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



Ioan Scheffel

December 19, 2022

## **Contents**

1	intr	oduction	J
2	Balancing Weights		
	2.1	Introduction	
	2.2	Estimating the Population Mean of Potential Outcomes	
	2.3	Application of Convex Optimization	
	2.4	Application of Matrix Concentration Inequalities	5
3	Convex Analysis		
	3.1	Basic Notions	6
	3.2	Relative Interior	6
	3.3	Conjugate Calculus	6
	3.4	Tseng Bertsekas	6
4	Random Matrix Inequalities		
	4.1	Matrix Analysis	9
	4.2	Matrix Concentration Inequalities via the Method of Exchangeable Pairs	9
	4.3	Matrix Khintchin Inequality	9
	4.4	Matrix Moment Inequality	
	4.5	Intrinsic Dimension	
5	Emp	pirical Processes	11
6	Sim	uple vet useful Calculations	12

## 1 Introduction

## 2 Balancing Weights

#### 2.1 Introduction

## 2.2 Estimating the Population Mean of Potential Outcomes

### 2.3 Application of Convex Optimization

**Assumption 2.1.** 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 int(dom(f)).
- (iii) The derivative of f on int(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  $t \in \mathbb{R}$ .

We consider the following optimization 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

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

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

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

**Theorem 2.1.** 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  $t \mapsto f^*(t) = t(f')^{-1}(t) - f((f')^{-1}(t))$  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

$$w_i^* = (f')^{-1}(\langle B(X_i), \lambda^{\dagger} \rangle), \quad i \in \{1, \dots, n\}$$
 (2.1)

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

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

subject to

$$\lambda_k \ge 0 \quad \text{for all } k \in \{1, \dots, K\}, \tag{2.2}$$

where

$$\mathbf{Q} := \begin{bmatrix} \mathbf{I}_n \\ \mathbf{B}(\mathbf{X}) \\ -\mathbf{B}(\mathbf{X}) \end{bmatrix}, \quad \mathbf{B}(\mathbf{X}) := \begin{bmatrix} B(X_1), \dots, B(X_n) \end{bmatrix}, \quad and \quad d := \begin{bmatrix} 0_n \\ \delta \\ \delta \end{bmatrix}. \quad (2.3)$$

2.4 Application of Matrix Concentration Inequalities

## 3 Convex Analysis

#### 3.1 Basic Notions

#### 3.2 Relative Interior

#### 3.3 Conjugate Calculus

#### 3.4 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

$$\mathbf{A}x \ge b,\tag{3.1}$$

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

**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 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

$$p \ge 0,\tag{3.2}$$

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

$$q(p) := \min_{x \in \mathbb{R}^m} f(x) + \langle p, b - \mathbf{A}x \rangle = \langle p, b \rangle - f^*(\mathbf{A}^\top p).$$
 (3.3)

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

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

$$d(p) = b - \mathbf{A}x$$
, where  $x := \nabla f^*(\mathbf{A}^\top p) = \operatorname{argsup}_{\xi \in \mathbb{R}^m} \langle p, \mathbf{A}\xi \rangle - f(\xi)$ . (3.4)

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

**Proposition 3.1.** (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

$$f: \mathbb{R}^n \to \mathbb{R}, \qquad x \mapsto \sup_{z \in Z} \phi(x, z)$$
 (3.5)

is convex and has directional derivative given by

$$f'(x;y) = \sup_{z \in Z(x)} \phi'(x,z;y), \tag{3.6}$$

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

$$Z(x) := \left\{ \overline{z} \in \mathbb{R}^m : \phi(x, \overline{z}) = \sup_{z \in Z} \phi(x, z) \right\}.$$
 (3.7)

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

$$\mathbf{A}p \in \partial f(x). \tag{3.8}$$

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

$$p = [p + d(p)]^+,$$
 (3.9)

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

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

$$\mathbf{A}x \ge b$$
 and  $p_i = 0$  for all  $i$  such that  $\sum_{j=1}^{m} a_{ij}x_j > b_i$ . (3.10)

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

**Definition 3.1.** [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

$$f_1(x) \le 0, \dots, f_r(x) \le 0, \qquad f_{r+1}(x) = 0, \dots, f_m(x) = 0,$$
 (3.11)

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 3.2.** 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 3.2.** (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  $i \in \{1, \dots, r\}$ .
- (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

#### 4.1 Matrix Analysis

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

$$f(\mathbf{A}) = \mathbf{Q} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} \mathbf{Q}^* \quad \text{where} \quad \mathbf{A} = \mathbf{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \mathbf{Q}^*. \quad (4.1)$$

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**.

**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  $f(\mathbf{A})$ , then  $f(\lambda)$  is an eigenvalue of  $f(\mathbf{A})$ .
- (ii)  $f(a) \le g(a)$  for all  $a \in I$  implies  $f(\mathbf{A}) \preceq g(\mathbf{A})$ .

# 4.2 Matrix Concentration Inequalities via the Method of Exchangeable Pairs

#### 4.3 Matrix Khintchin Inequality

**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  $\mathbf{E}[\|\mathbf{X}\|_{2p}^{2p}] < .$ 

**Theorem 4.2.** [MJC<sup>+</sup>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

$$\mathbf{E}[\mathbf{Y}_k] = 0$$
 and  $\mathbf{Y}_k^2 \leq \mathbf{A}_k^2$  almost surely for all  $k \geq 1$ . (4.2)

Then

$$\mathbf{E} \left[ \left\| \sum_{k \ge 1} \mathbf{Y}_k \right\|_{2p}^{2p} \right]^{1/(2p)} \le \sqrt{p - \frac{1}{2}} \left\| \left( \sum_{k \ge 1} (\mathbf{A}_k^2 + \mathbf{E}[\mathbf{Y}_k^2]) \right)^{1/2} \right\|_{2p}. \tag{4.3}$$

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

$$\mathbf{E} \left[ \left\| \sum_{k \ge 1} \xi_k \mathbf{A}_k \right\|_{2p}^{2p} \right]^{1/(2p)} \le \sqrt{2p - 1} \left\| \left( \sum_{k \ge 1} \mathbf{A}_k^2 \right)^{1/2} \right\|_{2p}. \tag{4.4}$$

#### 4.4 Matrix Moment Inequality

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

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{S}_{k}\right\|^{p}\right]^{1/p} \leq \left[\left\|\sum_{k\geq 1}\mathbf{E}[\mathbf{S}_{k}]\right\|^{1/2} + 2\sqrt{er}\mathbf{E}\left[\max_{k\geq 1}\left\|\mathbf{S}_{k}\right\|^{p}\right]^{1/(2p)}\right]^{2}.$$
 (4.5)

(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

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_{k}\right\|^{p}\right]^{1/p} \leq \sqrt{er}\left\|\left(\sum_{k\geq 1}\mathbf{E}[\mathbf{Y}_{k}^{2}]\right)^{1/2}\right\| + 2er\mathbf{E}\left[\max_{k\geq 1}\left\|\mathbf{S}_{k}\right\|^{p}\right]^{1/p}.$$
 (4.6)

#### 4.5 Intrinsic Dimension

**Definition 4.1.** For a positive-semidefinite matrix **S**, the **intrinic dimension** is the quantity

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

**Lemma 4.1.** (Intrinsic dimension) 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

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

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

$$\varphi(a) \le \left(1 - \frac{a}{L}\right)\varphi(0) + \frac{a}{L}\varphi(L) = \frac{a}{L}\varphi(L) \quad \text{for all } a \in [0, L].$$
 (4.7)

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

$$\operatorname{tr}(\varphi(\mathbf{S})) = \sum_{i=1}^{d} \varphi(\lambda_i) \le \frac{\sum_{i=1}^{d} \lambda_i}{\|\mathbf{S}\|} \varphi(\|\mathbf{S}\|) = \frac{\operatorname{tr}(\mathbf{S})}{\|\mathbf{S}\|} \varphi(\|\mathbf{S}\|) = \operatorname{intdim}(\mathbf{S}) \cdot \varphi(\|\mathbf{S}\|). \tag{4.8}$$

## 5 Empirical Processes

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

$$f \mapsto G_n f := \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbf{P}_X f \right),$$
 (5.1)

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

$$||G_n f||_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |G_n f|. \tag{5.2}$$

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

$$\mathbf{P}(|G_n f| > t) \le 2 \exp\left(-\frac{1}{4} \frac{t^2}{\mathbf{P}_X(f^2) + t \|f\|_{\infty} / \sqrt{n}}\right)$$
 (5.3)

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

$$\mathbf{P}(G_n f > t) \le e^{-\lambda t} \mathbf{E} \exp(\lambda G_n f)$$
(5.4)

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

$$\mathbf{E} \|G_n f\|_{\mathcal{F}} \lesssim \max_{f \in \mathcal{F}} \frac{\|f\|_{\infty}}{\sqrt{n}} \log \left(1 + |\mathcal{F}|\right) + \max_{f \in \mathcal{F}} \|f\|_{\mathbf{P}, 2} \sqrt{\log \left(1 + |\mathcal{F}|\right)}. \tag{5.5}$$

## 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

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

$$+ \frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2 f(x + \xi \Delta)}{\partial x_i^2} \Delta_i^2$$
(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

$$f(a^{T}(x+\Delta)) - f(a^{T}x) = f'(a^{T}x) \Delta^{T}a + \frac{1}{2}f''(a^{T}(x+\xi\Delta)) \Delta^{T}A \Delta,$$
 (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  $\xi \in [0, 1]$ 

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

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

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

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

**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

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

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**

- [BT03] Dimitri P. Bertsekas and John N. Tsitsiklis. Parallel and Distributed Computation:Numerical Methods. November 2003.
- [MJC<sup>+</sup>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.
- [Roc70] R. Tyrrell Rockafellar. Convex Analysis. Princeton University Press, 1970.
- [Tro15] Joel A. Tropp. An Introduction to Matrix Concentration Inequalities, January 2015.