^n} \langle x^*, x\rangle - (f + g)(x) = (f + g)^*(x^*) \end{aligned}$$

Taking the infimum over $x_1^*, x_2^* \in \mathbb{R}^n$ in the above display gives $(f^* \Box g^*)(x^*) \geq (f + g)^*(x^*)$. Let us prove now $\leq$ under the condition $\mathrm{ri}\,(\mathrm{dom}(f)) \cap \mathrm{ri}\,(\mathrm{dom}(g)) \neq$

$\emptyset$. The only case we need to consider is $(f+g)^*(x^*) < \infty$. Define two convex sets by

$$\Omega_1 := \left\{(x,\alpha,\beta) \in \mathbb{R}^{n+2} \colon \alpha \geq f(x)\right\} = \operatorname{epi}(f) \times \mathbb{R}, \tag{3.23}$$
$$\Omega_2 := \left\{(x,\alpha,\beta) \in \mathbb{R}^{n+2} \colon \beta \geq g(x)\right\}. \tag{3.24}$$

Similar to Lemma we get the representation

$$(f+g)^*(x^*) = \sigma_{\Omega_1 \cap \Omega_2}(x^*, -1, -1). \tag{3.25}$$

Indeed, the only thing we need to verify is $\operatorname{dom}(f) \cap \operatorname{dom}(g) = \operatorname{dom}(f+g)$. The inclusion $\subseteq$ is clear. Assume towards a contradiction that $(f+g)(x) < \infty$ and $f(x) = \infty$. Since $g(x) > -\infty$ it holds

$$\infty = \infty + g(x) = f(x) + g(x) = (f+g)(x) < \infty. \tag{3.26}$$

This is a contradiction. The same holds for $f$ and $g$ reversed. It follows the inclusion $\supseteq$ and equality. By the support function intersection rule there exist triples

$$(x_1^*, -\alpha_1, -\beta_1), (x_2^*, -\alpha_2, -\beta_2) \in \mathbb{R}^{n+2} \quad \text{such that} \quad (x^*, -1, -1) = (x_1^* + x_2^*, -(\alpha_1 + \alpha_2), -(\beta_1 + \beta_2 \tag{3.27}$$

and

$$(f+g)^*(x^*) = \sigma_{\Omega_1 \cap \Omega_2}(x^*, -1, -1) = \sigma_{\Omega_1}(x_1^*, -\alpha_1, -\beta_1) + \sigma_{\Omega_2}(x_2^*, -\alpha_2, -\beta_2). \tag{3.28}$$

Next we show $\beta_1 = \alpha_2 = 0$. Suppose towards a contradiction that $\beta_1 \neq 0$. We fix $(\overline{x}, \overline{\alpha}) \in \operatorname{epi}(f)$. Then

$$\sigma_{\Omega_1}(x_1^*, -\alpha_1, -\beta_1) = \sup_{(x,\alpha,\beta) \in \operatorname{epi}(f) \times \mathbb{R}} \langle x^*, x\rangle - \alpha\alpha_1 - \beta\beta_1 \geq \sup_{\beta \in \mathbb{R}} \langle x^*, \overline{x}\rangle - \overline{\alpha}\alpha_1 - \beta\beta_1 = \infty. \tag{3.29}$$

This contradicts $(f+g)^*(x^*) < \infty$. In a similar fashion we can derive a contradiction for $\alpha_2 \neq 0$. Employing Lemma and taking into account the structures of the sets $\Omega_1$ and $\Omega_2$ this implies

$$(f+g)^*(x^*) = \sigma_{\Omega_1 \cap \Omega_2}(x^*, -1, -1) = \sigma_{\Omega_1}(x_1^*, -1, 0) + \sigma_{\Omega_2}(x_2^*, 0, -1) \tag{3.30}$$
$$= \sigma_{\operatorname{epi}(f)}(x_1^*, -1) + \sigma_{\operatorname{epi}(g)}(x_2^*, -1) = f^*(x_1^*) + g^*(x_2^*) \geq (f^* \square g^*)(x^*). \tag{3.31}$$

This finishes the proof. $\square$

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

Write example on convex conjugates of $F(w) = \sum_{i=1}^{n} f(w_i)$ . See notes.

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 3.1.** (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 3.1.** *Problem 3.1 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 $\Omega$ is a convex set, we can choose*

$$g(x) = \delta_{\Omega}(x) := \tag{3.32}$$

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

**Problem 3.2.** (Dual) *Consider the same setting as in Problem 3.1. Using the convex conjugates of $f, g$ and the transpose of $A$ we define the* ***dual problem*** *of Problem 3.1 to be*

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

**Proposition 3.3.** *Consider the optimization problem 3.1 and its dual 3.2, where the functions $f$ and $g$ are not assumed to be convex. Define the* ***optimal values*** *of these problems by*

$$\widehat{p} := \inf_{x \in \mathbb{R}^n} f(x) + g(Ax) \quad \textit{and} \quad \widehat{d} := \sup_{y \in \mathbb{R}^m} -f^*(A^\top y) - g^*(y).$$

*Then we have the relationship $\widehat{d} \leq \widehat{p}$.*

***Proof.*** It holds

$$
\begin{aligned}
-f^*(A^\top y^*) - g^*(y^*) &= -\sup_{x\in\mathbb{R}^n} \langle A^\top y^*, x\rangle - f(x) - \sup_{y\in\mathbb{R}^m} \langle -y^*, y\rangle - g(y) \\
&= \inf_{x\in\mathbb{R}^n} f(x) - \langle y^*, Ax\rangle + \inf_{y\in\mathbb{R}^m} g(y) + \langle y^*, y\rangle \\
&\le \inf_{x\in\mathbb{R}^n} f(x) - \langle y^*, Ax\rangle + \inf_{x\in\mathbb{R}^n} g(Ax) + \langle y^*, Ax\rangle \\
&\le \inf_{x\in\mathbb{R}^n} f(x) - \langle y^*, Ax\rangle + g(Ax) + \langle y^*, Ax\rangle \\
&= \inf_{x\in\mathbb{R}^n} f(x) + g(Ax) = \widehat{p}
\end{aligned}
$$

The first equality is due to the definition of convex conjugates, the second equality due to $\langle A^\top y, x\rangle = \langle y, Ax\rangle$ and $\inf\{-B\} = -\sup\{B\}$ for all $B \subseteq \overline{\mathbb{R}}$ and the first inequality due to $\mathrm{Im}(A) \subseteq \mathbb{R}^m$. Taking the supremum with respect to all $y^* \in \mathbb{R}^m$ yields the result. □

Provide proof (Mordukhovich2022 p.293)

**Theorem 3.6.** *Let $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper convex functions and $0 \in ri(dom(g) - A(dom(f)))$ . Then the optimal values of (3.1) and (3.2) are equal, i.e.*

$$
\inf_{x\in\mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y\in\mathbb{R}^n} \left\{-f^*\left(A^T y\right) - g^*(-y)\right\}. \tag{3.33}
$$

***Proof.*** [MMN22, Theorem 4.63] □

Insert lemma in chapter 1.

**Lemma 3.2.** *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be convex. Then for all $y \in \mathbb{R}^n$ and $C > 0$*

$$
\inf_{\|\Delta\|=C} f(y+\Delta) - f(y) \ge 0 \quad \Longrightarrow \quad \exists y^* \in \mathbb{R}^n : y^* \textit{ is global minimum of } f \textit{ and } \|y^* - y\| \le C. \tag{3.34}
$$

***Proof.*** Since $\mathcal{C} := \{\|\Delta\| \le C\}$ is convex $f$ has a local minimum in $y + \mathcal{C} := \{y + \Delta \mid \|\Delta\| \le C\}$. Suppose towards a contradiction that $y^* \in y + \mathcal{C}$ is a local minimum, but not a global minimum and the left-hand side of (3.34) is true. Then it holds

$$
f(x) < f(y^*) \quad \text{for some } x \in \mathbb{R}^n \setminus y + \mathcal{C}. \tag{3.35}
$$

Furthermore since $y + \mathcal{C}$ is compact and contains $y^*$, the line segment $\mathcal{L}[y^*, x]$ contains a point on the boundary of $y + \mathcal{C}$, i.e.

$$
\theta x + (1-\theta) y^* = y + \Delta_x \quad \text{for some } \theta \in (0,1) \text{ and } \Delta_x \text{ with } \|\Delta_x\| = C. \tag{3.36}
$$

It follows

$$\begin{aligned} f(y^*) \leq f(y) \leq f(y+\Delta_x) = f(\theta x + (1-\theta)y^*) \\ \leq \theta f(x) + (1-\theta) f(y^*) < f(y^*), \end{aligned} \tag{3.37}$$

which is a contradiction. Thus every local minimum of $f$ in $y + \mathcal{C}$ is also a global minimum. The first inequality is due to $y^*$ being a local minimum of $f$ in $y + \mathcal{C}$, the second inequality is due to the left-hand side of (3.34) being true, the equality is due to (3.36), the third inequality is due to the convexity of $f$ and the strict inequality is due to (3.35). □

**Takeaways** Almost there

Add more meaning.

## 3.3 Tseng Bertsekas

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

Consider the following optimization problem

$$\underset{x \in \mathbb{R}^m}{\text{minimize}} \qquad f(x)$$

subject to the constraints

$$\mathbf{A}x \geq b, \tag{3.38}$$

Where $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, $\mathbf{A}$ is a given $n \times m$ matrix, and $b$ is a vector in $\mathbb{R}^n$.

Generalize also to take equality constraints. Write in unconstrained form to derive dual.

**Assumption 3.1.** *Assume that the map $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ has the following properties.*

*(i) $f$ is strictly convex.*

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

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

The dual optimization problem associated with $(P)$ is

$$\underset{p \in \mathbb{R}^n}{\text{maximize}} \qquad q(p)$$

subject to the constraints

$$p \geq 0, \tag{3.39}$$

where $q : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the concave function given by

$$q(p) := \min_{x \in \mathbb{R}^m} f(x) + \langle p, b - \mathbf{A}x \rangle = \langle p, b \rangle - f^*(\mathbf{A}^\top p). \tag{3.40}$$

The dual problem $(D)$ is a concave program with simple nonnegativity constraints. Furthermore, strong duality holds for $(P)$ and $(D)$ , i.e., the optimal value of $(P)$ equals the optimal value of $(D)$.

Since $f^*$ is real-valued and $f$ is strictly convex, $f^*$ and $q$ are continuously differentiable.

**Theorem 3.7.** [Roc70, Theorem 26.3] *A closed proper convex function is (essentially) strictly convex if and only if its conjugate is essentially smooth.*

Read and understand proof (p.270)

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

$$d(p) = b - \mathbf{A}x, \quad \text{where} \quad x := \nabla f^*(\mathbf{A}^\top p) = \operatorname{argsup}_{\xi \in \mathbb{R}^m} \langle p, \mathbf{A}\xi \rangle - f(\xi). \tag{3.41}$$

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

Read and understand proof (p.80)

**Proposition 3.4.** (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*

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad x \mapsto \sup_{z \in Z} \phi(x, z) \tag{3.42}$$

*is convex and has directional derivative given by*

$$f'(x;y) = \sup_{z\in Z(x)} \phi'(x,z;y), \tag{3.43}$$

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

$$Z(x) := \left\{ \overline{z} \in \mathbb{R}^m \colon \phi(x,\overline{z}) = \sup_{z\in Z} \phi(x,z) \right\}. \tag{3.44}$$

*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

$$\mathbf{A}p \in \partial f(x). \tag{3.45}$$

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

$$p = [p + d(p)]^+, \tag{3.46}$$

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.

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

$$\mathbf{A}x \geq b \quad \text{and} \quad p_i = 0 \quad \text{for all } i \text{ such that} \quad \sum_{j=1}^m a_{ij}x_j > b_i. \tag{3.47}$$

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

n

**Definition 3.6.** [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*

$$f_1(x) \leq 0, \ldots, f_r(x) \leq 0, \qquad f_{r+1}(x) = 0, \ldots, f_m(x) = 0, \tag{3.48}$$

*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 3.7.** *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 3.8.** (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* $\alpha_i$ *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** For strictly convex functions we can derive duality in terms of the optimal solutions.

# 4 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. We base our exposition on [MJC+14]. A lot will be exact copy of this paper, so no originality is claimed. Where it seemed fit, we conducted some calculations in more detail than presented in the paper.

We will first introduce the method of exchangable pairs and derive auxiliary theorems to establish the matrix Khintchin inequality. Then we will derive inequalities for moments of matrices, first for psd matrices and then via the Hermitian dilataition for general rectangular matrices. In a last step we will introduce the notion of intrinsic dimension to improve the bounds.

## 4.1 A Matrix Analysis Primer

The **trace** of a square matrix, denoted by tr, is the sum of its diagonal entries, i.e. $\operatorname{tr}(\mathbf{B}) = \sum_{j=1}^{d} b_{jj} \quad \text{for } \mathbf{B} \in \mathbb{M}_d$. The trace is unitarily invariant, i.e. $\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\mathbf{QBQ}^*) \quad \text{for all } \mathbf{B} \in \mathbb{M}_d$ 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 $\mathbf{A}$ :

$$f(\mathbf{A}) = \mathbf{Q} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} \mathbf{Q}^* \qquad \text{where} \qquad \mathbf{A} = \mathbf{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \mathbf{Q}^* = \sum_{i=1}^{d} \lambda_i \mathbf{Q}_{\bullet i} \mathbf{Q}_{\bullet i}^*. \tag{4.1}$$

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 $p \geq 1$.

**Proposition 4.1.** *Let $f, g : I \to \mathbb{R}$ be real-valued functions on an interval $I \subseteq \mathbb{R}$, and let $\mathbf{A} \in \mathbb{H}_d$ be a Hermitian matrix whose eigenvalues are contained in $I$.*

*(i) If $\lambda$ is an eigenvalue of of $\mathbf{A}$, then $f(\lambda)$ is an eigenvalue of $f(\mathbf{A})$.*

*(ii) $f(a) \leq g(a) \quad$ for all $a \in I \quad$ implies $\quad f(\mathbf{A}) \preccurlyeq g(\mathbf{A})$.*

**Takeaways** This Primer is not a prim number. Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

## 4.2 The Method of Exchangeable Pairs

We first define an exchangable pair.

**Definition 4.1.** *Let $Z$ and $Z'$ random variables taking values in a Polish space $\mathcal{Z}$. We say that $(Z, Z')$ is an **exchangable pair** if it has the same distribution as $(Z', Z)$. In particular, $Z$ and $Z'$ must share the same distribution.*

The following approach originates in the work of Charles Stein [Ste72] on normal approximation for a sum of dependent random variable. We will explain how some central ideas of this theory extends to matrices.

We can obtain a lot of information about the fluctuation of a random matrix $\mathbf{X}$ if we can construct a good exchangable pair $(\mathbf{X}, \mathbf{X}')$. With this motivation in mind, let us introduce a special class of exchangable pairs.

**Definition 4.2.** *Let $(Z, Z^{'})$ be an exchangable pair of random variables taking values in a Polish space $\mathcal{Z}$, and let $\mathbf{\Psi} : \mathcal{Z} \to \mathbb{H}_d$ be a measurable function. Define the random Hermitian matrices*

$$\mathbf{X} := \mathbf{\Psi}(Z) \quad \textit{and} \quad \mathbf{X}^{'} := \mathbf{\Psi}(Z^{'}). \tag{4.2}$$

*We say that $(\mathbf{X}, \mathbf{X}^{'})$ is a* ***matrix Stein pair*** *if there is a constant $\alpha \in (0, 1]$ for which*

$$\mathbf{E}[\mathbf{X} - \mathbf{X}^{'}|Z] = \alpha\mathbf{X} \qquad \textit{almost surely.} \tag{4.3}$$

*The constant $\alpha$ is called the* ***scale factor*** *of the pair. We always assume $\mathbf{E}\left[\|\mathbf{X}\|^2\right] < \infty$.*

A matrix Stein pair $(\mathbf{X}, \mathbf{X}^{'})$ has several useful propreties. First, $(\mathbf{X}, \mathbf{X}^{'})$ always forms an exchangable pair. Second, it must be the case that $\mathbf{E}[\mathbf{X}] = \mathbf{0}$. Indeed,

$$\mathbf{E}[\mathbf{X}] = \frac{1}{\alpha}\mathbf{E}[\mathbf{E}[\mathbf{X} - \mathbf{X}^{'}|Z]] = \frac{1}{\alpha}\mathbf{E}[\mathbf{X} - \mathbf{X}^{'}|] = \mathbf{0}.$$

A well-chosen matrix Stein pair $(\mathbf{X}, \mathbf{X}^{'})$ provides a surprisingly powerful tool for studying the random matrix $\mathbf{X}$. The technique depends on a fundamental technical lemma.

**Lemma 4.1.** *Suppose that $(\mathbf{X}, \mathbf{X}^{'})$ is a matrix Stein pair with scale factor $\alpha$. Let $\mathbf{F} : \mathbb{H}_d \to \mathbb{H}_d$ be a measurable function that satisfies the regularity condition $\mathbf{E}\left[\left\|(\mathbf{X} - \mathbf{X}^{'})\mathbf{F}(\mathbf{X})\right\|\right] < \infty$. Then*

$$\mathbf{E}[\mathbf{X} \cdot \mathbf{F}(\mathbf{X})] = \frac{1}{2\alpha}\mathbf{E}[(\mathbf{X} - \mathbf{X}^{'})(\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{X}^{'}))]. \tag{4.4}$$

In short, the randomness in the Stein pair furnishes an alternative expression for the expected product of $\mathbf{X}$ and a function $\mathbf{F}$. It allows us to estimate the expectation using the smoothness properties of the function $\mathbf{F}$ and the discrepancy between $\mathbf{X}$ and $\mathbf{X}^{'}$.

***Proof.*** *[MJC+14, Lemma 2.4]* Suppose that $(\mathbf{X}, \mathbf{X}^{'})$ constructed from an auxiliary exchangable pair $(Z, Z^{'})$. The defining property implies

$$\alpha \cdot \mathbf{E}[\mathbf{X} \cdot \mathbf{F}(\mathbf{X})] = \mathbf{E}[\mathbf{E}[\mathbf{X} - \mathbf{X}^{'}|Z] \cdot \mathbf{F}(\mathbf{X})] = \mathbf{E}[(\mathbf{X} - \mathbf{X}^{'})\mathbf{F}(\mathbf{X})] \tag{4.5}$$

$\square$

To each matrix Stein pair $(\mathbf{X}, \mathbf{X}^{'})$, we may associate a random matrix called the conditional variance of $\mathbf{X}$. The purpose of this section is to argue that the spectral norm of $\mathbf{X}$ is unlikely to be large, when the conditional variance is small.

**Definition 4.3.** *Suppose that* $(\mathbf{X}, \mathbf{X}^{'})$, *is a matrix Stein pair, constructed from an auxiliary exchangeable pair* $(Z, Z^{'})$. *The* ***conditional variance*** *is the random matrix*

$$\mathbf{\Delta_X} := \mathbf{\Delta_X}(Z) := \frac{1}{2\alpha}\mathbf{E}[(\mathbf{X} - \mathbf{X}^{'})^2 | Z], \tag{4.6}$$

*where* $\alpha$ *is the scale factor of the pair. We may take any version of the conditional expectation in this definition.*

The conditional variance $\mathbf{\Delta_X}$ can be regarded as a stochastic estimate for the variance of the random matrix $\mathbf{X}$. To see this, assume the second moment of $\mathbf{X}$ exists. Then it follows from Lemma with $\mathbf{F}(\mathbf{X}) = \mathbf{X}$

$$\mathbf{E}[\mathbf{\Delta_X}] = \mathbf{E}[\mathbf{X}^2]. \tag{4.7}$$

To verify the regularity condition, note that

$$\mathbf{E}[\|(\mathbf{X} - \mathbf{X}^{'})\mathbf{X}\|] \leq \mathbf{E}[\|\mathbf{X}\|^2] + \mathbf{E}[\|\mathbf{X}\| \cdot \|\mathbf{X}^{'}\|] \leq 2\mathbf{E}[\|\mathbf{X}\|^2] < \infty. \tag{4.8}$$

*Example* 4.1. *[MJC+ 14, Example 2.4]* ♢

**Takeaways** The conditional variance is cool. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 4.3 Matrix Khintchin Inequality and Applications

The goal of this section is to derive the matrix Khintchin inequality and show some important applications. For this we need an auxiliary theorem which is an extension of the *Burkholder-Davis-Gundy (BDG) inequality* from classical martingale theory [Bur73]. We prepare for the proof of this theorem by assembling some analytic tools.

**Proposition 4.2.** (Generalized Klein inequality) *Let $u_1, \dots, u_n$ and $v_1, \dots, v_n$ be real-valued functions on an interval $I$ of the real line. Suppose*

$$\sum_{k=1}^{n} u_k(a) v_k(b) \geq 0 \qquad \text{for all } a, b \in I. \tag{4.9}$$

*Then*

$$\overline{\operatorname{tr}} \left( \sum_{k=1}^{n} u_k(\mathbf{A}) v_k(\mathbf{B}) \right) \geq 0 \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}_d(I). \tag{4.10}$$

***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

$$\overline{\operatorname{tr}} \left( \sum_{k=1}^{n} u_k(\mathbf{A}) v_k(\mathbf{B}) \right) = \sum_{k=1}^{n} \sum_{i,j=1}^{d} \overline{\operatorname{tr}} \left( u_k(\lambda_i) \mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* v_k(\mu_j) \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right) \tag{4.11}$$

$$= \sum_{i,j=1}^{d} \overline{\operatorname{tr}} \left( \mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right) \sum_{k=1}^{n} u_k(\lambda_i) v_k(\mu_j) \geq 0 \tag{4.12}$$

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

$$\operatorname{tr}(\mathbf{PQ}) \geq \sum_{i=1}^{d} p_i q_{d-i+1} \geq 0 \qquad \text{for all } \mathbf{P}, \mathbf{Q} \in \mathbb{H}_d([0, \infty)), \tag{4.13}$$

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

**Lemma 4.2.** (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*

$$\overline{\mathrm{tr}}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (h(\mathbf{A}) - h(\mathbf{B}))] \leq \frac{1}{2} \overline{\mathrm{tr}}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))]. \tag{4.14}$$

*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

$$(g(a) - g(b)) \cdot (h(a) - h(b)) = (g(a) - g(b)) \cdot (a - b) \int_0^1 h'(\tau a + (1 - \tau)b) \mathrm{d}\tau \tag{4.15}$$

$$\leq (g(a) - g(b)) \cdot (a - b) \int_0^1 [\tau h'(a) + (1 - \tau)h'(b)] \mathrm{d}\tau \tag{4.16}$$

$$= \frac{1}{2} [(g(a) - g(b)) \cdot (a - b) \cdot (h'(a) + h'(b))]. \tag{4.17}$$

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

$$\begin{aligned}
&(g(a) - g(b)) \cdot (a - b) \cdot (h'(a) + h'(b)) \\
&\quad = [g(a) \cdot a \cdot h'(a)] + [g(a) \cdot a] \cdot h'(b) - b \cdot [h'(a) \cdot g(a)] - [b \cdot h'(b)] \cdot g(a) \\
&\qquad + [\text{ the same as above with } a \text{ and } b \text{ reversed }](a \rightleftarrows b)
\end{aligned} \tag{4.18}$$

Taking the trace yields

$$\begin{aligned}
&\mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{A} \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))] - \mathrm{tr}[\mathbf{B} \cdot (h'(\mathbf{A}) + h'(\mathbf{B})) \cdot g(\mathbf{A})] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\
&\quad = \mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{A} \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))] - \mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{B} \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\
&\quad = \mathrm{tr}[g(\mathbf{A}) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\
&\quad = \mathrm{tr}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h'(\mathbf{A}) + h'(\mathbf{B}))].
\end{aligned} \tag{4.19}$$

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. □

**Proposition 4.3.** (Hölder inequality for trace) *Let $p$ and $q$ be Hölder conjugate indices. Then*

$$\operatorname{tr}(\mathbf{BC}) \leq \|\mathbf{B}\|_p \|\mathbf{C}\|_q \qquad \textit{for all } \mathbf{B}, \mathbf{C} \in \mathbb{M}_d. \tag{4.20}$$

***Proof.*** [Bha97, Corollary IV.2.6] □

We are now ready to prove the auxiliary theorem.

**Theorem 4.1.** (Matrix BDG inequality) *Let $p = 1$ or $p \geq 3/2$. Suppose that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair where $\mathbf{E}[\|\mathbf{X}\|_{2p}^{2p}] < \infty$. Then*

$$\mathbf{E}[\|\mathbf{X}\|_{2p}^{2p}]^{1/(2p)} \leq \sqrt{2p-1}\ \mathbf{E}[\|\mathbf{\Delta_X}\|_p^p]^{1/(2p)}, \tag{4.21}$$

*where $\mathbf{\Delta_X}$ is the conditional variance .*

***Proof.*** *[MJC⁺14, §7.3]* Suppose that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair with scale factor $\alpha$. First, observe that the result for $p = 1$ already follows from $\mathbf{E}[\mathbf{\Delta_X}] = \mathbf{E}[\mathbf{X}^2]$. Therefore we may assume that $p \geq 3/2$. We introduce the notation for the quantity of interest,

$$E := \mathbf{E}[\|\mathbf{X}\|_{2p}^{2p}] = \mathbf{E}[\operatorname{tr}(|\mathbf{X}|^{2p})]. \tag{4.22}$$

We rewrite the expression for $E$ by peeling off a copy of $|\mathbf{X}|$. This yields

$$E = \mathbf{E}[\operatorname{tr}(|\mathbf{X}| \cdot |\mathbf{X}|^{2p-1})] = \mathbf{E}[\operatorname{tr}(\mathbf{X} \cdot \operatorname{sgn}(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1})]. \tag{4.23}$$

Apply the method of exchangable pairs with $\mathbf{F}(\mathbf{X}) = \operatorname{sgn}(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1}$ to reach

$$E = \frac{1}{2\alpha}\mathbf{E}[\operatorname{tr}((\mathbf{X} - \mathbf{X}') \cdot (\operatorname{sgn}(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1} - \operatorname{sgn}(\mathbf{X}') \cdot |\mathbf{X}'|^{2p-1}))] \tag{4.24}$$

Apply method of exchangeable pairs, generalized Klein inequality, trace Hölder □

**Theorem 4.2.** [MJC⁺14, Corollary 7.3] *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*

$$\mathbf{E}[\mathbf{Y}_k] = 0 \quad \textit{and} \quad \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2 \qquad \textit{almost surely for all } k \geq 1. \tag{4.25}$$

*Then*

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{p-\frac{1}{2}}\left\|\left(\sum_{k\geq 1}(\mathbf{A}_k^2+\mathbf{E}[\mathbf{Y}_k^2])\right)^{1/2}\right\|_{2p}. \tag{4.26}$$

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

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\xi_k\mathbf{A}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{2p-1}\left\|\left(\sum_{k\geq 1}\mathbf{A}_k^2\right)^{1/2}\right\|_{2p}. \tag{4.27}$$

**Theorem 4.3.** *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*

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{S}_k\right\|^p\right]^{1/p} \leq \left[\left\|\sum_{k\geq 1}\mathbf{E}[\mathbf{S}_k]\right\|^{1/2} + 2\sqrt{er}\,\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/(2p)}\right]^2. \tag{4.28}$$

*(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*

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|^p\right]^{1/p} \leq \sqrt{er}\left\|\left(\sum_{k\geq 1}\mathbf{E}[\mathbf{Y}_k^2]\right)^{1/2}\right\| + 2er\,\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/p}. \tag{4.29}$$

**Takeaways** This is so amazing Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

## 4.4 Generalzed Inequalities by Hermitian Dilataition

**Definition 4.4.** (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*

$$\mathfrak{H}(B) := \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} \tag{4.30}$$

**Theorem 4.4.** (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:*

$$v(\mathbf{S}) := \left\| \mathbf{E}[\mathbf{S}\mathbf{S}^\top] \right\| \vee \left\| \mathbf{E}[\mathbf{S}^\top \mathbf{S}] \right\| = \left\| \sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k \mathbf{A}_k^\top] \right\| \vee \left\| \sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k^T \mathbf{A}_k] \right\|. \tag{4.31}$$

*Then*

$$\left( \mathbf{E} \left[ \|\mathbf{S}\|^2 \right] \right)^{\frac{1}{2}} \leq \sqrt{2ev(\mathbf{S}) \log(d_1 + d_2)} + 4e \left( \mathbf{E}[\max_{k \leq n} \|\mathbf{A}_k\|^2] \right)^{\frac{1}{2}} \log(d_1 + d_2). \tag{4.32}$$

**Remark 4.1.** *Since* $\mathbf{E}[\|S\|] \le \mathbf{E}[\|S\|^2]^{\frac{1}{2}}$ *by the Cauchy-Schwarz inequality, Theorem 4.4 also holds with* $\mathbf{E}[\|S\|]$ *on the left-hand side of* (4.32). *To obtain a tail bound we can employ the Markov inequality and Theorem 4.4:*

$$\mathbf{P}[\,\|S\| \ge t] \\ \le \frac{\mathbf{E}[\|S\|]}{t} \le \frac{1}{t}\left(\sqrt{2ev(\mathbf{S})\log(d_1+d_2)} + 4e\left(\mathbf{E}[\max_{k\le n}\|\mathbf{A}_k\|^2]\right)^{\frac{1}{2}}\log(d_1+d_2)\right) \quad \text{for } t > 0. \tag{4.33}$$

*It might be possible to improve the* log *term employing an intrinsic dimension argument.* $\diamondsuit$

**Takeaways** Dilataition is so deep. Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

## 4.5 Intrinsic Dimension

**Definition 4.5.** *For a positive-semidefinite matrix* $\mathbf{S}$, *the* ***intrinic dimension*** *is the quantity*

$$\operatorname{intdim}(\mathbf{A}) := \frac{\operatorname{tr}\mathbf{A}}{\|\mathbf{A}\|}.$$

**Lemma 4.3.** (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 that*

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

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

$$\varphi(a) \le \left(1 - \frac{a}{L}\right)\varphi(0) + \frac{a}{L}\varphi(L) = \frac{a}{L}\varphi(L) \qquad \text{for all } a \in [0, L]. \tag{4.34}$$

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

$$\operatorname{tr}(\varphi(\mathbf{S})) = \sum_{i=1}^{d} \varphi(\lambda_i) \leq \frac{\sum_{i=1}^{d} \lambda_i}{\|\mathbf{S}\|} \varphi(\|\mathbf{S}\|) = \frac{\operatorname{tr}(\mathbf{S})}{\|\mathbf{S}\|} \varphi(\|\mathbf{S}\|) = \operatorname{intdim}(\mathbf{S}) \cdot \varphi(\|\mathbf{S}\|). \tag{4.35}$$

□

**Takeaways** Is it intrinsic or extrinsic? Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

# 5 Empirical Processes

## 5.1 A Primer on Empirical Processes

Let $(\mathbb{D}, d)$ be a metric space, and let $(\mathbf{P}_n)_{n\in\mathbb{N}}\mathbf{P}$ be (Borel) probability measures on $(\mathbb{D}, \mathcal{D})$, where $\mathcal{D}$ is the Borel $\sigma$-algebra on $\mathbb{D}$, the smallest $\sigma$-algebra containing all open sets. Then the sequence $\mathbf{P}_n$ **converges weakly** to $\mathbf{P}$, which we denote as $\mathbf{P}_n \rightsquigarrow \mathbf{P}$, if and only if

$$\int_{\mathbb{D}} f \mathrm{d}\mathbf{P}_n \to \int_{\mathbb{D}} f \mathrm{d}\mathbf{P} \qquad \text{for all } f \in C_b(\mathbb{D}). \tag{5.1}$$

Here $C_b(\mathbb{D})$ denotes the set of all bounded, continuous, real functions on $\mathbb{D}$. Equivalently, if $X_n$ and $X$ are $\mathbb{D}$-valued random variables with distribution $\mathbf{P}_n$ and $\mathbf{P}$ respectively, then $X_n \to X$ if and only if

$$\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)] \qquad \text{for all } f \in C_b(\mathbb{D}). \tag{5.2}$$

This definitions yield the classical theory of weak convergence. For a modern treatment see [Kle20].

The classical theory requires that $\mathbf{P}_n$ is defined, for each $n \in \mathbb{N}$, on the Borel $\sigma$-algebra $\mathcal{D}$, or, equivalently, that $X_n$ is a Borel measurable map for each $n \in \mathbb{N}$. If $(\Omega_n, \mathcal{A}_n, \mathbf{P}_n)$ are the underlying probability spaces on which the maps $X_n$ are defined, this means that $X_n^{-1}(D) \in \mathcal{A}_n$ for every Borel set $D \in \mathcal{D}$. This required measurability usually holds when $\mathbb{D}$ is a separable metric space such as $\mathbb{R}^k$ or $C([0,1])$ with the supremum metric.

However, this apparently modest requirement can and does easily fail when the metric space $\mathbb{D}$ is not separable.

*Example* 5.1. *[vdvW13, Problem 1.7.3]* Let $\mathbb{D} = D([0,1])$ be the **Skorohod space** of all right-continuous functions on $[0,1]$ with left limits endowed with the metric induced by the supremum norm. Define $X : [0,1] \to \mathbb{D},\ \omega \mapsto \mathbf{1}_{[\omega,1]}$. If we equip $[0,1]$ with the Borel $\sigma$-algebra $\mathcal{B}([0,1])$, then $X$ is not measurable. To see this, let $B_s$ be the open ball of radius $1/2$ in $\mathbb{D}$ around the function $\mathbf{1}_{[s,1]}$. Now $X(\omega) \in B_s$ if and only if $\omega = s$. Indeed, if $\omega \neq s$ there exists an $x$ between $\omega$ and $s$ such that the difference of the indicator functions is 1 at $x$. Conversely, if the distance is greater than $1/2$ at a point $x \in [0,1]$, it is because $x$ lies between $\omega$ and $s$ and the indicator functions have difference 1. Since arbitrary (even uncountable) unions of open sets are open, we get for every $S \subseteq [0,1]$ the open set $G := \bigcup_{s\in S} B_s \in \mathcal{D}$. It follows $X^{-1}(G) = S$ for all $S \subseteq [0,1]$. Since not all subsets of $[0,1]$ are measurable, we have $X^{-1}(\mathcal{D}) \nsubseteq \mathcal{B}([0,1])$. But then $X$ is not measurable. The $\sigma$-algebra $\mathcal{D}$ is to large. $\diamondsuit$

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $(\mathcal{X}, \Sigma)$ a measurable space. Let $X_j : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{X}, \Sigma), j = 1, \ldots, n$ be independent and identically-distributed (i.i.d.) random variables with probability distribution $\mathbf{P}_X$ and $\mathcal{F}$ a family of measurable functions $f : (\mathcal{X}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Consider the map

$$f \mapsto G_n f := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{P}_X f\right), \tag{5.3}$$

where $\mathbf{P}_X f := \int_{\mathcal{X}} f \mathrm{d}\mathbf{P}_X$. We call $(G_n f)_{f\in\mathcal{F}}$ the empirical process indexed by $\mathcal{F}$. Furthermore

$$\|G_n f\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |G_n f|. \tag{5.4}$$

## 5.2 Maximal Inequalities

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $(\mathcal{X}, \Sigma)$ a measurable space. Let $X_j : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{X}, \Sigma), j = 1, \ldots, n$ be independent and identically-distributed (i.i.d.) random variables with probability distribution $\mathbf{P}_X$ and $\mathcal{F}$ a family of measurable functions $f : (\mathcal{X}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Consider the map

$$f \mapsto G_n f := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbf{P}_X f\right), \tag{5.5}$$

where $\mathbf{P}_X f := \int_{\mathcal{X}} f \mathrm{d}\mathbf{P}_X$. We call $(G_n f)_{f\in\mathcal{F}}$ the empirical process indexed by $\mathcal{F}$. Furthermore

$$\|G_n f\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |G_n f|. \tag{5.6}$$

**Lemma 5.1.** (Bernstein Inequality for Empirical Processes) *For any bounded, measurable function $f$ it holds for all $t > 0$*

$$\mathbf{P}(|G_n f| > t) \leq 2\exp\left(-\frac{1}{4}\frac{t^2}{\mathbf{P}_X(f^2) + t\|f\|_\infty/\sqrt{n}}\right) \tag{5.7}$$

***Proof.*** By the Markov inequality it holds for all $\lambda > 0$

$$\mathbf{P}\left(G_n f > t\right) \leq e^{-\lambda t}\mathbf{E}\exp\left(\lambda G_n f\right) \tag{5.8}$$

□

**Lemma 5.2.** *For any finite class $\mathcal{F}$ of bounded, measurable, square-integrable functions, with $|\mathcal{F}|$ elements, it holds*

$$\mathbf{E}\left\|G_n f\right\|_{\mathcal{F}} \lesssim \max_{f\in\mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}} \log\left(1+|\mathcal{F}|\right) + \max_{f\in\mathcal{F}} \|f\|_{\mathbf{P},2}\sqrt{\log\left(1+|\mathcal{F}|\right)}. \tag{5.9}$$

**Lemma 5.3.** *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbf{P}f^2 < \delta^2$ for every $f$, we have, with $a(\delta) = \delta/\sqrt{\mathrm{Log} N_{[]}(\delta, \mathcal{F}, L_2(\mathbf{P}))}$, and $F$ an envelope function,*

$$\mathbf{E}^*_{\mathbf{P}}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J_{[]}(\delta, \mathcal{F}, L_2(\mathbf{P})) + \sqrt{n}\mathbf{P}^* F\left\{F > \sqrt{n}a(\delta)\right\}. \tag{5.10}$$

**Corollary 5.0.1.** *For any class $\mathcal{F}$ of measurable functions with envelope function $F$,*

$$\mathbf{E}^*_{\mathbf{P}}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J_{[]}(\|F\|_{\mathbf{P},2}, \mathcal{F}, L_2(\mathbf{P})). \tag{5.11}$$

## 5.3 Functional Delta Method

**Definition 5.1.** *A map $\phi : \mathbb{D}_\phi \to \mathbb{E}$, defined on a subset $\mathbb{D}_\phi$ of a normed space $\mathbb{D}$ that contains $\theta$, is called* ***Hadamard diffenertiable*** *at $\theta$ if there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that*

$$\left\|\frac{\phi(\theta + th_t) - \phi(\theta)}{t} - \phi'_\theta(h)\right\|_{\mathbb{E}} \to 0 \quad \textit{as } t \searrow 0 \textit{ for all } h_t \to h \tag{5.12}$$

*such that $\theta + th_t$ is contained in $\mathbb{D}_\phi$ for all small $t > 0$.*

**Theorem 5.1.** (Delta Method) *Let $\mathbb{D}$ and $\mathbb{E}$ be normed linear spaces. Let $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ be Hadamard differentiable it $\theta$ tangentially to $\mathbb{D}_0$. Let $T_n : \Omega_n \to \mathbb{D}_\phi$ be maps such that $r_n(T_n - \theta) \rightsquigarrow T$ for some sequence of numbers $r_n \to \infty$ and a random element $T$ that takes its values in $\mathbb{D}_0$. Then $r_n(\phi(T_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(T)$. If $\phi'_\theta$ is defined and continuous on the whole space $\mathbb{D}$, then we also have $r_n(\phi(T_n) - \phi(\theta)) = \phi'_\theta(r_n(T_n - \theta)) + o_{\mathbf{P}}(1)$.*

***Proof.*** [vdV98, Theorem 20.8] □

# 6 Simple yet useful Calculations

**Theorem 6.1.** (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*

$$\begin{aligned} f(x+\Delta) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\Delta_i + \sum_{\substack{i,j=1\\ i\neq j}} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i \partial x_j}\Delta_i\Delta_j \\ + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 f(x+\xi\Delta)}{\partial x_i^2}\Delta_i^2 \end{aligned} \tag{6.1}$$

**Corollary 6.1.1.** *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*

$$f(a^T(x+\Delta)) - f(a^Tx) = f^{'}(a^Tx)\,\Delta^T a + \frac{1}{2} f^{''}(a^T(x+\xi\Delta))\,\Delta^T A\ \Delta, \tag{6.2}$$

*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 $\xi \in [0,1]$

$$\frac{\partial^2 f(a^T(x+\xi\Delta))}{\partial x_i \partial x_j} = f^{''}(a^T(x+\xi\Delta))\,a_i a_j. \tag{6.3}$$

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

$$\Delta^T A\ \Delta = 2\sum_{\substack{i,j=1\\ i\neq j}}^{n} a_i a_j \Delta_i \Delta_j + \sum_{i=1}^{n} a_i^2 \Delta_i^2. \tag{6.4}$$

Plugging (6.3) and (6.4) into (6.1) yields (6.2). □

**Proposition 6.1.** *For all $x, y \in \mathbb{R}$ it holds*

$$|x+y| - |x| \geq -|y| \tag{6.5}$$

***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 $\mu$

# Bibliography


[Bha97] Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1997.

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

[Bur73] D. L. Burkholder. Distribution Function Inequalities for Martingales. *The Annals of Probability*, 1(1):19–42, 1973.

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

[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.

[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.

[Ste72] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 6.2:583–603, January 1972.

[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.

[vdV98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

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

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