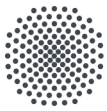
Todo list

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

Universität Stuttgart



Universität Stuttgart

Ioan Scheffel

January 4, 2023

Contents

1 Balancing Weights

1.1 Double Robustness

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

1.1.1 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*

$$\mathbf{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}[X_i])\right| \leq \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \geq 1 - \tau\,. \tag{1.1}$$

Reflection. Bernsteins's inequality yields better confidence. ♠

What about more refined concentration inequalities?cf Steinwart p225f

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

$$Y(1)\,|\,T = 1 \;\sim\; \mathbf{P}_{Y(1)}\,. \tag{1.2}$$

In practice, virtually every scenario violates this assumption. Compare the health of an asthmatic nonsmoker with that of an otherwise healthy smoker.

Find reference to more confounded scenarios.

Indeed, any unbalanced external influence on both T and $Y(1)$ ruins the above assessment, simply by imposing

$$\mathbf{E}[Y(T) \,|\, T = 1] \;\neq\; \mathbf{E}[Y(1)]\,. \tag{1.3}$$

Formulate better transition to propensity score weighting.

Statisticians have wrestled with this issue for nearly a century.

In experimental studies we usually specify treatment assignment as opposed to merely observing a unit receiving treatment.

The next statement makes use of the propensity score.

Theorem. *Consider the weighted mean estimator with weights*

$$w_i \;=\; \frac{1}{n}\frac{T_i}{\pi(X_i)}\,. \tag{1.4}$$

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

$$\mathbf{P}\left[\left|\sum_{i=1}^{n} w_i Y_i - \mathbf{E}[Y(1)]\right| \leq \sqrt{V}\frac{1}{\sqrt{\tau}}\frac{1}{\sqrt{n}}\right] \geq 1 - \tau\,. \tag{1.5}$$

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

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

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

$$(Y(0), Y(1)) \perp\!\!\!\perp T \,|\, X\,. \tag{1.6}$$

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

calculations go as follows.

$$\begin{aligned} n\,\mathbf{E}[w(T,X)\,Y(T)] &= \mathbf{E}\left[Y(T)\cdot(T\,/\,\pi(X))\right] \\ &= \mathbf{E}\left[Y(1)\,/\,\pi(X)\,|\,T=1\right]\cdot\mathbf{P}[T=1] \\ &= \int_{\mathcal{X}} \mathbf{E}\left[Y(1)\,|\,X=x,T=1\right]\cdot(\mathbf{P}[T=1]\,/\,\pi(x))\,\mathbf{P}_{X|T}(dx\,|\,1) \\ &= \int_{\mathcal{X}} [Y(1)|X=x]\,\mathbf{P}_X(dx) = \mathbf{E}[Y(1)]. \end{aligned} \tag{1.7}$$

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

$$\mathbf{E}\left[\Big(Y(T)\cdot(T\,/\,\pi(X))\Big)^2\right] \;=\; \mathbf{E}\Big[(Y(1))^2\,/\,\pi(X)\Big]\,. \tag{1.8}$$

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

Formulate better transition to weights with estimated propensity score.

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

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

Assumptions. *Let the following hold.*

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

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

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

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

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

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

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

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

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

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

$$c_\tau = \gamma_\tau / \tau \tag{1.9}$$

and uniform constant

$$C_{\mathbf{P}} = \max\left\{\sqrt{C_{\mathcal{F}}}C_w, C_Y C_w, \frac{C_Y}{C_\pi}\right\} \tag{1.10}$$

Proof. We consider the following error decomposition.

$$\begin{aligned}
&\sum_{i=1}^{n} w_i\, T_i\, Y_i \;-\; \mathbf{E}[Y(1)] \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)}\,(\,Y_i \;-\; \mathbf{E}[Y(1)]\,) \;+\; \sum_{i=1}^{n}\left(w_i \;-\; \frac{1}{n}\,\frac{1}{\pi(X_i)}\right) T_i\, Y_i\,. \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\pi(X_i)}\,(\,Y_i \;-\; \mathbf{E}[Y(1)]\,) \;+\; \frac{1}{\sqrt{n}}\,\mathbb{G}_n f_w \;-\; \mathbf{E}[\,f_w(T, X, Y)\,]\,.
\end{aligned} \tag{1.11}$$

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

2nd term

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

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

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

$$\mathbf{P}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}} \geq t\right] \;\leq\; \frac{1}{t}\,\mathbf{E}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}}\right] \;\leq\; \frac{1}{t}\,\mathbf{E}^*\left[\|\mathbb{G}_n\|_{\mathcal{F}_{n,\tau}}\right] \leq \frac{\delta_{n,\tau}}{t}\,, \tag{1.12}$$

and consequently

$$\mathbf{P}\left[\|\mathbb{G}_n\|^*_{\mathcal{F}_{n,\tau}} \leq \frac{1}{\tau} C_Y C_w \gamma_\tau r_n\right] \;\geq\; 1 - \tau \tag{1.13}$$

Next, note that

$$\|f_w\|_\infty \leq C_Y \left\| nw - \frac{1}{\pi(X)} \right\| \leq C_Y C_w \gamma_\tau r_n \tag{1.14}$$

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

$$\mathbf{P}\left[\mathbb{G}_n f_w \leq \frac{1}{\tau} C_Y C_w \gamma_\tau r_n\right] \geq 1 - 2\tau\,. \tag{1.15}$$

Streamline analysis of second term.

3rd term

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

$$\mathbf{E}[f_w] \leq C_Y C_w \gamma_\tau r_n (1 - \tau) + C_Y (n + \frac{1}{C_\pi})\tau \tag{1.16}$$

$$\leq C_Y \left(C_w \gamma_\tau r_n + \left(n + \frac{1}{C_\pi} \right) \tau \right)\,. \tag{1.17}$$

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

$$\mathbf{E}[f_w] \leq C_Y \left(C_w + 1 + \frac{1}{C_\pi} \right) \varepsilon_n\,. \tag{1.18}$$

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

Streamline analysis of third term.

□

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

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

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

1.2 Error Decompositions

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

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

$$\sum_{i=1}^{n} w_i T_i \Phi(Y_i) - \mathbf{E}[\Phi(Y(1))] = \frac{1}{n}\sum_{i=1}^{n} S_i + R_0 + R_1 + R_2, \tag{1.19}$$

where

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

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

$$f_\Phi(T, X, Y) = \frac{T}{\pi(X)} \left(\Phi(Y) - \mathbf{E}[\Phi(Y)|X]\right) + \mathbf{E}[\Phi(Y)|X]. \tag{1.20}$$

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

A plethora of applications of the delta method to estimates of the distribution function are to be found in [vdV00] and [vdvW13]. This includes Quantile estimation [vdV00, §21] [vdvW13, §3.9.21/24], survival analysis via Nelson-Aalen and Kaplan-Meier estimator [vdvW13, §3.9.19/31], Wilcoxon Test [vdvW13, §3.9.4.1], and much more. Maybe Boostrapping from the weighted distribution is also sensible .

1.3 Application of Convex Optimization

Assumption 1.1. *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* $\operatorname{int}(\operatorname{dom}(f))$.

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

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

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

Elaborate on assumptions. Give as a counterexample

$$f(x) := (1 + \delta_{x \geq 0})(x^2 + x \cdot \lambda)$$

We consider the following optimization problem.

Problem 1.1.

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

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 1.1. *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*

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

are the unique optimal solutions to (P) .

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

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

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

subject to

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

where

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

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

$$\begin{aligned}
\underset{w_1,\dots,w_n\in\mathbb{R}}{\text{minimize}} \quad & \sum_{i=1}^n T_i f(w_i) \\
w_i T_i \geq 0, \quad & i = 1,\dots,n, \\
\sum_{i=1}^n w_i T_i &= 1 \\
\sum_{i=1}^n w_i T_i B_k(X_i) \geq -\delta_k + \frac{1}{n}\sum_{i=1}^n B_k(X_i) \quad & k = 1,\dots,K \\
-\sum_{i=1}^n w_i T_i B_k(X_i) \geq -\delta_k - \frac{1}{n}\sum_{i=1}^n B_k(X_i) \quad & k = 1,\dots,K
\end{aligned}$$

Next we write

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

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

$$\begin{aligned}
\underset{w_1,\dots,w_n\in\mathbb{R}}{\text{minimize}} \quad & \sum_{i=1}^n T_i f(w_i) \\
1_n \cdot w &= 0 \\
\mathbf{Q}w &\geq d
\end{aligned}$$

The convex conjugate is

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

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

Continue with paraphrase of [WZ19, page 20]

□

1.4 Application of Matrix Concentration Inequalities

We extend the analysis to covariates with unbounded support. To the best of our knowledge this has not been done in this setting. Indeed, [WZ19, Assumption

1] only allows for covariates with bounded support. Our idea is to apply Matrix Moment Inequalities instead of Bernstein's inequality. We learned from this possibility from [Tro15, (6.1.6)]

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

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

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

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

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

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

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

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

$$1\,/\,\pi(X) \;=\; 1 \;+\; \exp(\,\vartheta_0 \,+\, \langle \vartheta, X\rangle\,)\,, \tag{1.21}$$

$$\mathbf{E}\left[\left(\frac{1}{\pi(X)}\right)^4\right] \;=\; \sum_{j=1}^{4} \binom{4}{j} e^{j\vartheta_0}\, M_X(\,j\,\vartheta\,), \tag{1.22}$$

where M_X is the momement-generating function of the random vector X (we assume it exists.). While the quantity in (1.21) may be unbounded when X has unbounded support, that in (1.22) remains bounded for moderate decay rates of the underlying distribution.

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

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

In particular, the expression in (1.22) is finite if X follows a multivariate normal distribution.

Likewise we give a negative example. Let $N = 1$ and $X \sim \mathrm{Exp}(\lambda)$, that is, only one exponentially distributed covariate is taken into account. The momement-generating function is then confined to take arguments $t < \lambda$, and thus (1.22) becomes pointless if $4\,\vartheta \geq \lambda$. ♢

Next, we recall the entity we want to examin.

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

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

$$\left|1 \;-\; \frac{T_i}{\pi_i}\right| \;\leq\; \left(1 \;\vee\; \frac{1-\pi_i}{\pi_i}\right) \;\leq\; 1 \;+\; \frac{1-\pi_i}{\pi_i} \;=\; \frac{1}{\pi_i}\,. \tag{1.23}$$

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

$$\begin{aligned}
\mathbf{E}\left[\max_{i\leq n}\|A_i\|^2\right] \;&=\; \mathbf{E}\left[\|A_{i^*}\|^2\right] \;\leq\; \mathbf{E}\left[\left(\frac{\|B(X_{i^*})\|}{\pi_{i^*}}\right)^2\right] \;/\; n^2 \\
&\leq\; \mathbf{E}\left[\left(\frac{1}{\pi_{i^*}}\right)^4\right]^{1/2} \cdot\; \mathbf{E}\left[\|B(X_{i^*})\|^4\right]^{1/2} \;/\; n^2 \\
&\leq\; K\,/\,n^2 \;\cdot\; \sqrt{C_\pi C_B}\,.
\end{aligned} \tag{1.24}$$

The first inequality comes from the bound (1.23). The Cauchy-Schwarz inequality provides the second inequality. In the last step we use the assumptions made at the start of the section. Paying the price of an extra n factor, the maximal inequality (1.24) yields a bound of the sum, that is,

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

Assumption 1.2. .

Assumption 1.3. .

Remark 1.1. *With Assumption we also get a bound on the fourth moment of $\|B(X_i)\|$. Indeed, by the convexity of $x \mapsto x^2$, the monotonicity and linearity of the expectation it holds*

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

♢

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

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

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

Thus

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

On the other hand

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

It follows

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

Thus we can apply Theorem 3.4 to get

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

2 Convex Analysis

In our application we want to analyse a convex optimization problem by its dual problem. In particular we want to obtain primal optimal solutions from dual solutions. To accomplish the task we need technical tools from convex analysis, mainly conjugate calculus and some KKT related results.

Our starting point is the support function intersection rule [MMN22, Theorem 4.23]. We give the details in the case of finite dimensions and refer for the rest of the proof to the book. The conjugate sum rule is applied to give first conjugate sum and then chain rule, which are vital to calculating convex conjugates. The proofs are omited, since the book is thorough enough. The well known Fenche-Rockafellar Duality theorem is a corollary of conjugate sum and chain rule. It gives general conditions under which dual and primal values coincide. The material we present is very well known, so we claim no originality. We paraphrase the approach of [MMN22] to Duality. As an introduction, we recommend this recently published book together with the classical reference [Roc70].

We finish the chapter with ideas from [TB91]. They provide the high-level ideas to obtain for strictly convex functions a dual relationship in terms of the optimal solutions. We will deliver the details that are omited in the paper.

2.1 A Convex Analysis Primer

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

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

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

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

Solve editorial issue with ball.

the **Euclidian unit ball** in $\mathbb{R}^n$. This is a closed convex set. For any $a \in \mathbb{R}^n$, the **ball with radius** $\varepsilon > 0$ **and center** a is given by $\left\{a + x \in \mathbb{R}^n \colon (\sum_{i=1}^n x_i^2)^{1/2} \leq \varepsilon\right\} = a+\varepsilon B$. For any set Ω in $\mathbb{R}^n$, the set of points x whose distance from Ω does not exceed ε is $\Omega+\varepsilon B$. The **closure** $\mathrm{cl}(\Omega)$ and **interior** $\mathrm{int}(\Omega)$ of Ω 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 Ω. A mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ is called **affine mapping** if there exist a linear mapping $L : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $b \in \mathbb{R}^m$ such that $A(x) = L(x) + b$ for all $x \in \mathbb{R}^n$. The image and inverse image/preimage of convex sets under affine mappings are also convex.

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

Definition 2.1. *Let $\Omega \subseteq \mathbb{R}^n$. We define the **relative interior** of Ω by*

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

Next we collect some useful properties of relative interiors.

Proposition 2.1. *Let C be a non-empty convex set in $\mathbb{R}^n$. 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) $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$ *if and only if* $0 \notin \operatorname{ri}(C_1 - C_2)$.

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

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

This proves $\subseteq$. Suppose $z_1 \in \operatorname{ri}(C_1)$ and $z_2 \in \operatorname{ri}(C_2)$. 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

$$z_2 + t_1(z_2 - x_2) = \theta(z_2 + t_2(z_2 - x_2)) + (1 - \theta)z_2 \in C_2. \tag{2.3}$$

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

(viii) [MMN22, Theorem 2.92]

□

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

Definition 2.2. *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 2.1. *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*

$$\sup_{x \in C_2} \langle x, b \rangle \leq \inf_{x \in C_1} \langle x, b \rangle \quad \text{and} \quad \inf_{x \in C_2} \langle x, b \rangle < \sup_{x \in C_1} \langle x, b \rangle. \tag{2.4}$$

Proof. [Roc70, Theorem 11.1] □

Theorem 2.2. (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 $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$.*

Proof. [Roc70, Theorem 11.3] □

Definition 2.3. *Given a nonempty subset $\Omega \subseteq \mathbb{R}^n$ the **support function** $\sigma_\Omega : \mathbb{R}^n \to \overline{\mathbb{R}}$ of Ω is defined by*

$$\sigma_\Omega(x^*) := \sup_{x \in \Omega} \langle x^*, x \rangle \qquad \text{for } x^* \in \mathbb{R}^n. \tag{2.5}$$

Definition 2.4. *Given functions $f_i : \mathbb{R}^n \to (-\infty, \infty]$ for $i = 1, \ldots, n$ the **infimal convolution** of these functions is defined as*

$$(f_1 \square \ldots \square f_m)(x) := \inf_{\substack{x_i \in \mathbb{R}^n \\ \sum_{i=1}^m x_i = x}} \sum_{i=1}^m f_i(x_i) \tag{2.6}$$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = (\sigma_{C_1} \square \sigma_{C_2})(x^*) \qquad \text{for all } x^* \in \mathbb{R}^n. \tag{2.7}$$

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

$$(\sigma_{C_1 \cap C_2})(x^*) = \sigma_{C_1}(x_1^*) + \sigma_{C_2}(x_2^*). \tag{2.8}$$

Proof. *[MMN22, Theorem 4.23]* 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 intersection of their

relative interiors is empty. The rest of the proof is as in the cited reference at the beginning of the proof. To this end, consider the sets

$$\Theta_1 \;:=\; C_1 \times [\,0,\infty) \quad \text{and} \quad \Theta_2 \;:=\; \{(x,\lambda) \in \mathbb{R}^n \;:\; x \in C_2 \text{ and } \lambda \leq \langle x^*, x\rangle - \alpha\}\;. \tag{2.9}$$

Simplify proof with properties of relative interiors.

Clearly, Θ_1 is convex by the convexity of C_1. To see that Θ_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 φ and Θ_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 Θ_2. Next we show that the relative interiors of Θ_1 and Θ_2 do not intersect, i.e., $\operatorname{ri}\Theta_1 \cap \operatorname{ri}\Theta_2 = \emptyset$. First note that

$$\operatorname{ri}(\Theta_1) = \operatorname{ri}(C_1) \times \operatorname{ri}([0,\infty)) \subseteq \operatorname{ri}(C_1) \times (0,\infty). \tag{2.10}$$

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

$$\operatorname{ri}(\Theta_2) \subseteq \{(x,\lambda) \in \mathbb{R}^n \colon x \in \operatorname{ri}(C_2) \text{ and } \lambda < \langle x^*, x\rangle - \alpha\}\,. \tag{2.11}$$

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

$$0 \leq (1+t)(\langle x^*, x\rangle - \alpha - \lambda) + t(\mu - \langle x^*, y\rangle + \alpha) < 0, \tag{2.12}$$

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

$$0 < \lambda < \langle x^*, x\rangle - \alpha \leq 0, \tag{2.13}$$

a contradiction.

Thus, convex separation is applicable. Confer [MMN22, Theorem 4.23] about the rest of the proof. □

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

2.2 Conjugate Calculus and Fenchel-Rockafellar Theorem

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

Definition 2.5. (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*

$$f^*(x^*) := \sup_{x \in \mathbb{R}^n} (x^*)^T x - f(x) \tag{2.14}$$

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

Note that f in Definition 2.5 does not have to be convex. On the other hand, the convex conjugate is always convex:

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

Proof. [MMN22, Proposition 4.2] □

Theorem 2.4. (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*

$$(g \circ A)^*(x^*) = \inf_{y^* \in (A^*)^{-1}(x^*)} g^*(y^*). \tag{2.15}$$

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

Proof. [MMN22, Proposition 4.28] □

Theorem 2.5. *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*

$$(f + g)^*(x^*) = (f^* \square g^*)(x^*) \tag{2.16}$$

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

$$(f+g)^*(x^*) = f^*(x_1^*) + g^*(x_2^*), \quad x^* = x_1^* + x_2^*. \tag{2.17}$$

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

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

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

$$g(x) = \delta_\Omega(x) := \tag{2.18}$$

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

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

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

Proposition 2.3. *Consider the optimization problem 2.1 and its dual 2.2, where the functions f and g are not assumed to be convex. Define the **optimal values** of these problems by*

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

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

Proof. [MMN22, Proposition 4.62] □

Theorem 2.6. *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 (2.1) and (2.2) are equal, i.e.*

$$\inf_{x\in\mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y\in\mathbb{R}^n} \left\{-f^*\left(A^T y\right) - g^*(-y)\right\}. \tag{2.19}$$

Proof. [MMN22, Theorem 4.63] □

Insert lemma in chapter 1.

Lemma 2.1. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be convex. Then for all $y \in \mathbb{R}^n$ and $C > 0$*

$$\inf_{\|\Delta\|=C} f(y+\Delta) - f(y) \geq 0 \quad \Longrightarrow \quad \exists y^* \in \mathbb{R}^n \colon y^* \text{ is global minimum of } f \text{ and } \|y^* - y\| \leq C. \tag{2.20}$$

Proof. Since $\mathcal{C} := \{\|\Delta\| \leq C\}$ is convex f has a local minimum in $y + \mathcal{C} := \{y + \Delta \mid \|\Delta\| \leq C\}$. 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 (2.20) is true. Then it holds

$$f(x) < f(y^*) \quad \text{for some } x \in \mathbb{R}^n \setminus y + \mathcal{C}. \tag{2.21}$$

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.

$$\theta x + (1-\theta)y^* = y + \Delta_x \quad \text{for some } \theta \in (0,1) \text{ and } \Delta_x \text{ with } \|\Delta_x\| = C. \tag{2.22}$$

It follows

$$\begin{aligned} f(y^*) \leq f(y) \leq f(y + \Delta_x) &= f(\theta x + (1-\theta)y^*) \\ &\leq \theta f(x) + (1-\theta) f(y^*) < f(y^*), \end{aligned} \tag{2.23}$$

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 (2.20) being true, the equality is due to (2.22), the third inequality is due to the convexity of f and the strict inequality is due to (2.21). □

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

2.3 Tseng Bertsekas

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

Consider the following optimization problem

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

subject to the constraints

$$\mathbf{A}x \geq b, \tag{2.24}$$

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 2.1. *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* $\mathrm{dom}(f)$.

(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

$$p \geq 0, \tag{2.25}$$

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

$$q(p) := \min_{x\in\mathbb{R}^m} f(x) + \langle p, b - \mathbf{A}x\rangle = \langle p, b\rangle - f^*(\mathbf{A}^\top p). \tag{2.26}$$

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 2.7. [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 ith coordinate by $d_i(p)$. Since q is continuously differentiable, $d_i(p)$ is continuous, and since q is concave, $d_i(p)$ as nonincreasing in p_i.

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

$$d(p) = b - \mathbf{A}x, \quad \text{where} \quad x := \nabla f^*(\mathbf{A}^\top p) = \operatorname{argsup}_{\xi \in \mathbb{R}^m} \langle p, \mathbf{A}\xi \rangle - f(\xi). \tag{2.27}$$

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

Read and understand proof (p.80)

Proposition 2.4. (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*

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad x \mapsto \sup_{z \in Z} \phi(x, z) \tag{2.28}$$

is convex and has directional derivative given by

$$f'(x; y) = \sup_{z \in Z(x)} \phi'(x, z; y), \tag{2.29}$$

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

$$Z(x) := \left\{ \overline{z} \in \mathbb{R}^m : \phi(x, \overline{z}) = \sup_{z \in Z} \phi(x, z) \right\}. \tag{2.30}$$

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

Note that x is the unique vector satisfying

$$\mathbf{A}p \in \partial f(x). \tag{2.31}$$

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

$$p = [p + d(p)]^+, \tag{2.32}$$

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

$$\mathbf{A}x \geq b \quad \text{and} \quad p_i = 0 \quad \text{for all } i \text{ such that } \sum_{j=1}^{m} a_{ij}x_j > b_i. \tag{2.33}$$

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

n

Definition 2.6. [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

$$f_1(x) \leq 0, \ldots, f_r(x) \leq 0, \qquad f_{r+1}(x) = 0, \ldots, f_m(x) = 0, \tag{2.34}$$

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

Definition 2.7. *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 2.8. (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 α_i of $\overline{\alpha}$ satisfy the following conditions.*

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

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

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

Proof. [Roc70, Theorem 28.3] □

Takeaways The Karush-Kuhn-Tucker conditions allow us to derive duality in terms of the optimal solutions for strictly convex functions.

3 Random Matrix Inequalities

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

3.1 A Matrix Analysis Primer

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

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

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

For each $p \geq 1$ the **Schatten p-norm** is defined as $\|\mathbf{B}\|_p := (\mathrm{tr}(|\mathbf{B}|^p))^{1/p}$ 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 3.1. *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 λ is an eigenvalue of of $\mathbf{A}$, then $f(\lambda)$ is an eigenvalue of $f(\mathbf{A})$.

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

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.

3.2 The Method of Exchangeable Pairs

We first define an exchangable pair.

Definition 3.1. *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 3.2. *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*

$$\mathbf{X} := \mathbf{\Psi}(Z) \quad \textit{and} \quad \mathbf{X}' := \mathbf{\Psi}(Z'). \tag{3.2}$$

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

$$\mathbf{E}[\mathbf{X} - \mathbf{X}'|Z] = \alpha\mathbf{X} \qquad \textit{almost surely.} \tag{3.3}$$

The constant α 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 3.1. *Suppose that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair with scale factor α. 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*

$$\mathbf{E}[\mathbf{X} \cdot \mathbf{F}(\mathbf{X})] = \frac{1}{2\alpha}\mathbf{E}[(\mathbf{X} - \mathbf{X}')(\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{X}'))]. \tag{3.4}$$

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

$$\alpha \cdot \mathbf{E}[\mathbf{X} \cdot \mathbf{F}(\mathbf{X})] = \mathbf{E}[\mathbf{E}[\mathbf{X} - \mathbf{X}'|Z] \cdot \mathbf{F}(\mathbf{X})] = \mathbf{E}[(\mathbf{X} - \mathbf{X}')\mathbf{F}(\mathbf{X})] \tag{3.5}$$

□

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 3.3. *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*

$$\boldsymbol{\Delta}_{\mathbf{X}} := \boldsymbol{\Delta}_{\mathbf{X}}(Z) := \frac{1}{2\alpha}\mathbf{E}[(\mathbf{X} - \mathbf{X}')^2 | Z], \tag{3.6}$$

where α is the scale factor of the pair. We may take any version of the conditional expectation in this definition.

The conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$ 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 $\mathbf{F}(\mathbf{X}) = \mathbf{X}$

$$\mathbf{E}[\boldsymbol{\Delta}_{\mathbf{X}}] = \mathbf{E}[\mathbf{X}^2]. \tag{3.7}$$

To verify the regularity condition, note that

$$\mathbf{E}[\|(\mathbf{X} - \mathbf{X}')\mathbf{X}\|] \leq \mathbf{E}[\|\mathbf{X}\|^2] + \mathbf{E}[\|\mathbf{X}\| \cdot \|\mathbf{X}'\|] \leq 2\mathbf{E}[\|\mathbf{X}\|^2] < \infty. \tag{3.8}$$

Example 3.1. *[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.

3.3 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 3.2. (Generalized Klein inequality) *Let $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ be real-valued functions on an interval I of the real line. Suppose*

$$\sum_{k=1}^{n} u_k(a) v_k(b) \geq 0 \qquad \text{for all } a, b \in I. \tag{3.9}$$

Then

$$\overline{\operatorname{tr}} \left(\sum_{k=1}^{n} u_k(\mathbf{A}) v_k(\mathbf{B}) \right) \geq 0 \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}_d(I). \tag{3.10}$$

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

$$\overline{\operatorname{tr}} \left(\sum_{k=1}^{n} u_k(\mathbf{A}) v_k(\mathbf{B}) \right) = \sum_{k=1}^{n} \sum_{i,j=1}^{d} \overline{\operatorname{tr}} \left(u_k(\lambda_i) \mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* v_k(\mu_j) \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right) \tag{3.11}$$

$$= \sum_{i,j=1}^{d} \overline{\operatorname{tr}} \left(\mathbf{P}_{\bullet i} \mathbf{P}_{\bullet i}^* \mathbf{Q}_{\bullet j} \mathbf{Q}_{\bullet j}^* \right) \sum_{k=1}^{n} u_k(\lambda_i) v_k(\mu_j) \geq 0 \tag{3.12}$$

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 $i, j \in \{1, \ldots, d\}$, we apply a well known extension of von Neumann's trace inequality [Ruh70, Lemma 1], namely

$$\operatorname{tr}(\mathbf{PQ}) \geq \sum_{i=1}^{d} p_i q_{d-i+1} \geq 0 \qquad \text{for all } \mathbf{P}, \mathbf{Q} \in \mathbb{H}_d([0, \infty)), \tag{3.13}$$

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

Lemma 3.2. (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

$$\overline{\mathrm{tr}}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (h(\mathbf{A}) - h(\mathbf{B}))] \leq \frac{1}{2}\,\overline{\mathrm{tr}}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))]. \tag{3.14}$$

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

$$(g(a) - g(b)) \cdot (h(a) - h(b)) = (g(a) - g(b)) \cdot (a - b) \int_0^1 h^{'}(\tau a + (1 - \tau)b)\mathrm{d}\tau \tag{3.15}$$

$$\leq (g(a) - g(b)) \cdot (a - b) \int_0^1 [\tau h^{'}(a) + (1 - \tau)h^{'}(b)]\mathrm{d}\tau \tag{3.16}$$

$$= \frac{1}{2}\,[(g(a) - g(b)) \cdot (a - b) \cdot (h^{'}(a) + h^{'}(b))]. \tag{3.17}$$

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

$$\begin{aligned}
&(g(a) - g(b)) \cdot (a - b) \cdot (h^{'}(a) + h^{'}(b)) \\
&\quad = [g(a) \cdot a \cdot h^{'}(a)] + [g(a) \cdot a] \cdot h^{'}(b) - b \cdot [h^{'}(a) \cdot g(a)] - [b \cdot h^{'}(b)] \cdot g(a) \\
&\qquad + [\text{ the same as above with } a \text{ and } b \text{ reversed }](a \rightleftarrows b)
\end{aligned} \tag{3.18}$$

Taking the trace yields

$$\begin{aligned}
&\mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{A} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] - \mathrm{tr}[\mathbf{B} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B})) \cdot g(\mathbf{A})] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\
&\quad = \mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{A} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] - \mathrm{tr}[g(\mathbf{A}) \cdot \mathbf{B} \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\
&\quad = \mathrm{tr}[g(\mathbf{A}) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))] + (\mathbf{A} \rightleftarrows \mathbf{B}) \\
&\quad = \mathrm{tr}[(g(\mathbf{A}) - g(\mathbf{B})) \cdot (\mathbf{A} - \mathbf{B}) \cdot (h^{'}(\mathbf{A}) + h^{'}(\mathbf{B}))].
\end{aligned} \tag{3.19}$$

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

$$\operatorname{tr}(\mathbf{BC}) \leq \|\mathbf{B}\|_p \|\mathbf{C}\|_q \qquad \text{for all } \mathbf{B}, \mathbf{C} \in \mathbb{M}_d. \tag{3.20}$$

Proof. [Bha97, Corollary IV.2.6] □

We are now ready to prove the auxiliary theorem.

Theorem 3.1. (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*

$$\mathbf{E}[\|\mathbf{X}\|_{2p}^{2p}]^{1/(2p)} \leq \sqrt{2p-1}\ \mathbf{E}[\|\mathbf{\Delta_X}\|_p^p]^{1/(2p)}, \tag{3.21}$$

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 α. 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,

$$E := \mathbf{E}[\|\mathbf{X}\|_{2p}^{2p}] = \mathbf{E}[\operatorname{tr}(|\mathbf{X}|^{2p})]. \tag{3.22}$$

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

$$E = \mathbf{E}[\operatorname{tr}(|\mathbf{X}| \cdot |\mathbf{X}|^{2p-1})] = \mathbf{E}[\operatorname{tr}(\mathbf{X} \cdot \operatorname{sgn}(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1})]. \tag{3.23}$$

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

$$E = \frac{1}{2\alpha}\mathbf{E}[\operatorname{tr}((\mathbf{X} - \mathbf{X}') \cdot (\operatorname{sgn}(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1} - \operatorname{sgn}(\mathbf{X}') \cdot |\mathbf{X}'|^{2p-1}))] \tag{3.24}$$

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

Theorem 3.2. [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*

$$\mathbf{E}[\mathbf{Y}_k] = 0 \quad \text{and} \quad \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2 \qquad \text{almost surely for all } k \geq 1. \tag{3.25}$$

Then

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{p-\frac{1}{2}}\left\|\left(\sum_{k\geq 1}(\mathbf{A}_k^2+\mathbf{E}[\mathbf{Y}_k^2])\right)^{1/2}\right\|_{2p}. \quad (3.26)$$

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

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\xi_k\mathbf{A}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{2p-1}\left\|\left(\sum_{k\geq 1}\mathbf{A}_k^2\right)^{1/2}\right\|_{2p}. \quad (3.27)$$

Theorem 3.3. *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

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{S}_k\right\|^p\right]^{1/p} \leq \left[\left\|\sum_{k\geq 1}\mathbf{E}[\mathbf{S}_k]\right\|^{1/2} + 2\sqrt{er}\,\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/(2p)}\right]^2. \quad (3.28)$$

(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

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|^p\right]^{1/p} \leq \sqrt{er}\left\|\left(\sum_{k\geq 1}\mathbf{E}[\mathbf{Y}_k^2]\right)^{1/2}\right\| + 2er\,\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/p}. \quad (3.29)$$

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.

3.4 Generalzed Inequalities by Hermitian Dilataition

Definition 3.4. (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

$$\mathfrak{H}(B) := \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix} \tag{3.30}$$

Theorem 3.4. (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:*

$$v(\mathbf{S}) := \left\| \mathbf{E}[\mathbf{S}\mathbf{S}^\top] \right\| \vee \left\| \mathbf{E}[\mathbf{S}^\top \mathbf{S}] \right\| = \left\| \sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k \mathbf{A}_k^\top] \right\| \vee \left\| \sum_{k=1}^{n} \mathbf{E}[\mathbf{A}_k^T \mathbf{A}_k] \right\|. \tag{3.31}$$

Then

$$\left(\mathbf{E}\left[\|\mathbf{S}\|^2 \right] \right)^{\frac{1}{2}} \leq \sqrt{2ev(\mathbf{S}) \log(d_1 + d_2)} + 4e \left(\mathbf{E}[\max_{k \leq n} \|\mathbf{A}_k\|^2] \right)^{\frac{1}{2}} \log(d_1 + d_2). \tag{3.32}$$

Remark 3.1. *Since* $\mathbf{E}[\|S\|] \leq \mathbf{E}[\|S\|^2]^{\frac{1}{2}}$ *by the Cauchy-Schwarz inequality, Theorem 3.4 also holds with* $\mathbf{E}[\|S\|]$ *on the left-hand side of* (3.32)*. To obtain a tail bound we can employ the Markov inequality and Theorem 3.4:*

$$\begin{aligned}&\mathbf{P}[\,\|S\| \geq t] \\ &\quad \leq \frac{\mathbf{E}[\|S\|]}{t} \leq \frac{1}{t}\left(\sqrt{2ev(\mathbf{S})\log(d_1+d_2)} + 4e\left(\mathbf{E}[\max_{k\leq n}\|\mathbf{A}_k\|^2]\right)^{\frac{1}{2}}\log(d_1+d_2)\right) \quad \textit{for } t>0.\end{aligned} \tag{3.33}$$

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

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.

3.5 Intrinsic Dimension

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

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

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

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

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

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

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

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

□

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

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

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

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

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

and consequently

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

♢

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

4 Empirical Processes

4.1 A Primer on Empirical Processes

Add outer probability calculus. [vdvW13] p.6

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 σ-algebra on $\mathbb{D}$, the smallest σ-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

$$\int_{\mathbb{D}} f \mathrm{d}\mathbf{P}_n \to \int_{\mathbb{D}} f \mathrm{d}\mathbf{P} \qquad \text{for all } f \in C_b(\mathbb{D}). \tag{4.1}$$

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

$$\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)] \qquad \text{for all } f \in C_b(\mathbb{D}). \tag{4.2}$$

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 σ-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 4.1. *[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 σ-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 ω 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 ω 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 σ-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

$$f \mapsto G_n f := \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{P}_X f \right), \tag{4.3}$$

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

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

4.2 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

$$f \mapsto G_n f := \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{P}_X f \right), \tag{4.5}$$

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

$$\|G_n f\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |G_n f| . \tag{4.6}$$

Lemma 4.1. (Bernstein Inequality for Empirical Processes) *For any bounded, measurable function f it holds for all $t > 0$*

$$\mathbf{P}\left(|G_n f| > t \right) \leq 2 \exp\left(-\frac{1}{4} \frac{t^2}{\mathbf{P}_X(f^2) + t \|f\|_\infty / \sqrt{n}} \right) \tag{4.7}$$

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

$$\mathbf{P}\left(G_n f > t\right) \leq e^{-\lambda t}\mathbf{E}\exp\left(\lambda G_n f\right) \tag{4.8}$$

□

Lemma 4.2. *For any finite class $\mathcal{F}$ of bounded, measurable, square-integrable functions, with $|\mathcal{F}|$ elements, it holds*

$$\mathbf{E}\left\|G_n f\right\|_{\mathcal{F}} \lesssim \max_{f\in\mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}} \log\left(1+|\mathcal{F}|\right) + \max_{f\in\mathcal{F}} \|f\|_{\mathbf{P},2} \sqrt{\log\left(1+|\mathcal{F}|\right)}. \tag{4.9}$$

Lemma 4.3. *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,*

$$\mathbf{E}^*_{\mathbf{P}}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J_{[]}(\delta, \mathcal{F}, L_2(\mathbf{P})) + \sqrt{n}\mathbf{P}^* F\left\{F > \sqrt{n}a(\delta)\right\}. \tag{4.10}$$

Corollary 4.0.1. *For any class $\mathcal{F}$ of measurable functions with envelope function F,*

$$\mathbf{E}^*_{\mathbf{P}}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J_{[]}(\|F\|_{\mathbf{P},2}, \mathcal{F}, L_2(\mathbf{P})). \tag{4.11}$$

4.3 Functional Delta Method

Definition 4.1. *A map $\phi : \mathbb{D}_\phi \to \mathbb{E}$, defined on a subset $\mathbb{D}_\phi$ of a normed space $\mathbb{D}$ that contains θ, is called **Hadamard diffenertiable** at θ if there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that*

$$\left\|\frac{\phi(\theta + th_t) - \phi(\theta)}{t} - \phi'_\theta(h)\right\|_{\mathbb{E}} \to 0 \quad \text{as } t \searrow 0 \text{ for all } h_t \to h \tag{4.12}$$

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

Theorem 4.1. (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 θ 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 ϕ'_θ 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] □

5 Simple yet useful Calculations

Theorem 5.1. (Multivariate Taylor Theorem) *Let $f \in C^2(\mathbb{R}^n, \mathbb{R})$. Then for all $x, \Delta \in \mathbb{R}^n$ there exists $\xi \in [0,1]$ such that it holds*

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

Corollary 5.1.1. *Let $f \in C^2(\mathbb{R})$. Then for all $a, x, \Delta \in \mathbb{R}^n$ there exist $\xi \in [0,1]$ such that it holds*

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

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

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

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

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

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

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

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

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

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

Notation Index

$\#A$ cardinality of the set A

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

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

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

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

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

$\mathbf{P}$ generic probability measure

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

$\mathbb{R}$ set of real numbers

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

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

Bibliography

[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.

[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.

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

[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.

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

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

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

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