*3* **Subadditive**: $\|\cdot\|$ *satisfies the triangle inequality* $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\|+\|\boldsymbol{y}\|$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$.

One very important family of norms are the $\ell^p$ norms. If we take $\mathbb{V} = \mathbb{R}^n$, and $p \in [1, \infty)$, we can write

$$\|\boldsymbol{x}\|_p = \left( \sum_i |x_i|^p \right)^{1/p}. \tag{A.9.1}$$

The most familiar example is the $\ell^2$ norm or "Euclidean norm"

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{\boldsymbol{x}^* \boldsymbol{x}},$$

which coincides with our usual way of measuring lengths. Two other cases are of almost equal importance: $p = 1$, and $p \to \infty$. Setting $p = 1$ in (A.9.1), we obtain

$$\|\boldsymbol{x}\|_1 = \sum_i |x_i|, \tag{A.9.2}$$

Finally, as $p$ becomes larger, the expression in (A.9.1) accentuates large $|x_i|$. As $p \to \infty$, $\|\boldsymbol{x}\|_p \to \max_i |x_i|$. We extend the definition of the $\ell^p$ norm to $p = \infty$ by defining

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|. \tag{A.9.3}$$

However, the $\ell^p$ norms are far from the only norms on vectors.

Example A.40. *The following are examples of norms:*

- *For* $p \geq 1$, $\|\boldsymbol{x}\|_p$ *is a norm.*
- *Every positive definite matrix* $\boldsymbol{P} \succ \boldsymbol{0}$ *defines a norm, via* $\|\boldsymbol{x}\|_{\boldsymbol{P}} = \sqrt{\boldsymbol{x}^* \boldsymbol{P} \boldsymbol{x}}$.
- *For* $\boldsymbol{x} \in \mathbb{R}^n$, *let* $[\boldsymbol{x}]_{(k)}$ *denote the* $k$*-th largest element of the sequence:* $|\boldsymbol{x}_1|, |\boldsymbol{x}_2|, \ldots, |\boldsymbol{x}_n|$. *Then*

$$\|\boldsymbol{x}\|_{[K]} = \sum_{k=1}^{K} [\boldsymbol{x}]_{(k)} \tag{A.9.4}$$

  *is a norm.*
- *For* $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, *the Frobenius norm* $\|\boldsymbol{X}\|_F = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X} \rangle}$ *is a norm.*

One fundamental result in the theory of normed spaces is that in finite dimensions, all norms are comparable:

Theorem A.41 (Equivalence of Norms). *Let* $\|\cdot\|_a$ *and* $\|\cdot\|_b$ *be two norms on a finite dimenisonal vector space* $\mathbb{V}$. *Then there exist* $\alpha, \beta > 0$ *such that for every* $\boldsymbol{v} \in \mathbb{V}$,

$$\alpha \|\boldsymbol{v}\|_a \leq \|\boldsymbol{v}\|_b \leq \beta \|\boldsymbol{v}\|_a. \tag{A.9.5}$$

It is important not to over-interpret this result. "Equivalence" here means that the values of the norms can be compared up to constants, as in (A.9.5). It does not mean that the norms behave in the same way – they may produce very different results when selected to define constraint sets, or as objective functions for optimization. For purposes of analysis, it is useful to note the following comparisons

LEMMA A.42 (Comparisons between $\ell^p$ Norms). *For all* $\boldsymbol{x} \in \mathbb{R}^n$,

- $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{n}\,\|\boldsymbol{x}\|_2$,
- $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\,\|\boldsymbol{x}\|_\infty$,
- $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_1 \leq n\,\|\boldsymbol{x}\|_\infty$.

To each norm, we can associate a *dual norm*. To do this precisely, we need to define a normed linear space. If $\mathbb{V}$ is a vector space and $\|\cdot\|$ is a norm on $\mathbb{V}$, we call the pair $(\mathbb{V}, \|\cdot\|)$ a *normed linear space*. A *linear functional* is a linear map $\boldsymbol{\phi} : \mathbb{V} \to \mathbb{R}$. Since linear combinations of linear functionals are again linear functionals, the space of all linear functionals on a given vector space $\mathbb{V}$ is itself a vector space (called the "topological dual" of $\mathbb{V}$). On this space, we can define another function

$$\|\boldsymbol{\phi}\|^* = \sup_{\boldsymbol{v} \in \mathbb{V},\ \|\boldsymbol{v}\| \leq 1} |\boldsymbol{\phi}(\boldsymbol{v})|. \tag{A.9.6}$$

As the notation suggests, $\|\boldsymbol{\phi}\|^*$ is a norm, if we restrict to $\boldsymbol{\phi}$ for which the supremum is finite:

DEFINITION A.43 (Dual Space and Dual Norm). *The normed dual of the space* $(\mathbb{V}, \|\cdot\|)$ *is the space* $(\mathbb{V}^*, \|\cdot\|^*)$*, where the dual norm* $\|\cdot\|^*$ *of a linear functional* $\boldsymbol{\phi} : \mathbb{V} \to \mathbb{R}$ *is defined as in* (A.9.6) *and*

$$\mathbb{V}^* = \left\{ \boldsymbol{\phi} : \mathbb{V} \to \mathbb{R} \text{ linear} \mid \|\boldsymbol{\phi}\|^* < +\infty \right\}. \tag{A.9.7}$$

This definition may seem somewhat abstract; for our purposes, the dual spaces and dual norms we encounter will have fairly concrete descriptions:

THEOREM A.44. *Let* $\langle \cdot, \cdot \rangle$ *denote the standard inner product on* $\mathbb{R}^n$ *(and by extension on* $\mathbb{R}^{m \times n}$*). Every linear functional* $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}$ *can be written as*

$$\boldsymbol{\phi}(\boldsymbol{x}) = \langle \boldsymbol{v}, \boldsymbol{x} \rangle, \tag{A.9.8}$$

*for some vector* $\boldsymbol{v} \in \mathbb{R}^n$*. Similarly, every linear functional* $\boldsymbol{\phi} : \mathbb{R}^{m \times n} \to \mathbb{R}$ *can be written as*

$$\boldsymbol{\phi}(\boldsymbol{X}) = \langle \boldsymbol{V}, \boldsymbol{X} \rangle, \tag{A.9.9}$$

*for some matrix* $\boldsymbol{V} \in \mathbb{R}^{m \times n}$*.*

The implication of this is that if we are considering a space $(\mathbb{R}^n, \|\cdot\|_\sharp)$, the dual space can be identified with $(\mathbb{R}^n, \|\cdot\|_\sharp^*)$, where

$$\|\boldsymbol{v}\|_\sharp^* = \sup_{\|\boldsymbol{x}\|_\sharp \leq 1} \langle \boldsymbol{v}, \boldsymbol{x} \rangle. \tag{A.9.10}$$

In particular, we have the following examples:

EXAMPLE A.45 (Duals of Common Norms). *Check the following:*

- *The dual of the $\ell^\infty$ norm is the $\ell^1$ norm.*
- *The dual of the $\ell^1$ norm is the $\ell^\infty$ norm.*
- *The $\ell^2$ and Frobenius norms are self-dual; i.e., $\|\cdot\|_2^* = \|\cdot\|_2$ and $\|\cdot\|_F^* = \|\cdot\|_F$.*
- *If $p, q \in [1, \infty)$, with $p^{-1} + q^{-1} = 1$, then $\|\cdot\|_p^* = \|\cdot\|_q$ and $\|\cdot\|_q^* = \|\cdot\|_p$.*

It is immediate from the definition that for any $\boldsymbol{x}, \boldsymbol{x}'$, and any norm $\|\cdot\|$,

$$\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \;\leq\; \|\boldsymbol{x}\| \, \|\boldsymbol{x}'\|^* \,. \tag{A.9.11}$$

If we take $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$, we obtain the Cauchy-Schwarz inequality.

*Matrix and Operator Norms.*
Even more interesting structure can arise when $\mathbb{V}$ is a space of matrices, e.g., $\mathbb{V} = \mathbb{R}^{m \times n}$, due to the interpretation of a matrix as a linear operator. For square matrices, many authors reserve the term "matrix norm" for a function $\|\cdot\|$ that satisfies the three criteria in Definition A.39, *and* is *submultiplicative*

$$\|\boldsymbol{A}\boldsymbol{B}\| \;\leq\; \|\boldsymbol{A}\| \, \|\boldsymbol{B}\| \,. \tag{A.9.12}$$

They use the term "vector norm on matrices" for functions on $\mathbb{V}$ that only satisfy Definition A.39. We will not emphasize this distinction in terminology. Nevertheless, the submultiplicative property (A.9.12) is often useful, and we will note it where it occurs.

The most important source of norms on matrices comes from the notion of a matrix as a linear operator:

DEFINITION A.46 (Operator Norm). *Let $(\mathbb{W}, \|\cdot\|_a)$ and $(\mathbb{W}', \|\cdot\|_b)$ be two normed linear spaces, and let $\mathcal{L} : \mathbb{W} \to \mathbb{W}'$. The* operator norm *of $\mathcal{L}$ is*

$$\|\mathcal{L}\|_{a \to b} = \sup_{\|\boldsymbol{w}\|_a \leq 1} \|\mathcal{L}[\boldsymbol{w}]\|_b \,. \tag{A.9.13}$$

Specializing the definition a bit, for an $m \times n$ matrix $\boldsymbol{A}$, if $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, we write

$$\|\boldsymbol{A}\|_{a \to b} = \sup_{\|\boldsymbol{x}\|_a \leq 1} \|\boldsymbol{A}\boldsymbol{x}\|_b \,. \tag{A.9.14}$$

The most important special case is

THEOREM A.47. *The norm of a matrix $\boldsymbol{A}$ as an operator from $\ell_n^2 = (\mathbb{R}^n, \|\cdot\|_2)$ to $\ell_m^2 = (\mathbb{R}^m, \|\cdot\|_2)$ is*

$$\|\boldsymbol{A}\|_{2 \to 2} = \sigma_1(\boldsymbol{A}). \tag{A.9.15}$$

Several other cases are of interest:

THEOREM A.48. *The norm of any matrix as an operator from* $(\mathbb{R}^n, \|\cdot\|_1)$ *to any normed space* $(\mathbb{R}^m, \|\cdot\|_\sharp)$ *is simply the largest* $\|\cdot\|_\sharp$ *of any column of* $\boldsymbol{A}$*:*

$$\|\boldsymbol{A}\|_{1\to\sharp} = \max_{j=1,\dots,n} \|\boldsymbol{A}\boldsymbol{e}_j\|_\sharp \,. \tag{A.9.16}$$

*The norm of any matrix as an operator from* $(\mathbb{R}^n, \|\cdot\|_\flat)$ *for any norm* $\|\cdot\|_\flat$ *into* $(\mathbb{R}^m, \|\cdot\|_\infty)$ *is the largest dual norm of any of the rows:*

$$\|\boldsymbol{A}\|_{\flat\to\infty} = \max_{i=1,\dots,m} \|\boldsymbol{e}_i^* \boldsymbol{A}\|_\flat^* \,, \tag{A.9.17}$$

*where the dual norm* $\|\cdot\|_\flat^*$ *is*

$$\|\boldsymbol{v}\|_\flat^* = \sup_{\|\boldsymbol{u}\|_\flat \le 1} \langle \boldsymbol{u}, \boldsymbol{v} \rangle \,. \tag{A.9.18}$$

For example, $\|\boldsymbol{A}\|_{1\to 1}$ is just the largest $\ell^1$ norm of any column of $\boldsymbol{A}$.

*Unitary Invariant Matrix Norms.*
It is interesting to note that the operator norm of a matrix $\boldsymbol{A}$ depends only on the singular values of $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_{2,2} = \sigma_1(\boldsymbol{A}) = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_\infty \,, \tag{A.9.19}$$

where $\boldsymbol{\sigma}(\boldsymbol{A})$ is the vector of singular values. In fact, the Frobenius norm $\|\boldsymbol{A}\|_F$ depends only on the singular values as well:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{A})^2} = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_2 \,. \tag{A.9.20}$$

This fact is not too difficult to observe from the orthogonal invariance of $\|\cdot\|_F$:

$$\forall\, \boldsymbol{A} \in \mathbb{R}^{m\times n},\ \boldsymbol{P} \in \mathsf{O}(m),\ \boldsymbol{Q} \in \mathsf{O}(n), \quad \|\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q}\|_F = \|\boldsymbol{A}\|_F \,. \tag{A.9.21}$$

This suggests a pattern. In fact, any $\ell^p$ norm of the singular values is a norm on matrices $\boldsymbol{A}$:

DEFINITION A.49 (Schatten $p$-Norm). *For* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$*, let* $\boldsymbol{\sigma}(\boldsymbol{A}) \in \mathbb{R}^{\min\{m,n\}}$ *denote the vector of singular values. For* $p \in [1, \infty]$*, the function*

$$\|\boldsymbol{A}\|_{S_p} = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_p \tag{A.9.22}$$

*is a norm on* $\mathbb{R}^{m\times n}$*.*

It is easy to recognize the operator norm and Frobenius norm as special cases. One other special case is of great interest – the Schatten 1-norm

$$\|\boldsymbol{A}\|_{S_1} = \sum_i \sigma_i(\boldsymbol{A}). \tag{A.9.23}$$

This is also sometimes called the *trace norm* or *nuclear norm*. We reserve a special notation

$$\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A}) \tag{A.9.24}$$

for this norm. The operator norm $\|\cdot\|_{2,2}$ and the nuclear norm $\|\cdot\|_*$ are dual norms.

We have defined several interesting, useful norms on matrices $\boldsymbol{A}$, by applying different vector norms to the singular values $\boldsymbol{\sigma}(\boldsymbol{A})$. Because the singular values are orthogonal invariant, i.e., for $\boldsymbol{P} \in \mathsf{O}(m)$, $\boldsymbol{Q} \in \mathsf{O}(n)$, $\boldsymbol{\sigma}(\boldsymbol{PAQ}) = \boldsymbol{\sigma}(\boldsymbol{A})$, norms defined in this way are also orthogonal invariant. It is natural to ask whether every function $\|\boldsymbol{\sigma}(\boldsymbol{A})\|$ generates a valid norm on $\mathbb{R}^{m\times n}$. It turns out that with several restrictions, this is true.

Definition A.50 (Symmetric Gauge Function). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is a* symmetric gauge function *if it satisfies the following three conditions:*

- **norm:** *$f$ is a norm on $\mathbb{R}^n$;*
- **permutation invariance:** *For every $\boldsymbol{x} \in \mathbb{R}^n$ and permutation matrix $\boldsymbol{\Pi}$, $f(\boldsymbol{\Pi x}) = f(\boldsymbol{x})$;*
- **symmetry:** *For every $\boldsymbol{x} \in \mathbb{R}^n$ and diagonal sign matrix $\boldsymbol{\Sigma}$ (i.e., matrix with diagonal entries $\pm 1$), $f(\boldsymbol{\Sigma x}) = f(\boldsymbol{x})$.*

Theorem A.51 (Von Neumann's Characterization of Unitary Invariant Norms). *Fix $m \geq n$. For $\boldsymbol{M} \in \mathbb{C}^{m\times n}$, let $\boldsymbol{\sigma}(\boldsymbol{M}) \in \mathbb{R}^n$ denote its vector of singular values. Then for every symmetric gauge function $f_\sharp$,*

$$\|\boldsymbol{M}\|_\sharp \doteq f_\sharp(\boldsymbol{\sigma}(\boldsymbol{M})) \tag{A.9.25}$$

*defines a unitary invariant matrix norm on $\mathbb{C}^{m\times n}$. Conversely, for every unitary invariant matrix norm $\|\boldsymbol{M}\|_\flat$ there exists a symmetric gauge function $f_\flat$ such that $\|\boldsymbol{M}\|_\flat = f_\flat(\boldsymbol{\sigma}(\boldsymbol{M}))$.*

# Appendix B Convex Sets and Functions

The notion of convexity arises when we try to formalize the property that "good local decisions lead to globally optimal solutions." Consider a generic unconstrained optimization problem

$$\min \quad f(\boldsymbol{x}). \tag{B.0.1}$$

Here $\boldsymbol{x} \in \mathbb{R}^n$ is the variable of optimization, and $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function, which we are trying to make as small as possible using a numerical algorithm. Figure B.1 displays two objective functions $f$. The one on the right has many peaks and valleys – it may be very difficult to find the lowest valley, corresponding to the global optimum $\boldsymbol{x}_\star$. Moreover, for the function $f$ on the right, local information around a point $\boldsymbol{x}$ is not particularly helpful for determining what direction to move to reach the global optimum. In contrast, the bowl-shaped function on the left is much more amenable to global optimization – a "gradient descent" type algorithm, that simply determined which direction to move by considering the slope of the graph of the function, would easily "ski" down to the global minimum.

The notion of *convexity* formalizes this property. Convexity is a geometric property. It is convenient to first introduce the notion of a convex set, and then extend this definition to functions.

## B.1 Convex Sets

A set $\mathsf{C}$ is said to be *closed* if it contains its boundary. More precisely, for any converging sequence of points $\{\boldsymbol{x}_k\}$ in $\mathsf{C}$, we must have:

$$\boldsymbol{x}_k \to \bar{\boldsymbol{x}} \quad \Rightarrow \quad \bar{\boldsymbol{x}} \in \mathsf{C}.$$

A set $\mathsf{C} \subseteq \mathbb{R}^n$ is *convex* if for every pair of points $\boldsymbol{x}, \boldsymbol{x}' \in \mathsf{C}$, the line segment obtained by joining the two points also lies entirely in $\mathsf{C}$:

Definition B.1 (Convex Set). $\mathsf{C} \subseteq \mathbb{R}^n$ *is* convex *if*

$$\forall\, \boldsymbol{x},\, \boldsymbol{x}' \in \mathsf{C}, \quad \alpha \in [0,1], \qquad \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{x}' \in \mathsf{C}. \tag{B.1.1}$$

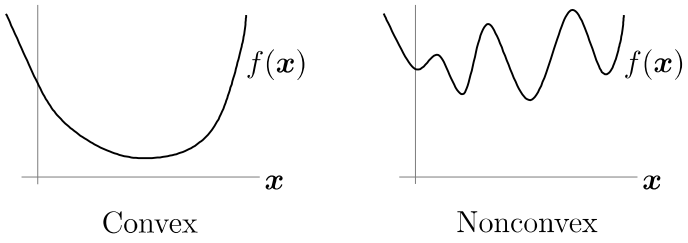


**Figure B.1 Two optimization problems** $\min f(\boldsymbol{x})$. The objective $f$ at left appears to be amenable to global optimization, while the one at right appears to be more challenging.

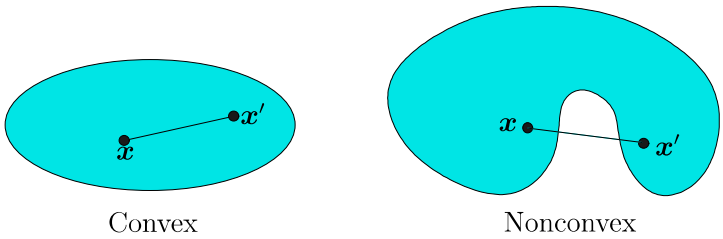


**Figure B.2 Convex and nonconvex sets.** A set is convex if we can select any pair of points $\boldsymbol{x}$, $\boldsymbol{x}'$ in the set, and the line segment joining them lies entirely within the set. The set to the left has this property, while the set to the right does not.

Figure B.2 gives an example of two sets, one of which is convex and one of which is not.

EXAMPLE B.2 (Convex sets). *Show that the following are convex:*

- *Every affine subspace.*
- *Every norm ball* $\mathsf{B}_{\|\cdot\|} = \{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leq 1\}$.
- *The empty set.*
- *Any intersection* $\mathsf{C} = \mathsf{C}_1 \cap \mathsf{C}_2$ *of two convex sets* $\mathsf{C}_1, \mathsf{C}_2$.

PROPOSITION B.3. *1 The intersection of a collection of convex sets* $\bigcap_i \mathsf{C}_i$ *is convex.*

*2 The image of a convex set under an affine transformation is convex.*

DEFINITION B.4 (Convex Hull). *The* convex hull *of any given set* $\mathsf{S}$ *is the minimal convex set containing* $\mathsf{S}$*, denoted as conv*$(\mathsf{S})$*. If* $\mathsf{S}$ *contains a finite number of* $\mathsf{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ *points, we have*

$$\operatorname{conv}(\mathsf{S}) \doteq \left\{ \sum_{i=1}^n \alpha_i \boldsymbol{x}_i \;\middle|\; \forall \alpha_i \geq 0 \text{ with } \sum_{i=1}^n \alpha_i = 1. \right\}. \tag{B.1.2}$$

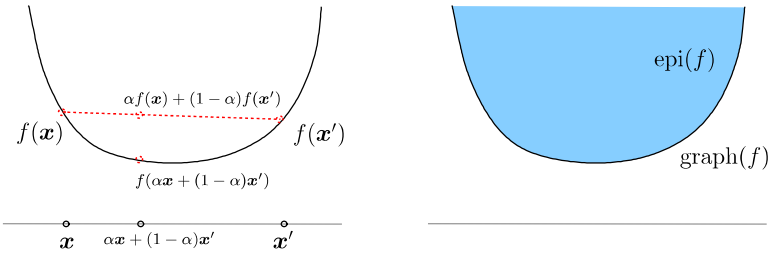


**Figure B.3 Convexity of functions**: a function $f$ is convex if its epigraph $\text{epi}(f) = \{(\boldsymbol{x}, t) \mid t \geq f(\boldsymbol{x})\}$ is a convex set (right). This is true if and only if for every pair of points $\boldsymbol{x}$, $\boldsymbol{x}'$ and scalar $\alpha \in [0,1]$, $f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{x}') \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}')$. The picture at right illustrates this inequality: the segment joining $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{x}', f(\boldsymbol{x}'))$ lies above the graph of $f$.

## B.2 Convex Functions

For a function $f : \mathcal{D} \to \mathbb{R}$ defined on a (convex) domain $\mathcal{D} \subseteq \mathbb{R}^n$, its *graph* is the set of pairs $(\boldsymbol{x}, f(\boldsymbol{x}))$ that can be generated by evaluating the function $f$ at every point:

$$\text{graph}(f) \doteq \{(\boldsymbol{x}, f(\boldsymbol{x})) \mid \boldsymbol{x} \in \mathcal{D}, f(\boldsymbol{x}) < +\infty\} \subseteq \mathbb{R}^{n+1}. \tag{B.2.1}$$

We give another name to everything that lies above the graph: the *epigraph*:

$$\text{epi}(f) \doteq \{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathcal{D}, t \in \mathbb{R}, f(\boldsymbol{x}) \leq t\} \subseteq \mathbb{R}^{n+1}. \tag{B.2.2}$$

We say that $f$ is a *convex function* if its epigraph is a convex set. Figure B.3 (right) illustrates this property. Figure B.3 (left) suggests an equivalent definition, which is sometimes easier to work with: $f$ is convex if for any pair of points $\boldsymbol{x}$ and $\boldsymbol{x}'$, the line segment joining $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{x}', f(\boldsymbol{x}'))$ lies entirely above the graph of $f$:

Definition B.5 (Convex Function). *A function $f : \mathcal{D} \to \mathbb{R}$ is* convex *if for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ and $\alpha \in [0,1]$,*

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}') \geq f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{x}'). \tag{B.2.3}$$

Notice that above definitions do not require $f$ to be differentiable. If $f$ is differentiable, the notion of convexity can be characterized in terms of its derivatives. Since the epigraph is convex, then the tangent plane at each point of the graph should lie beneath the graph. The following statement makes this precise:

Proposition B.6 (First-Order Condition). *Let $f : \mathcal{D} \to \mathbb{R}$ be differentiable. Then $f$ is convex if and only if it satisfies the condition:*

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^*(\boldsymbol{x}' - \boldsymbol{x})$$

*for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$.*

This is precisely the geometry of the "nice" function in Figure B.1 (left). From this picture, it is clear that convexity is very favorable for global optimization.[1] There also exist nonconvex functions that are easy to optimize – Chapter 7 provides a brief introduction to this emerging literature. However, if we want to talk about a class of functions, rather than a particular one, then there is a very beautiful motivation for studying convex functions. To appreciate this motivation, we need to first observe a useful fact: if $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are convex functions, then for any $\alpha, \beta \geq 0$, $h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})$ is also convex. If we let $\mathcal{F}$ be the largest class of continuously differentiable functions that satisfy the following three demands:

- every linear function $\phi(\boldsymbol{x}) = \boldsymbol{a}^* \boldsymbol{x} + b$ is in $\mathcal{F}$;
- every nonnegative combination $\alpha f_1(\boldsymbol{x}) + \beta f_2(\boldsymbol{x})$ of $f_1, f_2 \in \mathcal{F}$ is in $\mathcal{F}$;
- for every $f \in \mathcal{F}$, the stationarity condition $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$ implies that $\boldsymbol{x}_\star$ is a global optimizer of $f$,

then it turns out that the $\mathcal{F}$ is precisely the class of **convex**, continuously differentiable functions. You can interpret this as suggesting that for *global* solutions, convex functions really are the right general class of functions to study. For more details, see the book of Nesterov [Nes03].

You may also notice that in Figure B.3, the function $f(\boldsymbol{x})$ "curves upward": its second derivative is nonnegative at every point of the domain. For twice differentiable functions, this leads to a simpler condition for convexity: the function is convex if and only if its second derivative at any point, and in any direction is positive. The following makes this precise:

PROPOSITION B.7 (Second-Order Conditions). *Let $f : \mathcal{D} \to \mathbb{R}$ be twice differentiable. Then $f$ is convex if and only if its Hessian is positive semidefinite:*

$$\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}$$

*for all $\boldsymbol{x} \in \mathcal{D}$.*

The class of convex functions includes important examples such as linear functions and norms:

EXAMPLE B.8 (Convex Functions). *Show that the following are convex functions:*

- *Every affine function $f(\boldsymbol{x}) = \boldsymbol{a}^* \boldsymbol{x} + b$.*
- *Every norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$.*
- *Every semidefinite quadratic $f(\boldsymbol{x}) = \boldsymbol{x}^* \boldsymbol{P} \boldsymbol{x}$, with $\boldsymbol{P} \succeq \boldsymbol{0}$.*

Before continuing, we note one nice property of convex functions which will be useful for deriving an appropriate tractable replacement for the $\ell^0$ norm.

[1] Once you've internalized the definition a bit, you may begin to wonder to what extent the implication "convexity $\implies$ easy-to-optimize" is actually true. The convex functions that we encounter in this book will all possess special structure that makes them very amenable to efficient algorithms. However, this is not true of all convex functions – there exist convex functions that are NP-hard to optimize.

Definition B.9 (Convex Combination). *A convex combination of a set of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ is an expression of the form $\lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_k \boldsymbol{x}_k$, with $\lambda_i \geq 0$ for each $i$ and $\sum_i \lambda_i = 1$.*

Lemma B.10 (Jensen's Inequality). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. For any $k$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$, $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_+$, with $\sum_i \lambda_i = 1$,*

$$f\left(\sum_i \lambda_i \boldsymbol{x}_i\right) \leq \sum_i \lambda_i f(\boldsymbol{x}_i). \tag{B.2.4}$$

*Proof* The proof is by induction on $k$. For $k = 1$, there is nothing to show. Now suppose the claim is true for $1, \ldots, k-1$. Then

$$f\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i\right) \leq \left(\sum_{i=1}^{k-1} \lambda_i\right) f\left(\frac{\sum_{i=1}^{k-1} \lambda_i \boldsymbol{x}_i}{\sum_{i=1}^{k-1} \lambda_i}\right) + \lambda_k f(\boldsymbol{x}_k) \tag{B.2.5}$$

$$\leq \sum_{i=1}^{k} \lambda_i f(\boldsymbol{x}_i) \tag{B.2.6}$$

as desired. Above, the first step uses the definition of convexity, and the second uses the inductive hypothesis. □

With this lemma, it is easy to show that any $\alpha$-sublevel set of a convex function $f : \mathcal{D} \to \mathbb{R}$:

$$\mathsf{C}_\alpha = \{\boldsymbol{x} \in \mathcal{D} \mid f(\boldsymbol{x}) \leq \alpha\} \tag{B.2.7}$$

is a convex set. However, a function with all its sublevel sets being convex is not necessarily a convex function![2] A function is said to be a *closed* function, if each sublevel set is a closed set. We typically only consider closed convex functions, unless otherwise stated.

Proposition B.11. *We can use convex functions to generate other associated convex functions:*

1. *A function is convex if and only if it is convex when restricted to any line that intersects its domain.*
2. *A weighted sum of convex functions with nonnegative weights is convex.*
3. *If $f, g$ are convex functions and $g$ is non-decreasing in its univariate domain, then $h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ is convex.*
4. *Given a collection of convex functions $f_\alpha : \mathcal{D} \to \mathbb{R}$, $\alpha \in \mathsf{A}$, their point-wise supremum*

$$f(\boldsymbol{x}) \doteq \sup_{\alpha \in \mathsf{A}} f_\alpha(\boldsymbol{x})$$

*is also convex.*

Example B.12. *The maximal eigenvalue of a symmetric matrix is a (closed) convex function.*

[2] Such functions are called *quasi-convex*. Please find an example for yourself.

*Proof* To see that, the maximal eigenvalue function can be written as

$$\lambda_{\max}(\boldsymbol{X}) = \sup\{\boldsymbol{y}^* \boldsymbol{X} \boldsymbol{y}\}, \quad \|\boldsymbol{y}\|_2 = 1.$$

Since the function is the point-wise supremum of a set of linear functions with respect to $\boldsymbol{X}$, it is a convex function. □

*Convex Envelop and Conjugate.*
For any non-convex (closed) function $g : \mathcal{D} \to \mathbb{R}$ defined on a convex domain $\mathcal{D}$, it has a naturally associated convex function that bounds it from below:

Definition B.13 (Convex Envelope). *The convex envelope of a closed function $g$ is defined as*

$$\text{conv}g(\boldsymbol{x}) = \sup\{h(\boldsymbol{x}) \mid h(\boldsymbol{x}) \text{ convex } \& \ h(\boldsymbol{x}) \leq g(\boldsymbol{x}) \ \ \forall \boldsymbol{x} \in \mathcal{D}\}. \tag{B.2.8}$$

Let us define the (Fenchel) *conjugate* of a function $g(\boldsymbol{x})$ (not necessarily convex) as:

$$g^*(\boldsymbol{\lambda}) = \sup_{\boldsymbol{x}} \boldsymbol{\lambda}^* \boldsymbol{x} - g(\boldsymbol{x}). \tag{B.2.9}$$

The conjugate of a function $g$ is essentially the negated dual function of $g$ that we often see in the method of Largrange multipliers (see Section C.3).

Proposition B.14. *Assuming the conjugate is well-defined, we have the following:*

*1 The conjugate $g^*(\boldsymbol{\lambda})$ is always a convex function.*
*2 $g^{**}(\boldsymbol{x}) = \text{conv}g(\boldsymbol{x})$.*

*Strong Convexity.*
In this book, we sometimes are interested in stronger notion of convexity.

Definition B.15 (Strongly Convex Function). *A function $f : \mathcal{D} \to \mathbb{R}$ is* strongly convex *if $f$ is convex and for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ and $\alpha \in [0, 1]$,*

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}') \geq f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}') + \mu \frac{\alpha(1-\alpha)}{2} \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 \tag{B.2.10}$$

*for some $\mu > 0$.*

Notice that the above definition does not require $f$ to be differentiable. If $f$ is first or second-order differentiable, we have the following sufficient conditions for $f$ being strongly convex.

Proposition B.16. *For a differentiable convex function $f$ over $\mathcal{D}$, we have $f$ is strongly convex if either of the following conditions hold:*

*1 $f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^* (\boldsymbol{x}' - \boldsymbol{x}) + \mu \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2, \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$;*
*2 $\nabla^2 f(\boldsymbol{x}) \succeq \mu \cdot \boldsymbol{I}, \ \forall \boldsymbol{x} \in \mathcal{D}$;*

*for some $\mu > 0$.*

However, as we see in Section 3.3.2, we are interested in strong convexity in a restricted sense.

*Lipschitz Continuous Gradients.*
The functions we encounter in many optimization problems are often "smooth" in their landscape in the sense that their gradients do not vary so dramatically. One way to characterize such smoothness is the notion of Lipschitz continuous gradients.

DEFINITION B.17 (Lipschitz Continuous Gradient). *A differentiable function $f : \mathcal{D} \to \mathbb{R}$ has $L$-*Lipschitz continuous gradients *if $\nabla f(\boldsymbol{x})$ satisfies*

$$\|\nabla f(\boldsymbol{x}') - \nabla f(\boldsymbol{x})\|_2 \leq L\|\boldsymbol{x}' - \boldsymbol{x}\|_2, \quad \forall \boldsymbol{x}', \boldsymbol{x} \in \mathcal{D}, \tag{B.2.11}$$

*for some constant $L > 0$. The constant $L$ is called the Lipschitz constant of $\nabla f$.*

When the function $f$ is twice differentiable, then it is not difficult to prove from fundamental theorems of calculus (also see proof of Lemma 8.2) that $f$ has $L$-Lipschitz continuous gradients (over the domain $\mathcal{D}$) if we have

$$\|\nabla^2 f(\boldsymbol{x})\| \leq L, \quad \forall \boldsymbol{x} \in \mathcal{D}. \tag{B.2.12}$$

As we will see, when a convex function $f$ over a domain $\mathcal{D}$ is both strongly convex and smooth (in the sense of having Lipschitz continuous gradients), then it can be efficiently minimized over $\mathcal{D}$ by a simple gradient descent algorithm of the type:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k), \tag{B.2.13}$$

where the step size $t_k$ can be chosen to be between $\frac{1}{L}$ and $\frac{2}{L+\mu}$. Somewhat surprisingly, one can easily show (see Theorem D.4) that such a vanilla algorithm enjoys $\ell^2$ error contraction around the (global) minimum $\boldsymbol{x}_\star$:

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_2 \leq \rho \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \tag{B.2.14}$$

for some $\rho \leq 1 - \frac{\mu}{L} < 1$. That is the estimate error drops exponentially with the number of iterations.

## B.3 Subdifferentials of Nonsmooth Convex Functions

For smooth, convex functions $f$, the local information encoded in the gradient $\nabla f$ and Hessian $\nabla^2 f$ characterize both the local and global behavior of $f$, allowing us to give optimality conditions and construct minimization algorithms. Familiar, classical algorithms such as gradient ascent, Newton's method, and their variants, are all constructed using differential information. Moreover, as we saw in the previous section, these quantities play a critical role in characterizing convexity for smooth functions $f$.

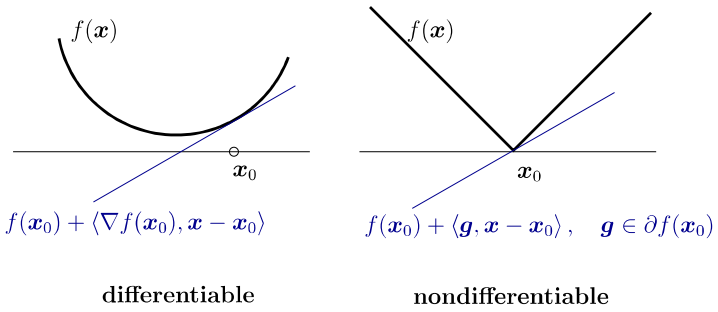


**Figure B.4** Differential and subdifferentials of convex functions.

It is a curious fact, then, that many of the most useful convex objective functions arising in high-dimensional data analysis are nondifferentiable: *their gradients and Hessians do not exist.* For example, the $\ell^1$ norm $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$ is nondifferentiable at any point $\boldsymbol{x} \in \mathbb{R}^n$ with fewer than $n$ nonzero entries. These are precisely the points that we care about for sparse estimation! This nonsmooth behavior is actually desirable from the statistical perspective. However, it forces us to make recourse to analytical tools that are general enough to handle nondifferentiable functions. Fortunately, for convex functions, the nondifferentiable theory rests on simple, geometrically intuitive ideas, which we describe in this section. For accessible introductions to the general theory of convexity, we recommend [Nem95, Nem07, Nes03, BV04].

The most important notion is that of a *subgradient* of a convex function, which provides a very satisfactory replacement for the gradient, when the function is not differentiable. Recall from Proposition B.6 that for convex, *differentiable* $f$,

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle\,, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}. \tag{B.3.1}$$

This inequality has a simple geometric interpretation, which we visualize in Figure B.4. We visualize the graph of the function $f : \mathbb{R}^n \to \mathbb{R}$. The graph is the collection of points of the form $(\boldsymbol{x}, f(\boldsymbol{x})) \in \mathbb{R}^{n+1}$. The graph of

$$h(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

is a hyperplane, which is tangent to the graph of $\boldsymbol{f}$ at $(\boldsymbol{x}, f(\boldsymbol{x}))$. The inequality (B.3.1) says that at all points $\boldsymbol{y}$ in the domain of the function $f$ this tangent hyperplane lies below (or more precisely, not above) the graph of $f$.

Figure B.4 (right) visualizes the graph of another convex function $f$, which is not differentiable at point $\boldsymbol{x}$. The gradient of $f$ *does not* exist at $\boldsymbol{x}$. Nevertheless, we can still define a nonvertical hyperplane $\mathcal{H} \subseteq \mathbb{R}^{n+1}$ that passes through $(\boldsymbol{x}, f(\boldsymbol{x}))$, and lies below the graph of $f$. This hyperplane has normal vector $(\boldsymbol{v}, -1)$, and can be expressed in notation as

$$\mathcal{H} = \{(\boldsymbol{y}, t) \mid t = f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle\}\,. \tag{B.3.2}$$

We say that $\boldsymbol{v} \in \mathbb{R}^n$ is a *subgradient* of $f$ at $\boldsymbol{x}$ if it defines a hyperplane that supports the graph of $f$ at $\boldsymbol{x}$, and lies below the graph everywhere:

Definition B.18 (Subgradient). *Let $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. A vector $\boldsymbol{v}$ is a* subgradient *of $f$ at $\boldsymbol{x} \in \mathcal{D}$ if for all $\boldsymbol{y} \in \mathcal{D}$,*

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle \,. \tag{B.3.3}$$

When $f$ is differentiable, from Proposition B.6 it is clear that $\boldsymbol{v} = \nabla f(\boldsymbol{x})$ satisfies (B.3.3). When $f$ is *nondifferentiable*, at a given point $\boldsymbol{x}$ there can be multiple distinct hyperplanes that support the graph of $f$, and hence, there can be multiple subgradients $\boldsymbol{v}$ (see Figure B.4). The collection of all subgradients is called the *subdifferential* of $f$ at $\boldsymbol{x}$, and is denoted $\partial f(\boldsymbol{x})$. Formally:

Definition B.19 (Subdifferential). *Let $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The* subdifferential *$\partial f(\boldsymbol{x})$ is the collection of all subgradients of $f$ at $\boldsymbol{x}$:*

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{v} \mid f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) \,+\, \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle \,,\; \forall \boldsymbol{y} \in \mathcal{D}\} \,. \tag{B.3.4}$$

Notice that if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\boldsymbol{x}$, its subdifferential at $\boldsymbol{x}$ is a singleton: $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. This coincides with the classical definition of differentials.

A number of functions of interest have relatively simple subdifferentials.

Example B.20. *As good exercises, the reader may try to verify the subdifferentials for the following functions:*

1 *The subdifferential for $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ with $\boldsymbol{x} \in \mathbb{R}^n$.*
2 *The subdifferential for $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$ with $\boldsymbol{x} \in \mathbb{R}^n$.*
3 *The subdifferential for $f(\boldsymbol{X}) = \sum_{j=1}^{n} \|\boldsymbol{X}\boldsymbol{e}_j\|_2$ with $\boldsymbol{X}$ a matrix in $\mathbb{R}^{n \times n}$.*
4 *The subdifferential for $f(\boldsymbol{x}) = \|\boldsymbol{X}\|_*$ with $\boldsymbol{X}$ a matrix in $\mathbb{R}^{n \times n}$.*

Below are some basic properties of subdifferentials.

Lemma B.21 (Monotonicity Property). *Given a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and any $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^n$ such that $\boldsymbol{v} \in \partial f(\boldsymbol{x})$ and $\boldsymbol{v}' \in \partial f(\boldsymbol{x}')$, we have*

$$\langle \boldsymbol{x} - \boldsymbol{x}', \boldsymbol{v} - \boldsymbol{v}' \rangle \geq 0. \tag{B.3.5}$$

*Proof* From the definition of subgradient (B.19), we have

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x} \rangle, \quad f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \langle \boldsymbol{v}', \boldsymbol{x} - \boldsymbol{x}' \rangle. \tag{B.3.6}$$

Adding these two inequalities together we obtain:

$$f(\boldsymbol{x}) + f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + f(\boldsymbol{x}') + \langle \boldsymbol{v} - \boldsymbol{v}', \boldsymbol{x}' - \boldsymbol{x} \rangle. \tag{B.3.7}$$

Canceling $f(\boldsymbol{x}) + f(\boldsymbol{x}')$ from both sides obtains the desired result. □

LEMMA B.22. *If a convex function $f(\boldsymbol{x})$ has Lipschitz continuous gradients with constant $L$, then for any $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we have:*

$$\langle \nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \geq \frac{1}{L} \|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\|_2^2 \geq 0. \tag{B.3.8}$$

*Proof* Let us define a function $h(\boldsymbol{z}) \doteq f(\boldsymbol{z}) - \boldsymbol{z}^* \nabla f(\boldsymbol{x})$. Then $h(\boldsymbol{z})$ is convex and is minimized at $\boldsymbol{z} = \boldsymbol{x}$ (as $\nabla h(\boldsymbol{x}) = \boldsymbol{0}$). Hence for any $\boldsymbol{z}$, we have

$$h(\boldsymbol{x}) \leq h\big(\boldsymbol{z} - \frac{1}{L}\nabla h(\boldsymbol{z})\big) \leq h(\boldsymbol{z}) + \langle \nabla h(\boldsymbol{z}), -\frac{1}{L}\nabla h(\boldsymbol{z}) \rangle + \frac{L}{2}\|\frac{1}{L}\nabla h(\boldsymbol{z})\|_2^2.$$

The last inequality comes from the fact that the function $f(\boldsymbol{x})$ (and hence $h(\boldsymbol{z})$) has Lipschitz continuous gradients with constant $L$. This gives

$$h(\boldsymbol{x}) \leq h(\boldsymbol{z}) - \frac{1}{2L}\|\nabla h(\boldsymbol{z})\|_2^2. \tag{B.3.9}$$

Now applying the inequality to $\boldsymbol{x} = \boldsymbol{x}_1, \boldsymbol{z} = \boldsymbol{x}_2$ as well as the reverse case $\boldsymbol{x} = \boldsymbol{x}_2, \boldsymbol{z} = \boldsymbol{x}_1$, we get

$$f(\boldsymbol{x}_1) - \boldsymbol{x}_1^* \nabla f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_2) - \boldsymbol{x}_2^* \nabla f(\boldsymbol{x}_1) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2,$$
$$f(\boldsymbol{x}_2) - \boldsymbol{x}_2^* \nabla f(\boldsymbol{x}_2) \leq f(\boldsymbol{x}_1) - \boldsymbol{x}_1^* \nabla f(\boldsymbol{x}_2) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\|_2^2.$$

Adding these two together gives the desired bound (B.3.8). □

# Appendix C Optimization Problems and Optimality Conditions

*"Since the fabric of the universe is most perfect and the work of a most wise Creator, nothing at all takes place in the universe in which some rule of maximum or minimum does not appear."*

– Leonhard Euler

## C.1 Unconstrained Optimization

The mathematical model of an (unconstrained) optimization problem can be generally described by a domain or constraint set $\mathcal{D}$ in $\mathbb{R}^n$ and an objective function $f : \mathcal{D} \to \mathbb{R}$ that maps an element of $\mathcal{D}$ to a real value. The optimization problem seeks an optimal solution $\boldsymbol{x}_\star \in \mathcal{D}$ such that the value of $f$ is minimized:

$$f(\boldsymbol{x}_\star) \leq f(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \mathcal{D}.$$

In particular, if $\mathcal{D} = \mathbb{R}^n$, it is called an unconstrained optimization problem.

Definition C.1 (Local and Global Minima). *A variable $\boldsymbol{x}_\star$ is a local minimum of $f$ if there exists a neighborhood $\mathsf{B}(\varepsilon, \boldsymbol{x}_\star) \doteq \{\boldsymbol{x} \in \mathcal{D} \mid \|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 < \varepsilon\}$ for some $\varepsilon > 0$ such that*

$$f(\boldsymbol{x}_\star) \leq f(\boldsymbol{x}), \quad \textit{for all } \boldsymbol{x} \in \mathsf{B}(\varepsilon, \boldsymbol{x}_\star).$$

*The variable $\boldsymbol{x}_\star$ is a global minimum of $f$ if $\mathsf{B}(\varepsilon, \boldsymbol{x}_\star) = \mathcal{D}$. The above local and global minima are said to be strict if the corresponding inequalities are also strict for $\boldsymbol{x} \neq \boldsymbol{x}_\star$.*

If the objective function $f$ is differentiable, then conditions for the optimality can be expressed in terms of its derivatives. In particular, if $\boldsymbol{x}_\star$ is a local minimum, then within a small neighborhood $\mathsf{B}(\varepsilon, \boldsymbol{x}_\star)$, for any given vector $\boldsymbol{v} \in \mathbb{R}^n$, we have

$$f(\boldsymbol{x}_\star + t \cdot \boldsymbol{v}) \geq f(\boldsymbol{x}_\star)$$

for sufficiently small $t > 0$ such that $t \cdot \boldsymbol{v} \in \mathsf{B}(\varepsilon, \boldsymbol{0})$. Hence we have

$$\lim_{t \to 0} \frac{f(\boldsymbol{x}_\star + t \cdot \boldsymbol{v}) - f(\boldsymbol{x}_\star)}{t} = \nabla f(\boldsymbol{x}_\star)^* \boldsymbol{v} \geq 0.$$

Notice that this must be true for both $\boldsymbol{v}$ and $-\boldsymbol{v}$. Then for the inequality to hold for all $\boldsymbol{v} \in \mathbb{R}^n$, we must have

$$\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}. \tag{C.1.1}$$

Definition C.2 (Stationary Point or Critical Point). *A point $\boldsymbol{x}_\star$ that satisfies the condition $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$ is referred to as a stationary point of $f(\boldsymbol{x})$. A stationary point is also known as a critical point.*

If $f$ is twice continuously differentiable and $\boldsymbol{x}_\star$ is a stationary point with $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$, we have:

$$f(\boldsymbol{x}_\star + t \cdot \boldsymbol{v}) \approx f(\boldsymbol{x}_\star) + \frac{1}{2}\boldsymbol{v}^* \nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{v}t^2 + o(t^2).$$

If $\boldsymbol{x}_\star$ is a local minimum, we have

$$f(\boldsymbol{x}_\star + t \cdot \boldsymbol{v}) - f(\boldsymbol{x}_\star) \geq 0 \quad \Rightarrow \quad \frac{1}{2}\boldsymbol{v}^* \nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{v}t^2 \geq 0$$

for all $\boldsymbol{v} \in \mathbb{R}^n$. This implies the matrix $\nabla^2 f(\boldsymbol{x}_\star)$ is necessarily positive semidefinite, namely,

$$\nabla^2 f(\boldsymbol{x}_\star) \succeq \boldsymbol{0}. \tag{C.1.2}$$

A stationary point satisfying the above condition is also called a *second-order* stationary point.

It is then not difficult to show the following sufficient condition for local minima:

Proposition C.3 (Second-Order Sufficient Optimality Condition). *Let $f : \mathcal{D} \to \mathbb{R}$ be twice continuously differentiable. If $\boldsymbol{x}_\star$ satisfies the conditions*

$$\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0} \quad \textit{and} \quad \nabla^2 f(\boldsymbol{x}_\star) \succ \boldsymbol{0},$$

*Then $\boldsymbol{x}_\star$ is a strict local minimum of $f(\boldsymbol{x})$.*

In general, a local minimum is not necessarily a global minimum in the domain of $f(\boldsymbol{x})$. Therefore, the global minimum can be found by exhaustively comparing the values of $f$ at all local minima. However, when the objective function $f$ is convex, the following proposition shows that any local minimum is also a global minimum.

Proposition C.4 (Global Optimality of Convex Functions). *Let $f : \mathcal{D} \to \mathbb{R}$ be a convex function over convex set $\mathcal{D}$. Then*

1 *A local minimum of $f$ is also a global minimum. Furthermore, if $f$ is strictly convex, then the global minimum, if it exists, is unique.*
2 *A point $\boldsymbol{x}_\star \in \mathcal{D}$ is a global minimum of $f$ if $\boldsymbol{0} \in \partial f(\boldsymbol{x}_\star)$. In the case that $f$ is differentiable, $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$ implies that $\boldsymbol{x}_\star$ is a global minimum.*

Finally, we note that given an objective function $f$, a local minimum need not exist. For example, the simple scalar function $f(x) = x$ does not have a minimal value in the domain of real numbers as $\inf_{x\in\mathbb{R}} f(x) = -\infty$. Therefore, a sufficient condition for $f$ to have at least one local minimum is that the set $\{f(\boldsymbol{x})|\boldsymbol{x} \in \mathcal{D}\}$ is bounded below. Alternatively, according to the Weierstrass theorem, if $f$ is continuous and the domain set $\mathcal{D} \subseteq \mathbb{R}^n$ is compact (i.e.closed and bounded), then $f$ has at least one local minimum.

## C.2 Constrained Optimization

In the previous section, the constraint set of the optimization problems is assumed to be any general set. However, in most optimization problems considered in this book, the constraints are formulated as equality or inequality conditions. For example, the domain $\mathcal{D} \subset \mathbb{R}^n$ of a polyhedron can be specified by a set of equality and inequality conditions. *Lagrange multipliers* are a set of supportive variables to facilitate the derivation of optimality conditions for such constrained optimization problems. Arguably, Lagrange multiplier theory is the most influential theory in constrained optimization. In duality theory that we will discuss in the next section, the same Lagrange multiplier variables are also called *dual variables*, which will play a central role as the optimization variables of the *dual problems*.

First, we consider the optimization problem with equality constraints:

$$\min f(\boldsymbol{x}) \quad \text{subject to} \quad h_i(\boldsymbol{x}) = 0,\ i = 1, \dots, m, \tag{C.2.1}$$

where $f$ and each $h_i$ are assumed to be continuously differentiable.[1] Conveniently, we further assume the gradients of the equality conditions at any feasible solution $\boldsymbol{x}'$ (that satisfies the equality constraints)

$$\nabla h_1(\boldsymbol{x}'), \nabla h_2(\boldsymbol{x}'), \dots, \nabla h_m(\boldsymbol{x}')$$

are linearly independent. Such a solution $\boldsymbol{x}'$ is also called *regular*.

The optimality conditions for (C.2.1) can be conveniently derived in terms of the Lagrangian function $\mathcal{L} : \mathbb{R}^{n+m} \to \mathbb{R}$ as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \doteq f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{h}(\boldsymbol{x}) \rangle, \tag{C.2.2}$$

where $\lambda_i$ are the Lagrange multipliers for the equality conditions, and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_m]^* \in \mathbb{R}^m$ is the corresponding Lagrange multiplier vector; and for brevity, we denote $\boldsymbol{h} = [h_1, h_2, \dots, h_m]^*$ as a map from $\mathbb{R}^n$ to $\mathbb{R}^m$.

The basic Lagrange multiplier theory states the following necessary condition for the optimality of a regular solution.

[1] In the main text, we need to generalize to cases when $f$ is not differentiable.

PROPOSITION C.5 (Necessary Conditions for Optimality). *Let $\boldsymbol{x}_\star$ be a local minimum of function $f(\boldsymbol{x})$ subject to $h_i(\boldsymbol{x}) = 0$, $i = 1, \ldots, m$. Further assume $\boldsymbol{x}_\star$ is regular. Then there exists a Lagrange multiplier vector $\boldsymbol{\lambda}_\star = (\lambda_{\star,1}, \lambda_{\star,2}, \ldots, \lambda_{\star,m}) \in \mathbb{R}^m$, such that*

$$\begin{array}{rcl} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) & = & \nabla f(\boldsymbol{x}_\star) + \sum_{i=1}^m \lambda_{\star,i} \nabla h_i(\boldsymbol{x}_\star) = \boldsymbol{0}, \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) & = & \boldsymbol{h}(\boldsymbol{x}_\star) = \boldsymbol{0}. \end{array} \tag{C.2.3}$$

*Furthermore, if $f$ and $\boldsymbol{h}$ are twice continuously differentiable, we have*

$$\begin{array}{rcll} \boldsymbol{v}^* \nabla^2_{\boldsymbol{xx}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) \boldsymbol{v} & = & \boldsymbol{v}^* \Big( \nabla^2 f(\boldsymbol{x}_\star) + \sum_{i=1}^m \lambda_{\star,i} \nabla^2 h_i(\boldsymbol{x}_\star) \Big) \boldsymbol{v} & \\ & \geq & 0, \quad \forall \boldsymbol{v} : \boldsymbol{v}^* \nabla h_i(\boldsymbol{x}_\star) = 0, \ i = 1, \ldots, m. & \end{array} \tag{C.2.4}$$

In (C.2.4), the conditions for vector $\boldsymbol{v} \in \mathbb{R}^n$ that satisfies $\boldsymbol{v}^* \nabla h_i(\boldsymbol{x}_\star) = 0$ can be understood as follows. If we consider a new point $\boldsymbol{x}' = \boldsymbol{x}_\star + t \cdot \boldsymbol{v}$ for some small $t \in \mathbb{R}$, due to the fact that $\boldsymbol{v}^* \nabla h_i(\boldsymbol{x}_\star) = 0$, a small variation along $\boldsymbol{v}$ will not change the value of $\boldsymbol{h}(\boldsymbol{x}') \approx \boldsymbol{0}$. Therefore, we can define

$$\mathsf{V}(\boldsymbol{x}_\star) = \{ \boldsymbol{v} \mid \boldsymbol{v}^* \nabla h_i(\boldsymbol{x}_\star) = 0, \ i = 1, \ldots, m \}. \tag{C.2.5}$$

as the *subspace of first-order feasible variations*.

In summary, the first-order condition (C.2.3) implies the gradient $\nabla f(\boldsymbol{x}_\star)$ is orthogonal to $\mathsf{V}(\boldsymbol{x}_\star)$, which resembles the first-order condition $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$ in unconstrained optimization. The second-order condition (C.2.4) implies the Hessian of the Lagrangian function $\mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star)$ is positive semidefinite when constrained in $\mathsf{V}(\boldsymbol{x}_\star)$.

PROPOSITION C.6 (Sufficient Conditions). *Assume $f$ and $\boldsymbol{h}$ are twice continuously differentiable. Let $(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) \in \mathbb{R}^{n+m}$ satisfy*

$$\begin{array}{rcll} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) & = & \boldsymbol{0}, & \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) & = & \boldsymbol{0}, & \\ \boldsymbol{v}^* \nabla^2_{\boldsymbol{xx}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star) \boldsymbol{v} & > & 0, \quad \forall \boldsymbol{v} \in \mathsf{V}(\boldsymbol{x}_\star), \boldsymbol{v} \neq \boldsymbol{0}. & \end{array} \tag{C.2.6}$$

*Then $\boldsymbol{x}_\star$ is a strict local minimum of $f(\boldsymbol{x})$ subject to $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$.*

## C.3 Basic Duality Theory

Recall the Lagrangian function for the above equality-constrained optimization problem:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \doteq f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i h_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{h}(\boldsymbol{x}) \rangle, \tag{C.3.1}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_m]^* \in \mathbb{R}^m$ are the Lagrangian multipliers.

In duality theory, the vector $\boldsymbol{\lambda}$ is also called the *dual variables* for the so-called *dual function*:

$$q(\boldsymbol{\lambda}) \doteq \inf_{\boldsymbol{x} \in \mathcal{D}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{C.3.2}$$

Correspondingly, $f(\boldsymbol{x})$ is referred to as the *primal function* and $\boldsymbol{x}$ the *primal variables*.

A simple property of the dual function $q$ is that it is a concave function regardless whether the primal problem is convex or not, since $q$ is the point-wise infimum of a family of affine functions with respect to $(\boldsymbol{\lambda})$.

Another important property of the dual function is that $q(\boldsymbol{\lambda})$ is a lower bound of $f(\boldsymbol{x}')$ for any feasible solution $\boldsymbol{x}'$. In particular, $q(\boldsymbol{\lambda})$ is a lower bound of the optimal value $f(\boldsymbol{x}_\star)$. This can be easily verified since for a feasible $\boldsymbol{x}'$ satisfying $\boldsymbol{h}(\boldsymbol{x}') = \boldsymbol{0}$, we have

$$q(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{h}(\boldsymbol{x})\rangle \leq \inf_{\boldsymbol{x}\in\mathcal{D}, \boldsymbol{h}(\boldsymbol{x})=\boldsymbol{0}} f(\boldsymbol{x}) \leq f(\boldsymbol{x}').$$

For the dual function $q(\boldsymbol{\lambda})$ to provide a meaningful lower bound for $f(\boldsymbol{x}_\star)$, it is natural to avoid trivial cases when $q(\boldsymbol{\lambda}) = -\infty$. So we normally restrict the domain of the dual function $q$ to:

$$\mathcal{C} \doteq \{\boldsymbol{\lambda} \mid q(\boldsymbol{\lambda}) > -\infty\}. \tag{C.3.3}$$

More specifically, the dual variables $(\boldsymbol{\lambda})$ that satisfy above conditions are called *dual feasible solutions*.

A very useful concept in duality theory is the so-called *duality gap* between the primal and dual functions

$$f(\boldsymbol{x}) - q(\boldsymbol{\lambda}). \tag{C.3.4}$$

Since the dual function $q(\boldsymbol{\lambda})$ is a lower bound of the primal function $f(\boldsymbol{x})$, in particular of its minimal value $f(\boldsymbol{x}_\star)$. The duality gap is always nonnegative (over the set of feasible solutions). More importantly, when the duality gap is zero, namely, there exists a feasible solution $\boldsymbol{x}_\star$ and $\boldsymbol{\lambda}_\star$ such that $f(\boldsymbol{x}_\star) = q(\boldsymbol{\lambda}_\star)$, then $\boldsymbol{x}_\star$ is the optimal primal solution and $\boldsymbol{\lambda}_\star$ is the optimal dual solution.

Naturally, when we want to achieve the best lower-bound estimation of the minimal value, we can consider the following optimization problem in the dual space:

$$\max_{\boldsymbol{\lambda}} q(\boldsymbol{\lambda}). \tag{C.3.5}$$

The problem (C.3.5) is called the *Lagrange dual problem* associated with the original *primal problem* (C.2.1).

Since the optimal solution $q(\boldsymbol{\lambda}_\star)$ is the best lower-bound approximation of the global minimum $f(\boldsymbol{x}_\star)$, the following inequality condition holds trivially:

$$q(\boldsymbol{\lambda}_\star) \leq f(\boldsymbol{x}_\star). \tag{C.3.6}$$

The condition is known as the *weak duality condition*. Furthermore, when the equality can be obtained in (C.3.6), the duality gap between $f$ and $q$ becomes zero, and we say the primal and dual function pair satisfy the *strong duality condition*.

The strong duality condition can be achieved for convex objective functions subject to linear constraints.

THEOREM C.7 (Strong Duality Theorem). *Let the objective function $f(\boldsymbol{x})$ in (C.2.1) be convex and $\boldsymbol{h}(\boldsymbol{x})$ be linear. If the optimal value $f_\star$ is finite, then the optimal solution for its dual problem exists and there is no duality gap.*

Under the strong duality condition, the minimal value of $f$ can be found by optimizing the dual problem $q(\boldsymbol{\lambda})$, and the optimal primal solution can also be obtained by minimizing the Lagrangian function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_\star)$ over $\boldsymbol{x}$. In other words, the optimal $(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star)$ is the saddle point of the Lagrangian function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ that solves the following program:

$$\max_{\boldsymbol{\lambda}} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{C.3.7}$$

In the above, we have assumed all functions are differentiable. In this book, we often need to optimize a convex function that is not differentiable and the type of constraints are in the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$.

LEMMA C.8 (Dual Certificate). *Let $f : \mathbb{R}^n \to \mathbb{R}$ convex, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, and let $\boldsymbol{x}_\star$ be some point satisfying $\boldsymbol{A}\boldsymbol{x}_\star = \boldsymbol{y}$. If there exists $\boldsymbol{\nu}$ such that*

$$\boldsymbol{A}^*\boldsymbol{\nu} \in \partial f(\boldsymbol{x}_\star), \tag{C.3.8}$$

*then $\boldsymbol{x}_\star$ is a solution to the optimization problem*

$$\begin{array}{ll} \min & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{array} \tag{C.3.9}$$

*Proof* Consider any $\boldsymbol{x}'$ satisfying $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{y}$. By the subgradient inequality (B.3.3),

$$\begin{aligned} f(\boldsymbol{x}') &\geq f(\boldsymbol{x}_\star) + \langle \boldsymbol{A}^*\boldsymbol{\nu}, \boldsymbol{x}' - \boldsymbol{x}_\star \rangle \\ &= f(\boldsymbol{x}_\star) + \langle \boldsymbol{\nu}, \boldsymbol{A}(\boldsymbol{x}' - \boldsymbol{x}_\star) \rangle \\ &= f(\boldsymbol{x}_\star), \end{aligned} \tag{C.3.10}$$

since $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{x}_\star$. Thus, $\boldsymbol{x}_\star$ is optimal. □

# Appendix D Methods for Optimization

In this chapter, we review classical approaches to solving optimization problems of the form

$$\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}), \tag{D.0.1}$$

in which we seek to minimize an objective function $f$ over some domain $\mathcal{D}$. All of the algorithms we describe are *iterative methods* of optimization, which produce a sequence of points

$$\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \ldots \tag{D.0.2}$$

starting from some initialization $\boldsymbol{x}_0$. The goal is to generate a sequence $\{\boldsymbol{x}_k\}$ which quickly converges to a minimizer $\boldsymbol{x}_\star$ of $f$ over $\mathcal{D}$. The total time an iterative method requires to produce an acceptable answer depends chiefly on two quantities:

1 **per iteration cost**: how much computation it takes to generate the next point $\boldsymbol{x}_{k+1}$ given the previous points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k$.
2 **convergence rate**: how quickly the iterate $\boldsymbol{x}_k$ improve in quality. This dictates how many iterations are required to produce a sufficiently accurate solution. This may be measured either in terms of the distance of the iterate $\boldsymbol{x}_k$ to a minimizer,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2, \tag{D.0.3}$$

or in terms of the sub-optimality in objective value:

$$|f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)|, \tag{D.0.4}$$

or its gradient:[1]

$$\|\nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_\star)\|_2 = \|\nabla f(\boldsymbol{x}_k)\|_2. \tag{D.0.5}$$

The above two cost quantities are usually in tension: we can have fast convergence rate at the price of very expensive iterations, or we can have very cheap iterations at the price of a relatively slow convergence. Hence, the overall complexity of an optimization algorithm is typically measured as:

$$\text{complexity} = \text{per iteration cost} \times \#\text{ of iterations}, \tag{D.0.6}$$

[1] when we are only interested in converging to stationary point of the objective function with $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$.

subject to a prescribed accuracy in $\boldsymbol{x}$ or the objective value $f(\boldsymbol{x})$.

In the era of big data or large models, many practical problems involve optimizing over very large number of model parameters or training over large-scale datasets. Due to computation limitations, we typically can only afford to do fairly simple calculations in each iteration. Hence we are mainly interested in methods that achieve the fastest possible convergence rate out of methods that only work with *first-order* information (values of $f(\boldsymbol{x})$ and $\nabla f(\boldsymbol{x})$). Sometimes due to memory limitation and time requirement, we need to store the data and conduct the calculation over many *parallel* processes or a *distributed* network of machines. To reduce communication cost and delay, we often prefer algorithms that are amenable to parallel or distributed implementation and require minimal exchange of data and information across different processes or machines. In this appendix, we sketch basic ideas of some of the most popular and effective techniques that enhance the performance of first-order methods, especially those that are suitable for solving large-scale problems. We also provide references where the reader can find more complete exposition and analysis of these techniques.

## D.1 Gradient Descent

Perhaps the simplest iterative method of optimization is *gradient descent*, also known as the gradient method, which applies to *differentiable* functions $f : \mathbb{R}^n \to \mathbb{R}$. The method was first introduced by Cauchy in 1847 to solve systems of equations [Cau47]. It comes from the simplest idea that from the current state $\boldsymbol{x}_k$, one would like to take a small step $t \geq 0$ in the direction $\boldsymbol{v} \in \mathbb{R}^n$ to $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t \cdot \boldsymbol{v}$ such that the value of $f$ decreases:

$$f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k).$$

Since $f$ is differentiable, we know that up to first-order approximation:

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) = f(\boldsymbol{x}_k + t \cdot \boldsymbol{v}) - f(\boldsymbol{x}_k) \approx t \cdot \nabla f(\boldsymbol{x}_k)^* \boldsymbol{v}.$$

The gradient $\nabla f(\boldsymbol{x}_k)$ points in the direction of steepest increase of the objective $f$; the negative gradient is the direction of steepest decent. So in order for $f(\boldsymbol{x}_{k+1})$ to be smaller than $f(\boldsymbol{x}_k)$, it is natural to take the direction in which the value of $f$ drops the fastest: $\boldsymbol{v} \propto -\nabla f(\boldsymbol{x}_k)$. Hence the *gradient descent* is also know as the *steepest descent*.

Therefore, gradient descent generates its next iterate by stepping in the direction of the negative gradient

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k). \tag{D.1.1}$$

Here, $t_k \geq 0$ is a scalar, often called the *step size*.[2] The step size $t_k$ can either be determined analytically from the properties of the function $f$, or numerically by

[2] or the *learning rate* in learning algorithms.

performing a *line search*, which produces an approximate solution[3] to the one dimensional problem:

$$\min_{t\geq 0} \; f\left(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)\right). \tag{D.1.2}$$

*Convergence of Gradient Descent.*
A principal virtue of gradient descent is that for many problems, $\nabla f$ can be computed efficiently. To understand the overall properties of the method, we need to know how many iterations it requires to obtain a solution of a given desired quality. This depends in turn on the properties of the objective function $f$.

We begin by assuming that $f$ is a convex, differentiable function, and that the gradient $\nabla f(\boldsymbol{x})$ is $L$-Lipschitz:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\|_2 \leq L\,\|\boldsymbol{x} - \boldsymbol{x}'\|_2\,, \quad \forall\, \boldsymbol{x}, \boldsymbol{x}'. \tag{D.1.3}$$

This condition states that the gradient does not change too rapidly as we move from point to point. Intuitively, this means that a first-order model for the objective function generated by taking a Taylor expansion at point $\boldsymbol{x}$ will be valid over a relatively large portion of the space. Indeed, it turns out that under these hypotheses, we can take $t_k$ to a uniform

$$t_k = \tfrac{1}{L},$$

and smaller $L$ allows larger steps. Moreover, it can be shown that with this choice,

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \tfrac{1}{2L}\,\|\nabla f(\boldsymbol{x}_k)\|_2^2 \\ &\leq f(\boldsymbol{x}_k). \end{aligned} \tag{D.1.4}$$

Thus, with this choice, the gradient method is a *descent method*: it strictly decreases the objective at each iteration, until $\boldsymbol{x}_k$ reaches a minimizer. The following theorem gives an overall control on the rate of convergence, measured in function values:

THEOREM D.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with $\nabla f(\boldsymbol{x})$ $L$-Lipschitz. Let $\mathsf{X}_\star \neq \varnothing$ denote the set of minimizers of $f$, and $f_\star$ the minimum value of $f$ over $\mathbb{R}^n$. Consider the gradient method with constant step size $t_k = \frac{1}{L}$. Then*

$$f(\boldsymbol{x}_k) - f_\star \leq \frac{L}{2}\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{k}. \tag{D.1.5}$$

*Moreover, as $k \to \infty$, $\boldsymbol{x}_k \to \mathsf{X}_\star$.*

A proof of this theorem (actually a more generalized version) can be found in Chapter 8, Section 8.2.

Several aspects of this result are worth noting. First, the suboptimality in

[3] Typically, this is done by *backtracking*: starting from some nominal value of $t$, we reduce $t$ until the function value decreases adequately, say satisfying the Armijo rule.

function values decreases as $1/k$. In particular, as $k \to \infty$, $f(\boldsymbol{x}_k) \to f_\star$. Second, the rate of convergence depends on the Lipschitz constant $L$ – the smaller $L$ is, faster $f$ approaches $f_\star$. Finally, the rate of convergence depends on the distance of the initialization to $\boldsymbol{x}_\star$. A strength of this result is that it is nonasymptotic (the bound works for all $k$, not just $k$ large) and does not depend on dimension $n$. For applications, we care not just about function values, but about the quality of the iterates $\{\boldsymbol{x}_k\}$. Here, we are guaranteed that $\boldsymbol{x}_k$ approaches $\mathsf{X}_\star$. However, no general, dimension-independent bound on the rate of convergence is known.

## D.2 Rates of Convergence and Acceleration

How good is the gradient method? More generally, if we restrict ourselves to relatively simple methods that only use gradient and function value information, what rate can we obtain? This fundamental question motivates the study of lower bounds for the computational efficiency of methods. This requires a model of computation. One simple model for first-order methods assumes that at each iteration, the next point $\boldsymbol{x}_{k+1}$ is generated based only on the previous points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k$, their function values $f(\boldsymbol{x}_0), \ldots, f(\boldsymbol{x}_k)$, and gradients $\nabla f(\boldsymbol{x}_0), \ldots, \nabla f(\boldsymbol{x}_k)$:

$$f(\boldsymbol{x}_{k+1}) = \mathcal{F}_{k+1}\big(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k, f(\boldsymbol{x}_0), \ldots, f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_0), \ldots, \nabla f(\boldsymbol{x}_k)\big). \quad \text{(D.2.1)}$$

This is sometimes referred to as a *black box model*, since the method only accesses the function $f$ through its value and gradient at a finite discrete set of points.[4]

It has been shown that (see [Nes03]):

THEOREM D.2 (Convergence Rate of Gradient Descent). *For every $L$ and $R$, there exists a convex differentiable function $f$ with $\nabla f$ $L$-Lipschitz, and an initial point $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2 \leq R$ such that*

$$f(\boldsymbol{x}_k) - f_\star \geq c\frac{LR^2}{k^2}, \quad \text{(D.2.2)}$$

*where $c > 0$ is a numerical constant.*

This result can be read as saying that for the class of functions with Lipschitz continuous gradients, the best generic rate of convergence that any gradient-like method can achieve is $O(1/k^2)$. Notice that Theorem D.1 implies that the gradient method converges at a rate of $O(1/k)$. For large $k$, this is *much* worse!

[4] This is fundamentally different from having access to those values over a continuous set, since any algorithm that relies on such assumption is in fact, strictly speaking, not computable. Sometimes we may use the continuous time dynamics such as the negated gradient-flow $\dot{\boldsymbol{x}} = -\nabla f(\boldsymbol{x})$ to study qualitative behaviors of certain algorithms, such as what type of critical points they converge to. Such dynamics however do not directly translate to implementable algorithms through naive discretization of the time, because many of the quantitative properties of the dynamics would not necessarily be preserved by such discretization.

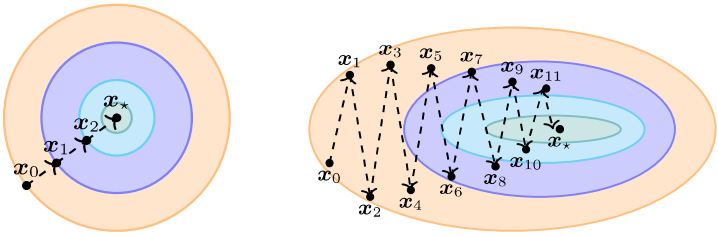


**Figure D.1** Illustration of the iteration behaviors of gradient descent. Left: A quadratic function with spherical level sets. Right: A quadratic function with more ellipsoidal level sets.

Could the gradient method be suboptimal? Figure D.1 shows the behavior of gradient descent on two different problems. The figure plots the level sets $S_\beta = \{\boldsymbol{x} \mid f(\boldsymbol{x}) = \beta\}$ of the objective $f$ as well as the iterates $\{\boldsymbol{x}_k\}$. Because the gradient $\nabla f(\boldsymbol{x})$ is orthogonal to the level set containing $\boldsymbol{x}$, the gradient method moves orthogonal to the level sets. At left, we show a function $f(\boldsymbol{x})$ whose level sets are nearly circular. The gradient method makes rapid progress. At right is a function $f(\boldsymbol{x})$ whose level sets are more elongated. The iterates "chatter" repeatedly changing direction and making slow progress towards $\boldsymbol{x}_\star$.

*The Heavy Ball Method.*
The bad behavior in Figure D.1 can be mitigated by preventing the steps $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ from changing direction too rapidly. An intuitive way to accomplish this is to treat the iterate $\boldsymbol{x}_k$ as the trajectory of a particle with some amount of momentum, which causes it to continue moving in the same direction. This suggests an update of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k) + \beta_k \left(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right). \tag{D.2.3}$$

Because this emulates the trajectory of a particle with nonzero mass, this method is aptly called the *heavy ball method*, first introduced by Polyak in 1964 [Pol64]. This method is also sometimes known as the *momentum method*, as the second term can be viewed as carrying some momentum from the previous iteration. This is the basis for the popular momentum-based ADAM algorithm for training modern neural networks [KB14]. Figure D.2 compares the heavy ball method to the gradient method on an ill-conditioned quadratic. Notice that the heavy ball method takes far fewer iterations to reach the vicinity of $\boldsymbol{x}_\star$.

*Nesterov's Accelerated Method.*
Although the heavy ball method improves over the gradient method, its worst case rate of convergence is still $O(1/k)$. However, by using momentum in a clever way, it *is* possible to achieve a better rate of convergence of $O(1/k^2)$, which matches the lower bound in Theorem D.2. This means, perhaps surprisingly,

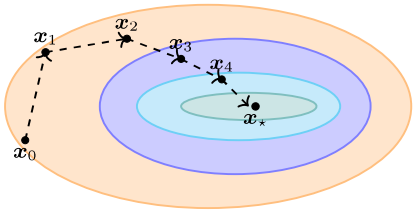


**Figure D.2** Illustration of gradient descent with the heavy ball method.

that there is a gradient-like method that is fundamentally better than gradient descent!

The method that achieves this optimal rate is known as *Nesterov's accelerated gradient method*. Strictly speaking, it is not a momentum method. Rather, it uses two sequences of iterates $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{p}_k\}$. The auxiliary point $\boldsymbol{p}_k$ is extrapolated from $\boldsymbol{x}_k$ in a form similar to that in the heavy ball method:

$$\boldsymbol{p}_{k+1} \doteq \boldsymbol{x}_k + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}).$$

At each iteration, we move to this new point, compute the gradient at this point, and descend from it (instead of $\boldsymbol{x}_k$):

$$\boldsymbol{x}_{k+1} = \boldsymbol{p}_{k+1} - \alpha \nabla f(\boldsymbol{p}_{k+1}). \tag{D.2.4}$$

As we will show in Section 8.3 of Chapter 8, with properly chosen weights $\beta_k$ and $\alpha$, the gradient method is indeed accelerated and can achieve the optimal convergence rate of $O(1/k^2)$, for the class of functions with Lipschitz continuous gradients,.

THEOREM D.3 (Convergence Rate of Accelerated Gradient Method). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with $\nabla f(\boldsymbol{x})$ being $L$-Lipschitz. Let $\mathsf{X}_\star \neq \varnothing$ denote the set of minimizers of $f$ and $f_\star$ the minimum value of $f$ over $\mathbb{R}^n$. The iterates $\{\boldsymbol{x}_k\}$ produced by the accelerated gradient method satisfy*

$$f(\boldsymbol{x}_k) - f_\star \leq \frac{L}{2} \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{(k+1)^2}. \tag{D.2.5}$$

*Moreover, as $k \to \infty$, $\boldsymbol{x}_k \to \boldsymbol{x}_\star$.*

Recently several work try to understand such acceleration by characterizing the stability of continuous ordinary differential equations associated with such iterations [SBC14] (and many subsequent work [KBB16, KBB15, WWJ16]). A more detailed survey and discussion can be found in Section 8.3 of Chapter 8.

*Strongly Convex Functions.*

Notice that Theorem D.2 characterizes the best possible rate of convergence for gradient-like methods for the class of functions with Lipschitz continuous gradients; and Theorem D.3 states that this rate can be achieved with the accelerated gradient methods. Nevertheless, this does not means that this is the

best one can do for more restricted classes of functions with better properties. If, in addition to Lipschitz continuous gradients, the functions satisfy additional properties such as strongly convex defined in Appendix B, it is long known in the optimization literature that gradient-descent type methods can converge at a *linear rate* [BV04].

THEOREM D.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable strongly convex function with constant $\mu$ and $\nabla f(\boldsymbol{x})$ being $L$-Lipschitz. Let $f_\star$ be the minimum value of $f$ over $\mathbb{R}^n$. Then the iterates $\{\boldsymbol{x}_k\}$ produced by the gradient-descent $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$ with $t = \frac{1}{L}$ satisfy*

$$f(\boldsymbol{x}_k) - f_\star \leq \frac{L}{2} e^{-\alpha k} \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 \tag{D.2.6}$$

*for some constant $\alpha > 0$.*

*Proof* We here give a proof to this simple fact as it helps explain why gradient descent converges very fast for many statistical learning problems in practice – the objective (loss) functions often concentrate on a function that is both strongly convex and smooth, as the size of random samples increase.

First, notice that at the optimal solution $\boldsymbol{x}_\star$ we have $\nabla f(\boldsymbol{x}_\star) = \boldsymbol{0}$. According to Lemma 8.2, we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \leq \frac{L}{2} \|\boldsymbol{x}_k - \boldsymbol{x}_\star)\|_2^2.$$

Also, due to the strong convex and smooth assumption, we also have:

$$\mu \cdot \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L \cdot \boldsymbol{I}, \quad \forall \boldsymbol{x}. \tag{D.2.7}$$

From the gradient descent rule, and with the fundamental theorem of calculus, we have

$$\begin{aligned} \boldsymbol{x}_{k+1} - \boldsymbol{x}_\star &= \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k) - \boldsymbol{x}_\star \\ &= \boldsymbol{x}_k - \boldsymbol{x}_\star - t\Big(\int_0^1 \nabla^2 f(\boldsymbol{x}_\star + \tau(\boldsymbol{x}_k - \boldsymbol{x}_\star))d\tau\Big)(\boldsymbol{x}_k - \boldsymbol{x}_\star). \end{aligned}$$

This gives:

$$\begin{aligned} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_2 &\leq \left\| \boldsymbol{I} - t\int_0^1 \nabla^2 f(\boldsymbol{x}_\star + \tau(\boldsymbol{x}_k - \boldsymbol{x}_\star))d\tau \right\| \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \\ &\leq (1 - t\mu)\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2. \end{aligned}$$

If we choose $t = 1/L$, then $(1 - \frac{\mu}{L}) < 1$, we have contraction of $\boldsymbol{x}_k - \boldsymbol{x}_\star$ in $\ell^2$ norm. So we have

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \leq \big(1 - \frac{\mu}{L}\big)^k \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2, \quad \forall k.$$

Or equivalently

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 \leq \big(1 - \frac{\mu}{L}\big)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2, \quad \forall k.$$

Now, let $\alpha = -\log(1 - \frac{\mu}{L})^2 > 0$, we obtain the desired result. □

As we see from the above proof, we may also set the step size to be $t = \frac{2}{L+\mu}$ and get slightly better contraction factor.

According to the above theorem, $f(\boldsymbol{x}_k) - f_\star$ converges to zero exponentially in the order of $O(e^{-\alpha k})$, much faster than $O(1/k^2)$. In this book, the class of functions that we often encounter are not necessarily globally strongly convex. Nevertheless, they may satisfy certain weaker notion of strong convexity, such as *restricted strong convexity* or local strong convexity. We will see that under such conditions, one may also expect gradient-like methods to achieve linear rate of convergence around the global minimum.

*Nondifferentiable Functions.*
The main assumption of gradient descent methods is that the objective function $f(\boldsymbol{x})$ is differentiable in $\boldsymbol{x}$. In this book, we often need to minimize functions that are not everywhere differentiable, such as functions involving the $\ell^1$ norm $\|\boldsymbol{x}\|_1$. In such cases, we need to generalize the notion of gradient to "subgradients" (see Definition 2.12 in Chapter 2). Essentially, subgradients at a point $\boldsymbol{x}$ is the set of vectors $\boldsymbol{u} \in \mathbb{R}^n$ such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{x} \rangle, \quad \forall \boldsymbol{y}.$$

We often denote the set of subgradients as $\partial f(\boldsymbol{x})$. To minimize such a function $f(\boldsymbol{x})$, we may generalize the gradient descent method by replacing the gradient $\nabla f(\boldsymbol{x})$ with any subgradient:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \boldsymbol{g}_k, \quad \boldsymbol{g}_k \in \partial f(\boldsymbol{x}).$$

A main disadvantage of such subgradient descent methods is their relatively poor convergence rate. In general, the convergence rate of subgradient descent for non-smooth objective functions is

$$f(\boldsymbol{x}_k) - f_\star = O\big(1/\sqrt{k}\big).$$

The reader can refer to [Nem95, Nem07, Nes03] for more detailed analysis of subgradient descent algorithms.

It is worth noticing the significant difference in convergence rates for the same gradient-descent algorithm being applied to two extreme subclasses of convex functions: the strongly convex functions versus nondifferentiable ones. For the former, gradient descent converges linearly $O(e^{-\alpha k})$, and yet for the latter it converges much slower with a rate $O(1/\sqrt{k})$.

Nevertheless, as we will see in Chapter 8, in many of our problems, the objective function $f(\boldsymbol{x})$ is of the form $f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ with $f_1$ being smooth and $f_2$ nonsmooth. If for the nonsmooth part $f_2$, the so called *proximal operator*:

$$\min_{\boldsymbol{x}} f_2(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{w}\|_2^2 \tag{D.2.8}$$

has a closed-form solution or can be solved efficiently, then the subgradient descent method can be properly modified so that it would enjoy the same conver-

gence rate as the smooth case. See the *proximal gradient* method in Section 8.2 of Chapter 8.

## D.3 Constrained Optimization

It is very common in practice that we want to minimize a function $f(\boldsymbol{x})$ while the desired solution $\boldsymbol{x}_\star$ is constrained to some subset $\mathsf{C} \subset \mathbb{R}^n$:

$$\begin{array}{ll} \min & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathsf{C}. \end{array} \tag{D.3.1}$$

Solutions in the subset $\mathsf{C}$ are called *feasible* solutions. Notice that if we still apply the gradient descent method to minimize $f(\boldsymbol{x})$. Then, after each descent iteration

$$\boldsymbol{p}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k),$$

even if $\boldsymbol{x}_k$ is feasible, the new state $\boldsymbol{p}_{k+1}$ may step outside of the constrained set: $\boldsymbol{p}_{k+1} \notin \mathsf{C}$. A natural and simple fix to this issue is to "project" $\boldsymbol{p}_{k+1}$ back to the set $\mathsf{C}$:

$$\boldsymbol{x}_{k+1} = \mathcal{P}_{\mathsf{C}}[\boldsymbol{p}_{k+1}] = \arg\min_{\boldsymbol{x} \in \mathsf{C}} \tfrac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{p}_{k+1} \right\|_2^2, \tag{D.3.2}$$

where $\boldsymbol{x}_{k+1}$ is the point in $\mathsf{C}$ closest to $\boldsymbol{p}_{k+1}$. This will ensure the new iterate $\boldsymbol{x}_{k+1}$ is always feasible. This method is called *projected gradient descent*, and we use it to provide a simplest algorithm for minimizing the $\ell^1$ norm in Chapter 2. This simple method is also the inspiration for other first-order constrained optimization methods such as the classic *Frank-Wolfe* method [FW56] that we study in Section 8.6 of Chapter 8.

One disadvantage of such projected gradient descent methods is their relatively poor convergence rate or computational efficiency per iteration[5]. In the case the constraints are equality constraints: $\mathsf{C} = \{\boldsymbol{x} \mid \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}\}$, one could try to convert the constrained optimization

$$\begin{array}{ll} \min & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \end{array} \tag{D.3.3}$$

to an unconstrained one by penalizing any deviation of $\boldsymbol{h}(\boldsymbol{x})$ from $\boldsymbol{0}$:

$$\min f(\boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{h}(\boldsymbol{x})\|_2^2. \tag{D.3.4}$$

This is known as the *penalty method*. One can show that as $\mu \to +\infty$, the solution to the unconstrained optimization approaches that of the constrained one. However, in practice, as $\mu$ becomes large, the unconstrained problem becomes increasingly harder to solve as its gradient Lipschitz constant becomes increasingly large. See Section 8.4 of Chapter 8 for an example.

[5] unless the constraint set $\mathsf{C}$ is nice so that projection onto it or optimization over it is easy. That is precisely the assumption of the Frank-Wolfe method.

As we have discussed in Appendix C, another way to convert the constrained optimization problem is through the Lagrangian formulation. The optimal (feasible) solution $\boldsymbol{x}_\star$ to the above constrained optimization is also the optimal solution $(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star)$ to the unconstrained optimization:

$$\max_{\boldsymbol{\lambda}} \min_{\boldsymbol{x}} \ \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \tag{D.3.5}$$

where the Lagrangian function is defined as

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \doteq f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{h}(\boldsymbol{x}) \rangle.$$

It is natural to consider solving the above min-max problem through the following alternating optimization scheme:

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \ \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_k), \tag{D.3.6}$$

$$\boldsymbol{\lambda}_{k+1} = \arg\max_{\boldsymbol{\lambda}} \ \mathcal{L}(\boldsymbol{x}_{k+1}, \boldsymbol{\lambda}). \tag{D.3.7}$$

Although the saddle point of the Lagrangian is the desired optimal solution, there is no guarantee that each step of the above iteration would produce feasible iterates nor the process is guaranteed to converge. As we see in Section 8.4 of Chapter 8, even for some simple problems, the above subproblems might fail to have a solution (the value of the objective function can be unbounded).

To remedy this problem, one could augment the Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ with an extra quadratic penalty term for the constraint:

$$\mathcal{L}_\mu(\boldsymbol{x}, \boldsymbol{\lambda}) \doteq f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{h}(\boldsymbol{x}) \rangle + \frac{\mu}{2} \|\boldsymbol{h}(\boldsymbol{x})\|_2^2,$$

which is known as the *augmented Lagrangian* [Hes69, Roc73, Pow69]. As we will see in Section 8.4 of Chapter 8, the augmented Lagrangian leads to much better conditioned subproblems for the alternating scheme:

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \ \mathcal{L}_\mu(\boldsymbol{x}, \boldsymbol{\lambda}_k), \tag{D.3.8}$$

$$\boldsymbol{\lambda}_{k+1} = \arg\max_{\boldsymbol{\lambda}} \ \mathcal{L}_\mu(\boldsymbol{x}_{k+1}, \boldsymbol{\lambda}), \tag{D.3.9}$$

and the sequence of iterates $\{(\boldsymbol{x}_k, \boldsymbol{\lambda}_k)\}$ typically converge to the desired optimal solution $(\boldsymbol{x}_\star, \boldsymbol{\lambda}_\star)$ for a properly chosen $\mu$ or a sequence $\{\mu_k\}$.

## D.4 Block Coordinate Descent and ADMM

In many optimization problems we may encounter in practice, the dimension of $\boldsymbol{x}$ could be so high that we might not even afford to conduct gradient descent to minimize $f(\boldsymbol{x})$ for all the variables together. Very often the objective function $f(\boldsymbol{x})$ has certain decomposable structures such as a finite sum:

$$\min f(\boldsymbol{x}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}^i). \tag{D.4.1}$$

For example, the $\ell^1$-norm function $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$ is such a decomposable function. In such cases, we may conduct the so-called *block coordinate descent* to take advantage of such decomposable structures by iteratively minimizing the objective function with respect to one block of variables at a time.

More specifically, assume the domain $\mathcal{D}$ can be written as a Cartesian product

$$\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_m,$$

where each $\mathcal{D}_i \subseteq \mathbb{R}^{n_i}$, $n_1+n_2+\cdots+n_m = n$. The variables can be also partitioned into $m$ blocks as $\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^m) \in \mathbb{R}^n$ with each $\boldsymbol{x}^i \in \mathcal{D}_i$. The block coordinate descent scheme proceeds as follows:

1 Initialize $\boldsymbol{x}_0 = (\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \ldots, \boldsymbol{x}_0^m)$.
2 In the $k$-th iteration, for every $i = 1, \ldots, m$,

$$\boldsymbol{x}_k^i = \arg\min_{\bar{\boldsymbol{x}} \in \mathcal{D}_i} f\left(\boldsymbol{x}_k^1, \ldots, \boldsymbol{x}_k^{i-1}, \bar{\boldsymbol{x}}, \boldsymbol{x}_{k-1}^{i+1}, \ldots, \boldsymbol{x}_{k-1}^m\right).$$

3 Repeat Step 2 until the solution converges.

In the literature, the convergence of block coordinate descent methods can be proven under different conditions. A most natural condition is when the objective function $f(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{i-1}, \boldsymbol{x}^i, \boldsymbol{x}^{i+1}, \ldots, \boldsymbol{x}^m)$ is *strictly convex* with respect to each block $\boldsymbol{x}^i$. This guarantees the minimal solution $\boldsymbol{x}_\star^i$ is also unique. For a more detailed discussion about conditions under which such methods converge, the reader is referred to [Ber03].

In compressive sensing or statistical learning,[6] very often we need to deal with an objective function $f(\boldsymbol{x})$ that is a sum of multiple terms:

$$f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_m(\boldsymbol{x}). \tag{D.4.2}$$

To obtain more scalable algorithms such that we can optimize each term in a parallel or distributed fashion, we could rewrite this problem in terms of a set of local variables $\boldsymbol{x}^i \in \mathbb{R}^n$ and one global variable $\boldsymbol{z}$:

$$\min \quad \sum_{i=1}^m f_i(\boldsymbol{x}^i) \quad \text{subject to} \quad \boldsymbol{x}^i = \boldsymbol{z}, \quad i = 1, \ldots, m. \tag{D.4.3}$$

In the literature, this is also known as the *consensus optimization*. To solve such a constrained optimization problem, we could apply the above block descent method to its augment Lagrangian:

$$\mathcal{L}(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m, \boldsymbol{z}, \boldsymbol{\lambda}) = \sum_{i=1}^m f_i(\boldsymbol{x}^i) + \langle \boldsymbol{\lambda}^i, \boldsymbol{x}^i - \boldsymbol{z} \rangle + \frac{\mu}{2}\|\boldsymbol{x}^i - \boldsymbol{z}\|_2^2. \tag{D.4.4}$$

[6] Say training a deep neural networks over a very large set of training samples, where $\boldsymbol{x}$ are the network parameters.

This leads to the following iterative process:

$$\begin{aligned}
\boldsymbol{x}_{k+1}^i &= \arg\min_{\boldsymbol{x}^i} f_i(\boldsymbol{x}^i) + \langle \boldsymbol{\lambda}^i, \boldsymbol{x}^i - \boldsymbol{z} \rangle + \frac{\mu}{2}\|\boldsymbol{x}^i - \boldsymbol{z}\|_2^2, \\
\boldsymbol{z}_{k+1} &= \frac{1}{m}\sum_{i=1}^m \big(\boldsymbol{x}_{k+1}^i + \frac{1}{\mu}\boldsymbol{\lambda}_k^i\big), \\
\boldsymbol{\lambda}_{k+1}^i &= \boldsymbol{\lambda}_k^i + \mu\big(\boldsymbol{x}_{k+1}^i - \boldsymbol{z}_{k+1}\big).
\end{aligned}$$

This is known as the *Alternating Direction Method of Multipliers* (ADMM). Notice that the above scheme is rather amenable to distributed implementation as each local process can solve in parallel a subproblem for $\boldsymbol{x}^i$ and then share the information through the common variable $\boldsymbol{z}$.

Although ADMM has been a very popular scheme widely used by practitioners, the analysis for its convergence and convergence rate is far from trivial. In Chapter 8, we will study the ADMM scheme for the case with $m = 2$ in great detail, as it is closely applicable to our problems (such as the Robust PCA problem considered in Chapter 5). Convergence analysis for more general cases remains largely open research topics. For a more detailed exposition of ADMM and more general variants, the reader may refer to the recent manuscript of [BPC+11].

# Appendix E Facts from High-Dimensional Statistics

*"God tirelessly plays dice under laws which he has himself prescribed."*
– Albert Einstein

In this appendix, we recount a few facts about high-dimensional statistics and concentration of measure which are used throughout the text. The results that we quote are examples of a pervasive phenomenon: functions of many independent random variables often concentrate sharply about their expectations. In this section we give only a brief account of a few concentration inequalities that are used throughout the text, starting in with classical scalar inequalities in Section E.1 and their counterparts for matrices in Section E.2. We refer the reader to the recent texts [BLM13, Ver08, Ver18, Wai19] for deeper and more through accounts of high-dimensional probability and its applications.

## E.1 Basic Concentration Inequalities

Our first concentration inequality pertains to sums of independent bounded random variables $X_1, \dots, X_m$. For simplicity, we assume that the $X_i$ have zero mean.

THEOREM E.1 (Hoeffding's Inequality). *Let $X_1, \dots, X_m$ be independent random variables, with $\mathbb{E}[X_i] = 0$, and $|X_i| \le R$ almost surely,*

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} X_i\right| > t\right] \le 2\exp\left(-\frac{t^2}{2mR^2}\right). \tag{E.1.1}$$

This theorem implies that the sum $\sum_i X_i$ exhibits a *subgaussian* tail: the tail probability decays as $e^{-ct^2}$. The proof is an application of the *exponential moment method* (sometimes referred to the Cramer-Chernoff method), in which we apply Markov's inequality[1] to the nonnegative random variable $\exp(\lambda \sum_i X_i)$. This general approach yields not only Hoeffding's inequality, but many other classical concentration inequalities. We illustrate the method by proving Theorem E.1 below:

[1] Recall that Markov's inequality states that for a nonnegative random variable $Y$, $\mathbb{P}[Y > t] < \mathbb{E}[Y]/t$.

*Proof* We calculate

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{m} X_i > t\right] &= \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^{m} X_i\right) > \exp(\lambda t)\right] \\ &\le e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m} X_i\right)\right] \\ &= e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{m} e^{\lambda X_i}\right] \\ &= e^{-\lambda t}\prod_{i=1}^{m}\mathbb{E}\left[e^{\lambda X_i}\right]. \end{aligned} \tag{E.1.2}$$

Using that for $s \in [-R, R]$, $e^{\lambda s} \le 1 + \lambda s + \frac{1}{2}\lambda^2 R^2$, we have that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda X_i}\right] &\le \mathbb{E}\left[1 + \lambda X_i + \tfrac{1}{2}\lambda^2 R^2\right] \\ &= 1 + \tfrac{1}{2}\lambda^2 R^2 \\ &\le \exp\left(\tfrac{1}{2}\lambda^2 R^2\right). \end{aligned} \tag{E.1.3}$$

Plugging in to (E.1.2), we get that

$$\mathbb{P}\left[\sum_{i=1}^{m} X_i > t\right] \le \exp\left(-\lambda t + \frac{m}{2}\lambda^2 R^2\right). \tag{E.1.4}$$

Minimizing the exponent, by setting $\lambda = t/mR^2$, we obtain the claimed result, (E.1.1). □

Hoeffding's inequality gives a convenient tool for controlling sums of bounded random variables, which we use several times throughout the text. As mentioned above, it shows that the sum exhibits a subgaussian tail. In many cases the "variance" suggested by this tail, $mR^2$ is larger than the true variance if, e.g., $\mathbb{E}[X_i^2] = \sigma^2$ with $\sigma \ll R$. The classical Bernstein inequality also accounts for variance information:

THEOREM E.2 (Bernstein's Inequality). *Let $X_1, \ldots, X_m$ be independent random variables, with $\mathbb{E}[X_i] = 0$, $|X_i| \le R$ almost surely, and $\mathbb{E}[X_i^2] \le \sigma^2$. Then*

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} X_i\right| > t\right] \le 2\exp\left(-\frac{t^2/2}{m\sigma^2 + 3Rt}\right). \tag{E.1.5}$$

In essence, it says that for small $t$, the tail behaves $e^{-ct^2/m\sigma^2}$, i.e., Gaussian with standard deviation $m\sigma^2$, while for large $t$, the tail is *subexponential*, $e^{-ct/R}$. The proof of Bernstein's inequality proceeds under exactly the same lines as the proof of Hoeffding's inequality, up to line (E.1.2), but uses slightly different calculations to control the moment generating function $\mathbb{E}[e^{\lambda X_i}]$.

*Concentration for norms of Gaussian vectors.*
Using similar reasoning, we can obtain the following useful bound on the $\ell^2$ norm of a Gaussian vector, which is used throughout Chapter 3 to establish embedding results such as the Johnson-Lindenstrauss lemma and the Restricted Isometry Property:

LEMMA E.3. *Let $\boldsymbol{g} = (g_1, \dots, g_m)$ with the $g_i$ independent $\mathcal{N}(0, 1/m)$ random variables. Then for any $t \in [0, 1]$,*

$$\mathbb{P}\left[\left|\|\boldsymbol{g}\|_2^2 - 1\right| > t\right] \leq 2\exp\left(-\frac{t^2 m}{8}\right). \tag{E.1.6}$$

This lemma again follows from the proof scheme of Theorem E.1, noting that $\|\boldsymbol{g}\|_2^2 = \sum_{i=1}^m g_i^2$ is a sum of independent random variables and using the following expression for the moment generating function of the random variable $h_i = g_i^2$:

$$\mathbb{E}\left[e^{\lambda h_i}\right] = \left(1 - \frac{2\lambda}{m}\right)^{-1/2} \qquad \lambda < \frac{m}{2}, \tag{E.1.7}$$

and making an appropriate choice of $\lambda$.

*General concentration results for Lipschitz functions.*
The basic concentration results described above show that sums $f(X_1, \dots, X_m) = \sum_{i=1}^m X_i$ of independent random concentrate sharply about their expectations $\mathbb{E}[f(X_1, \dots, X_m)] = \sum_{i=1}^m \mathbb{E}[f(X_i)]$. Depending on the assumptions on the random variables $X_i$ the tail probability of the random variable $f(X_1, \dots, X_m) - \mathbb{E}[f(X_1, \dots X_m)]$ is either subgaussian or subexponential, i.e., it is dominated by either $e^{-ct^2}$ or $e^{-ct}$. In fact, this behavior can be observed not only for sums of random variables, but for much more general functions $f(X_1, \dots, X_m)$. At the slogan level, *sufficiently "nice" functions of many random variables* concentrate sharply about their expectations.

For example, suppose that $f$ satisfies a Lipschitz condition

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^m, \tag{E.1.8}$$

which controls how rapidly $f$ changes as the vector $\boldsymbol{x}$ changes. Then if $g_1, \dots, g_m$ are Gaussian random variables, $f(g_1, \dots, g_m)$ concentrates about its expectation:

THEOREM E.4 (Gauss-Lipschitz Concentration). *Let $f : \mathbb{R}^m \to \mathbb{R}$ by an $L$-Lipschitz function, and let $g_1, \dots, g_m \sim_{\text{iid}} \mathcal{N}(0, 1)$. Then*

$$\mathbb{P}\left[|f(g_1, \dots, g_m) - \mathbb{E}\left[f(g_1, \dots, g_m)\right]| > t\right] < 2\exp\left(-\frac{t^2}{2L}\right). \tag{E.1.9}$$

This theorem states that the random variable $f(g_1, \dots, g_m)$ has a subgaussian tail, which acts like a Gaussian random variable with variance $L$. The smaller the Lipschitz constant $L$ (i.e., the nicer the function $f$), the sharper the concentration about the expectation. The the orientation of the random vector $\boldsymbol{g} = (g_1, \dots, g_m)$ uniform: $\boldsymbol{u} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$ is uniformly distributed on the sphere $\mathbb{S}^{m-1}$. It should be no

surprise, then, that Lipschitz functions of uniformly distributed random vectors on the sphere also concentrate:

THEOREM E.5 (Concentration on the Sphere). *Let $f : \mathbb{S}^{m-1} \to \mathbb{R}$ be an $L$-Lipschitz function and let $\boldsymbol{u} \sim \text{uni}(\mathbb{S}^{m-1})$ be uniformly distributed on the sphere. Then*

$$\mathbb{P}\left[\,|f(\boldsymbol{u}) - \text{median} f(\boldsymbol{u})| > t\,\right] < 2\exp\left(-\frac{mt^2}{8L}\right). \tag{E.1.10}$$

This result again shows subgaussian concentration with variance proportional to the Lipschitz constant $L$. The result happens to be phrased in terms of the median, rather than the mean. However, brief calculations using (E.1.10) show that the median is close to the mean ($|\text{median}(f) - \mathbb{E}[f]| \leq C/\sqrt{L}$) and so $f(\boldsymbol{u})$ is typically within $O(1/\sqrt{L})$ of its expectation as well. In our book, this result has been used to construct incoherent matrices in Chapter 3.

These results on Lipschitz concentration have generalizations to other spaces [Led01]. They also have generalizations to other distributions. One powerful related result is Talagrand's inequality for convex Lipschitz functions on the cube [Tal95]. Finally, it is possible to show concentration under other hypotheses on the function $f$ – see [BLM13].

## E.2 Matrix Concentration Inequalities

The basic concentration inequalities in Section E.1 have natural generalizations from sums of independent random scalars to sums of independent random *matrices*. The basic concentration in equalities in Section E.1 are obtained by the exponential moment method, illustrated in the proof of Theorem E.1. This elegant approach can be used to derive a number of classical probability inequalities, by using different assumptions to get different bounds on the moment generating function. However, our interest is not just in scalar random variables, but in matrices, or even operators. Is there any natural way to generalize this approach? Remarkably, the answer is yes. Since the crucial step above is exponentiating and then applying Markov's inequality, we might hope to simply replace the scalar exponential with the matrix exponential. Surprisingly, it is *almost* that easy.

*Facts about the matrix exponential.*
Before carrying the above argument over to the matrix case, let us recall a few facts about matrices and matrix exponentials. Recall that a symmetric matrix $\boldsymbol{M}$ is semidefinite ($\boldsymbol{M} \succeq \boldsymbol{0}$) iff for all $\boldsymbol{x}$, $\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x} \geq 0$. We write

$$\boldsymbol{A} \succeq \boldsymbol{B},$$

whenever

$$\boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}.$$

The matrix exponential is the function

$$\exp(\boldsymbol{M}) \;=\; \sum_{n=0}^{\infty} \frac{\boldsymbol{M}^n}{n!} \;=\; \boldsymbol{I} + \boldsymbol{M} + \boldsymbol{M}^2/2 + \dots. \tag{E.2.1}$$

Since a symmetric matrix $\boldsymbol{M}$ has a complete orthonormal basis of eigenvector, $\boldsymbol{M} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$, with $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, the exponential of a symmetric matrix has a particularly simple form:

$$\exp(\boldsymbol{M}) \;=\; \boldsymbol{V}\exp(\boldsymbol{\Lambda})\boldsymbol{V}^* = \boldsymbol{V}\begin{bmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_n} \end{bmatrix}\boldsymbol{V}^* \succeq \boldsymbol{0}. \tag{E.2.2}$$

The exponential of a symmetric matrix is always semidefinite.

The matrix exponential satisfies many of the natural properties also satisfied by the scalar exponential. It differs in important ways, however, because matrix multiplication is not precisely analogous to scalar multiplication. In particular, in general, matrix multiplication does not commute: $\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A}$. This causes the property $\exp(s+t) = \exp(s)\exp(t)$ to fail for matrices:

$$\text{In general,} \quad \exp(\boldsymbol{A}+\boldsymbol{B}) \neq \exp(\boldsymbol{A})\exp(\boldsymbol{B}). \tag{E.2.3}$$

The only exception occurs when $\boldsymbol{A}$ and $\boldsymbol{B}$ do commute: $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$.

If our imagined program is to replace the scalar exponential with the matrix exponential in the proof of Bernstein's inequality, this fact is very bad news. The proof used in a very critical way, the fact that $\exp(s+t) = \exp(s)\exp(t)$. Fortunately, there is a weak analogue of this property that does hold for matrices, given by the following result of Golden [Gol65] and Thompson [Tho04]:

THEOREM E.6 (Golden-Thompson Inequality). *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be self-adjoint matrices. Then*

$$\operatorname{trace}[\exp(\boldsymbol{A}+\boldsymbol{B})] \;\leq\; \operatorname{trace}[\exp(\boldsymbol{A})\exp(\boldsymbol{B})]. \tag{E.2.4}$$

Before proceeding, we also note that for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\operatorname{trace}[\boldsymbol{A}\boldsymbol{B}] \;\leq\; \|\boldsymbol{A}\|\operatorname{trace}[\boldsymbol{B}]. \tag{E.2.5}$$

*Matrix Bernstein inequality.*

Let us apply the above results to demonstrate a probability inequality for matrices:

THEOREM E.7 (Matrix Bernstein Inequality). *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be $d \times d$ independent, identically distributed self-adjoint random matrices, with $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$ and $\|\boldsymbol{X}_i\| \leq 1$ almost surely. Then*

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\right) > t\right] \;\leq\; d\exp\left(-\min\left\{\frac{t^2}{4n}, \frac{t}{2}\right\}\right). \tag{E.2.6}$$

*Proof* Note that

$$\lambda_{\max}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\right) > t \iff \lambda_{\max}\left(\lambda\sum_{i}\boldsymbol{X}_i\right) > e^{\lambda t}$$
$$\implies \operatorname{trace}\left(\exp\left(\lambda\sum_{i}\boldsymbol{X}_i\right)\right) > e^{\lambda t}.$$

So,

$$\begin{aligned}
&\mathbb{P}\left[\lambda_{\max}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\right) > t\right] \le \mathbb{P}\left[\operatorname{trace}\left(\exp\left(\lambda\sum_{i}\boldsymbol{X}_i\right)\right) > e^{\lambda t}\right]\\
&\le e^{-\lambda t}\,\mathbb{E}\operatorname{trace}\left(\exp\left(\lambda\sum_{i=1}^{n}\boldsymbol{X}_i\right)\right)\\
&\le e^{-\lambda t}\,\mathbb{E}\operatorname{trace}\left(\exp\left(\lambda\boldsymbol{X}_n\right)\exp\left(\lambda\sum_{i=1}^{n-1}\boldsymbol{X}_i\right)\right)\\
&\le e^{-\lambda t}\operatorname{trace}\left(\mathbb{E}\left[\exp(\lambda\boldsymbol{X}_n)\right]\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1}\boldsymbol{X}_i\right)\right]\right)\\
&\le e^{-\lambda t}\,\left\|\mathbb{E}\left[\exp(\lambda\boldsymbol{X}_n)\right]\right\|\operatorname{trace}\left(\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1}\boldsymbol{X}_i\right)\right]\right)\\
&\le e^{-\lambda t}\,\left\|\mathbb{E}\left[\exp(\lambda\boldsymbol{X}_n)\right]\right\|\mathbb{E}\operatorname{trace}\left(\exp\left(\lambda\sum_{i=1}^{n-1}\boldsymbol{X}_i\right)\right)\\
&\le e^{-\lambda t}\prod_{i=2}^{n}\left\|\mathbb{E}\left[\exp(\lambda\boldsymbol{X}_i)\right]\right\|\mathbb{E}\operatorname{trace}\left(\exp(\lambda\boldsymbol{X}_1)\right)\\
&\le de^{-\lambda t}\,\left\|\mathbb{E}\exp(\lambda\boldsymbol{X})\right\|^{n}. \qquad \text{(E.2.7)}
\end{aligned}$$

To bound the "matrix moment generating function"

$$M_{\boldsymbol{X}}(\lambda) = \mathbb{E}\left[\exp(\lambda\boldsymbol{X})\right], \qquad \text{(E.2.8)}$$

we use a matrix variant of the scalar inequality $1 + s \le \exp(s) \le 1 + s + s^2$, namely, for any self-adjoint matrix $\boldsymbol{S}$ satisfying $-\boldsymbol{I} \preceq \boldsymbol{S} \preceq \boldsymbol{I}$, we have

$$\boldsymbol{I} + \boldsymbol{S} \preceq \exp(\boldsymbol{S}) \preceq \boldsymbol{I} + \boldsymbol{S} + \boldsymbol{S}^2. \qquad \text{(E.2.9)}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda\boldsymbol{X})\right] &\preceq \mathbb{E}\left[\boldsymbol{I} + \lambda\boldsymbol{X} + \lambda^2\boldsymbol{X}^2\right] &\text{(E.2.10)}\\
&\preceq \boldsymbol{I} + \lambda^2\mathbb{E}[\boldsymbol{X}^2] &\text{(E.2.11)}\\
&\preceq \boldsymbol{I} + \lambda^2\boldsymbol{I}. &\text{(E.2.12)}
\end{aligned}$$

So, $\left\|\mathbb{E}\exp(\lambda\boldsymbol{X})\right\| \le \left\|\boldsymbol{I} + \lambda^2\boldsymbol{I}\right\| = 1 + \lambda^2 \le \exp(\lambda^2)$. From this, we obtain

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\right) > t\right] \le de^{-\lambda t}e^{\lambda^2 n}. \qquad \text{(E.2.13)}$$

The proof then concludes as in the scalar case. □

The matrix Bernstein inequality can also be expressed in the following form that we will use in the book.

THEOREM E.8 (Matrix Bernstein Inequality [Tro12]). *Suppose that* $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots$ *are independent random matrices of dimension* $n_1 \times n_2$*, with* $\mathbb{E}[\boldsymbol{W}_j] = \mathbf{0}$*, and* $\|\boldsymbol{W}_j\| \leq R$ *almost surely. Define*

$$\sigma^2 = \max\left\{ \left\| \sum_j \mathbb{E}[\boldsymbol{W}_j \boldsymbol{W}_j^*] \right\|, \left\| \sum_j \mathbb{E}[\boldsymbol{W}_j^* \boldsymbol{W}_j] \right\| \right\}. \tag{E.2.14}$$

*Then*

$$\mathbb{P}\left[ \left\| \sum_j \boldsymbol{W}_j \right\| \geq t \right] \leq (n_1 + n_2) \exp\left( \frac{-t^2/2}{\sigma^2 + Rt/3} \right). \tag{E.2.15}$$

# Bibliography


[AAZB+16] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv:1611.01146v2*, 2016. 418

[AB99] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. 24

[AB09] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, USA, 1st edition, 2009. 197

[ABSS93] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, pages 724–733. IEEE, 1993. 67

[ADG+16] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989, 2016. 537

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014. 297

[AHP06] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006. 479

[AK95] E. Amaldi and V. Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Computer Science*, 147:181–210, 1995. 67

[AK98] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209:237–260, 1998. 67

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 594–598, USA, 1998. Society for Industrial and Applied Mathematics. 198

[ALMT13] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *CoRR*, abs/1303.6672, 2013. 248

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014. 132, 252, 253, 254, 259, 264

[Ame11] Dennis Amelunxen. *Geometric analysis of the condition of the convex feasibility problem.* PhD Thesis, Univ. Paderborn, 2011. 248

[AMS09] Pierre-Antoine Absil, Robert Mahoney, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2009. 290, 470

[ANR74] N. Ahmed, T. Natarajan, and K. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 23(1):90–93, 1974. 40

[ANW10] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010. 471

[ANW12] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–197, 2012. 227, 264

[ARR14] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014. 287, 296

[ARV20] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 575

[AW18] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 486, 533, 555

[AZ17] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017. 361, 362

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019. 271, 536

[B+15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. 470

[B+20] Tom Brown et al. Language models are few-shot learners. *arXiv:2005.14165v4*, 05 2020. 27

[Bal10] Radu V. Balan. On signal reconstruction from its spectrogram. In *Information Sciences and Systems (CISS), 44th Annual Conference on*, pages 1–4. IEEE, 2010. 277

[Bar72] H. B. Barlow. Single unites and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394, 1972. 10

[BB20] Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. *arXiv preprint arXiv:2005.08099*, 2020. 9, 67, 198

[BBC09] S. Becker, J. Bobin, and E. Candès. NESTA: a fast and accurate first-order method for sparse recovery. *preprint*, 2009. 233

[BBE17] Tamir Bendory, Robert Beinert, and Yonina C Eldar. Fourier phase retrieval: Uniqueness and algorithms. In *Compressed Sensing and its Applications*, pages 55–91. Springer, 2017. 299

[BBV16] Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on learning theory*, pages 361–382, 2016. 287

[BC14] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014. 536

[BCE06] Radu Balana, Pete Casazzab, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345 – 356, 2006. 277

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. 537

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008. 87

[BDP+07] Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K. Satapathy, and J. Friso van der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A*, 63(4):306–314, Jul. 2007. 277

[BEGK11] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2011. 405

[Bel73] E. Beltrami. Sulle funzioni bilineari. *Giornale di Mathematiche di Battaglini*, 11:98–106, 1873. 20

[Ber82] D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1982. 202, 335

[Ber03] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2003. 370, 632

[BGM+08] Andrei Belitski, Arthur Gretton, Cesare Magri, Yusuke Murayama, Marcelo A. Montemurro, Nikos K. Logothetis, and Stefano Panzeri. Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *Journal of Neuroscience*, 28(22):5696–5709, 2008. 577

[BGNR17] Bowen Baker, Otkrist Gupta, N. Naik, and R. Raskar. Designing neural network architectures using reinforcement learning. *ArXiv*, abs/1611.02167, 2017. 537

[BGW20] Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. *arXiv preprint arXiv:2008.11245*, 2020. 572

[Bha96] R. Bhatia. *Matrix Analysis*. Springer, 1996. 583

[BHK97] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 476

[Bir11] H. Birkholz. A unifying approach to isotropic and anisotropic total variation denoising models. *Journal of Computational and Applied Mathematics*, pages 2502–2514, 2011. 443

[BJ03] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):218–233, 2003. 42, 199, 207, 476, 504

[BJMO12] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, January 2012. 244, 364

[BJPD17] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017. 537

[BJS19] Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. 302

[BJT19] Albert S Berahas, Majid Jahani, and Martin Takáč. Quasi-Newton methods for deep learning: Forget the past, just sample. *arXiv preprint arXiv:1901.09997*, 2019. 537

[BK96] P.N. Belhumeur and D.J. Kriegman. What is the set of images of an object under all possible lighting conditions? In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 270–277, 1996. 504

[BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 537

[BKLN08] S. C. Blaakmeer, E. A. M. Klumperink, D. M. W. Leenaerts, and B. Nauta. Wideband balun-LNA with simultaneous output balancing, noise-canceling and distortion-canceling. *IEEE J. Solid-State Circuits*, 43(6):1341–1350, June 2008. 456

[BKM67] E. M. L. Beale, M. G. Kendall, and D. W. Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54(3/4):357–366, 1967. 17

[BKM16] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Ann. Statist.*, 44(2):813–852, 04 2016. 17

[BKN04] F. Bruccoleri, E. A. M. Klumperink, and B. Nauta. Wide-band CMOS low-noise amplifier exploiting thermal noise canceling. *IEEE Journal of Solid-State Circuits*, 39(2):275–282, 2004. 456

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015. 300

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. 634, 637

[BLO05] James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005. 270

[BLWY06] Pratik Biswas, Tzu-Chen Lian, Ta-Chung Wang, and Yinyu Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks (TOSN)*, 2(2):188–220, 2006. 285

[BM03] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003. 281

[BM04] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal on Computer Vision*, 56(3):221–255, 2004. 516

[BM13] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013. 557, 561, 573

[BMC+06] Rahim Bagheri, A. Mirzaei, S. Chehrazi, M. E. Heidari, Minjae Lee, M. Mikhemar, Wai Tang, and A. A. Abidi. An 800-MHz-6-GHz software-defined wireless receiver in 90-nm CMOS. *IEEE J. Solid-State Circuits*, 41(12):2860–2876, Dec 2006. 457

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. 141

[BNO03] D. Bertsekas, A. Nedic, and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Nashua, NH, 2003. 309

[BNS16] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016. 285

[BO69] Richard L Bishop and Barrett O'Neill. Manifolds of negative curvature. *Transactions of the American Mathematical Society*, 145:1–49, 1969. 300

[Bos50] R. Boscovich. *De calculo probabilitatum que respondent diversis valoribus summe errorum post plures observationes, quarum singule possient esse erronee certa quadam quantitate*. 1750. 12, 132

[Bot82] Raoul Bott. Lectures on Morse theory, old and new. *Bulletin of the American mathematical society*, 7(2):331–358, 1982. 271, 279

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 361

[Bou] Jean-Yves Bouguet. Camera calibration toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/. 525, 530

[Bou16] Nicolas Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016. 287

[Bou20] Nicolas Boumal. An introduction to optimization on smooth manifolds. Available online, Aug 2020. 419

[BPC+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 339, 363, 633

[BR83] G. Binnig and H. Rohrer. Surface imaging by scanning tunneling microscopy. *Ultramicroscopy*, pages 157–160, 1983. 462

[Bro71] D. C. Brown. Close-range camera calibration. *Photogrammetric Engineering*, 37(8):855–866, 1971. 528

[BRT09] Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. 112

[BSBC00] M. Bertalmio, G. Sapiro, C. Ballester, and V. Caselles. Image inpainting. In *Proceedings of ACM SIGGRAPH Conference*, 2000. 512

[BSFG09] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (SIGGRAPH)*, 28(3), 2009. 513, 514

[BSGF10] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized PatchMatch correspondence algorithm. *European Conference on Computer Vision (ECCV)*, 2010. 513

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Science*, 2(1):183–202, 2009. 233, 327, 328, 330, 363, 470

[BTR12] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *CoRR*, abs/1204.0562, 2012. 264

[BUF07] K. Block, M. Uecker, and J. Frahm. Undersampled radial MRI with multiple coils. iterative image reconstruction using a total variation constraint. *Magnetic Resonance in Medicine*, 57:1086–1098, 2007. 443

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. 27, 301, 309, 310, 613, 628

[BV18] Stephen Boyd and Lieven Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, 2018. 16, 583

[BWY17] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017. 298

[CAB+16] G. Cruz, D. Atkinson, C. Buerger, T. Schaeffter, and C. Prieto. Accelerated motion corrected three-dimensional abdominal MRI using total variation regularized SENSE reconstruction. *Magnetic Resonance in Medicine*, 75:1484–1498, 2016. 443

[CAD+18] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018. 574

[Can06] E. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, 2006. 7, 23

[Can08] E. Candès. The restricted isometry property and its implications for compressed sensing. *Comte Rendus de l'Academie des Sciences, Paris, Serie I*, 346:589–592, 2008. 87, 92, 132

[Cau47] Augustin Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25:536–538, 1847. 54, 268, 372, 623

[CB19] Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In *Advances in Neural Information Processing Systems*, pages 5987–5997, 2019. 301

[CC17] Yuxin Chen and Emmanuel J Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on pure and applied mathematics*, 70(5):822–883, 2017. 280

[CCD+19] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019. 286, 302

[CCFM18] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:1–33, 03 2018. 301, 413

[CCS08] J. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. preprint, 2008. [Online]. Available: http://arxiv.org/abs/0810.3286, 2008. 232

[CD91] M. Frank Callier and A. Charles Desoer. *Linear System Theory*. Springer-Verlag, 1991. 1, 2

[CD13] E. Candès and M. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013. 113

[CD16] Yair Carmon and John C. Duchi. Gradient descent efficiently finds the cubic-regularized non-convex Newton step. *arXiv:1612.00547*, 2016. 382

[CD19] Yair Carmon and John Duchi. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM Journal on Optimization*, 29(3):2146–2178, 2019. 382

[CDDD19] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019. 302

[CDHS17] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017. 390, 418

[CDHS18] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization, and arXiv:1611.00756v1*, 28:1751–1772, November 2018. 418

[CDS98] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition for basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998. 18

[CDS01] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001. 107, 362, 484

[CESV13] Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1), 2013. 275, 277

[CF80] T. M. Cannon and E. E. Fenimore. Coded Aperture Imaging: Many Holes Make Light Work. *Optical Engineering*, 19(3):283 – 289, 1980. 442

[CGT00] Andrew R. Conn, Nicholas I.M. Gould, and Philippe L. Toint. *Trust region methods*, volume 1. SIAM, 2000. 271, 300, 418, 420

[CGW19] Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant CNNs on homogeneous spaces. In *Advances in Neural Information Processing Systems*, pages 9142–9153, 2019. 555

[CHC08] G. Vogiatzis C. Hernández and R. Cipolla. Multi-view photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008. 504, 505

[Che95] S. Chen. *Basis Pursuit*. PhD thesis, Stanford University, Stanford, CA, 1995. 362

[Che13] Yudong Chen. Incoherence-optimal matrix completion. *Information Theory, IEEE Transactions on*, 61, 10 2013. 188

[CHM+14] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surface of multilayer networks. *arXiv preprint arXiv:1412.0233*, 2014. 271

[Cho17] Franois Chollet. Xception: Deep learning with depthwise separable convolutions. *preprint arXiv:1610.02357*, 2017. 537, 572

[CHS87] T.S. Chiang, C.R. Hwang, and S.J. Sheu. Diffusions for global optimization in $\mathbb{R}^n$. *SIAM Journal Control and Optimization*, 25:737–752, 1987. 403

[CJG+15] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. PCANet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015. 474, 557, 559

[CJSC13] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013. 182, 222

[CLC19] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. 28, 265, 269, 271, 272, 275, 282, 286, 299

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011. 24, 286, 300

[CLS15a] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015. 277, 280

[CLS15b] Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. 275, 278, 280

[CM73] J. F. Claerbout and F. Muir. Robust modeling of erratic data. *Geophysics*, 38(5):826–844, 1973. 15

[CMH+17] Jacopo Cavazza, Pietro Morerio, Benjamin D. Haeffele, Connor Lane, Vittorio Murino, and René Vidal. Dropout as a low-rank regularizer for matrix factorization. *CoRR*, abs/1710.05092, 2017. 317

[CMH+18] Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a low-rank regularizer for matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 435–444, 2018. 537

[CMM+20] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. 575

[CMP10] Anwei Chai, Miguel Moscoso, and George Papanicolaou. Array imaging using intensity-only measurements. *Inverse Problems*, 27(1):015005, 2010. 277

[Cor06] John V. Corbett. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 57(1):53–68, 2006. 277

[CP86] Thomas F Coleman and Alex Pothen. The null space problem I. complexity. *SIAM Journal on Algebraic Discrete Methods*, 7(4):527–537, 1986. 67

[CP10] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. 188

[CP11] E. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *Annals of Statistics*, 2011. 164, 165

[CPW12] V. Chandrasekaran, P. Parrilo, and A. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012. 9

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009. 140, 284, 285, 300

[Cro78] L. Cromme. Strong uniqueness: A far-reaching criterion for the convergence analysis of iterative procedures. *Numerishe Mathematik*, 29:179–193, 1978. 517

[CRPW12] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundation of Computational Mathematics*, 12(6):805–849, 2012. 264

[CRT06a] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. 87, 103

[CRT06b] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Math*, 59(8):1207–1223, 2006. 23, 92, 108, 114, 132

[CSD+09] V. Cevher, A. Sankaranarayanan, M. Duarte, D. Reddy, R. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2009. 205

[CSL+20] Sky Cheung, John Shin, Yenson Lau, Zhengyu Chen, Ju Sun, Yuqian Zhang, Marvin Mller, Ilya Eremin, John Wright, and Abhay Pasupathy. Dictionary learning in Fourier-transform scanning tunneling spectroscopy. *Nature Communications*, 11:1081, 02 2020. 294, 462, 464, 465, 471

[CSPW09] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Sparse and low-rank matrix decompositions. In *IFAC Symposium on System Identification*, 2009. 222, 232

[CSV13] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. 275, 277, 300

[CT81] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *SIGGRAPH Comput. Graph.*, 15(3):307–316, 1981. 494, 498

[CT91] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, 1991. 22, 541, 544

[CT05] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. 15, 23, 87, 91, 132, 574

[CT07] Emmanuel Candès and Terence Tao. The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *The Annals of Statistics*, pages 2313–2351, 2007. 112

[CT09] E. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2009. 174

[CT17] Le Chang and Doris Tsao. The code for facial identity in the primate brain. *Cell*, 169:1013–1028.e14, 06 2017. xi, 538, 577

[CW98] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998. 462

[CW05] P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4:1168–1200, 2005. 363

[CW16] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016. 555, 561, 573

[CYY+20] Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Deep networks from the principle of rate reduction. *arXiv preprint arXiv:2010.14765*, 2020. 560, 561, 572

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014. 361, 362

[DD18] Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *arXiv preprint arXiv:1810.07590*, 2018. 302

[DDF+90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R.Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. 141

[DDM04] I. Daubechies, M. Defrise, and C. Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Math*, 57:1413–1457, 2004. 363

[DDMP18] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018. 302

[DDP17] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017. 280

[DDT+08] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008. 442

[DE03] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell^1$ minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2197–2202, March 2003. 132

[DF87] Chris Dainty and James R. Fienup. Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, pages 231–275, 1987. 277

[DFL+88] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI*, pages 281–286, 1988. 141, 199

[DIIM04] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, SCG '04, pages 253–262, New York, NY, USA, 2004. ACM. 96

[DJL+17] Simon Du, Chi Jin, Jason Lee, Michael Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. 291, 301, 400, 412

[DLL+19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019. 271, 536

[DMA97] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Journal of Constructive Approximation*, 13:57–98, 1997. 67

[DO19] Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019. 331

[Don00] D. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 2000. [Online]. Available: http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html, 2000. 21

[Don05] David L Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. *Stanford Technical Report 2005-04*, 2005. 132, 264

[Don06a] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 7, 23

[Don06b] D. Donoho. For most large underdetermined systems of linear equations the minimal $\ell_1$-norm solution is also the sparsest solution. *Communications on Pure and Applied Math*, 59(6):797–829, 2006. 23, 87

[DR16] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016. 281, 282, 284, 285

[DR19] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019. 302

[DS07] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007. 298

[DT09] David Donoho and Jared Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009. 26, 132, 264

[DT10] David L Donoho and Jared Tanner. Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling. *IEEE Transactions on Information Theory*, 56(4):2002–2016, 2010. 26, 132, 264

[DTZ16] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. *arXiv preprint arXiv:1609.00368*, 2016. 298

[DY16] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66:889–916, 2016. 340, 363

[EA03] Chris Eliasmith and Charles Anderson. *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*. Cambridge, MA, 01 2003. 577

[EA06] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006. 292

[EB92] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992. 363

[Eck12] J. Eckstein. Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Technical Report*, 2012. 335, 341, 363

[Efr66] M. Efroymson. Stepwise regression a backward and forward look. In *Eastern Regional Meetings of the Institute of Mathematical Statistics*, 1966. 17

[Ela10] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010. 291

[EMS18] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global nonconvex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9671–9680, 2018. 270

[ESQD05] M. Elad, J. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Applied and Computational Harmonic Analysis*, 19:340–358, 2005. 257, 484

[ETT+17] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017. 486, 533, 555

[Ext] http://vision.ucsd.edu/∼leekc/extyaledatabase/extyaleb.html. 235

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. 20

[FB81] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–385, 1981. 504

[FHB01] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference (ACC)*, 2001. 2, 188

[FHB03] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 3, pages 2156–2162. IEEE, 2003. 301, 542

[FHB04] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *American Control Conference*, 2004. 146, 188, 232

[Fie87] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America A*, 4(12), 1987. 10

[Fie13] James R Fienup. Phase retrieval algorithms: a personal tour. *Applied optics*, 52(1):45–56, 2013. 277

[FJZT17] Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks: (almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017. 299

[FKT15] Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *Conference on Learning Theory*, pages 696–709, 2015. 67

[FlSM09] M. J. Fadili, J. l. Starck, and F. Murtagh. Inpainting and zooming using sparse representations. *The Computer Journal*, pages 64–79, 2009. 512

[FLZZ20] Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical Foundations of Data Science*. CRC Press, 2020. xiii, 16, 31

[FMR16] S. Fusi, E. Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, 2016. 557

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhauser, Springer, 2013. xiii

[FS20] Albert Fannjiang and Thomas Strohmer. The numerics of phase retrieval. *arXiv preprint arXiv: 2004.05788*, 2020. 275, 277, 280

[Fuc04] J. Fuchs. On sparse representation in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004. 132

[FW56] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956. 351, 362, 630

[FZS21] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv:2101.03961*, 01 2021. 27, 537, 553, 554

[Gab78] K. R. Gabriel. Least squares approximation of matrices by additive and multiplicative models. *J. R. Statist. Soc. B*, 40:186–196, 1978. 20

[Gau09] C.F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Carl Friedrich Gauss Werke. sumtibus F. Perthes et I. H. Besser, 1809. 13

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 537

[GB14] M. Grant and S. Boyd. CVX: MATLAB software for disciplined convex programming (web page and software). 2009. [Online]. Available: http://stanford.edu/∼boyd/cvx, 2014. 202, 232

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 24, 268, 535

[GBK01] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001. 207, 208, 477, 481, 482, 504, 507

[GBLJ19] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3196–3206, 2019. 574

[GBW19] Dar Gilboa, Sam Buchanan, and John Wright. Efficient dictionary learning with gradient descent. *ICML*, 2019. 273, 275, 301, 413

[GCW18] Cristina Garcia-Cardona and Brendt Wohlberg. Convolutional dictionary learning: A comparative review and new algorithms. *IEEE Transactions on Computational Imaging*, 4(3):366–381, 2018. 296

[GEBS18] R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro. Tradeoffs between convergence speed and reconstruction accuracy in inverse problems. *IEEE Transactions on Signal Processing*, 66(7):1676–1690, 2018. 554, 576

[GH86] S. Geman and C.R. Hwang. Diffusions for global optimization. *SIAM Journal Control and Optimization*, 24:1031–1043, 1986. 403

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015. 271, 275, 287, 297, 301

[GHY14] Guoyong Gu, Bingsheng He, and Xiaoming Yuan. Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: a unified approach. *Computational Optimization and Applications*, 59(1-2):135–161, 2014. 341

[GJ79] Michael R Garey and David S Johnson. *Computers and Intractability*. W. H. Freeman, 1979. 52

[GJ90] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990. 52

[GJP95] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995. 537

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1233–1242, 2017. 275, 282

[GK12] Christine Grienberger and Arthur Konnerth. Imaging calcium in neurons. *Neuron*, 73(5):862–885, 2012. 462

[GKXS18] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *International Conference on Learning Representations*, 2018. 537

[GL10] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 399–406, 2010. 537, 554, 561

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016. 275, 285

[GLM17] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017. 299

[GLSS18] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018. 538, 574

[GM90] Saul B. Gelfand and Sanjoy K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. Technical Report LIDS-P-1937, Massachusetts Institute of Technology, 1990. 403

[GM09] D. Goldfarb and S. Ma. Convergence of fixed point continuation algorithms for matrix rank minimization. *preprint*, 2009. 204, 232

[GM17] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. *Advances in Neural Information Processing Systems*, 2017. 275, 287, 298

[GMOV18] Weihao Gao, Ashok Vardhan Makkuva, Sewoong Oh, and Pramod Viswanath. Learning one-hidden-layer neural networks under general input distributions. *arXiv preprint arXiv:1810.04133*, 2018. 299

[GN03] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(13), 2003. 132

[GN16] Lee-Ad Gottlieb and Tyler Neylon. Matrix sparsification and the sparse null space problem. *Algorithmica*, 76(2):426–444, 2016. 67

[Gol65] Sidney Golden. Lower bounds for the Helmholtz function. *Physical Review*, 137(4B):B1127, 1965. 638

[Gol67] R. Gold. Optimal binary sequences for spread spectrum multiplexing (corresp.). *IEEE Trans. Inf. Theory*, 13(4):619–621, October 1967. 457

[Gol80] Donald Goldfarb. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical programming*, 18(1):31–40, 1980. 271, 300, 385

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 268

[Gro10] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 2010. 188

[GS12] Surya Ganguli and Haim Sompolinsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual review of neuroscience*, 35:485–508, 2012. 11

[GV96] G. Golub and C. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3 edition, 1996. 583

[GVR12] Laurent Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *Pacific Journal of Optimization*, 8:667–698, 2012. 363

[GWL+10] A. Ganesh, J. Wright, X. Li, E. Candès, and Y. Ma. Dense error correction for low-rank matrices via principal component pursuit. In *International Symposium on Information Theory (ISIT)*, 2010. 222, 502

[GZ19] David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *ArXiv*, abs/1904.07174, 2019. 9, 198

[Haj90] Prabhat Hajela. Genetic search-an approach to the nonconvex optimization problem. *AIAA journal*, 28(7):1205–1210, 1990. 270

[Hay94a] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of Optical Society of America A*, 11(11):3079–3089, 1994. 505

[Hay94b] Simon S Haykin. *Blind deconvolution*. Prentice Hall, 1994. 464

[HBZ+17] T. Haque, Mathew Bajor, Yudong Zhang, Jianxun Zhu, Zarion Jacobs, Robert Kettlewell, J. Wright, and Peter R. Kinget. A direct RF-to-information converter for reception and wideband interferer detection employing pseudo-random LO modulation. In *IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, 2017. 460

[HCL06] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, pages 1735–1742, 2006. 546

[Hes69] Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969. 333, 363, 631

[HFLM+18] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 544

[HFW+19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 546

[HH08] F. Herrmann and G. Hennenfent. Non-parametric seismic data recovery with curvelet frames. *Geophysical Journal International*, 173(1):233–248, 2008. 66

[HHS17] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017. 537

[HJ85] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge Press, 1985. 583

[HKSS15] David J Herzfeld, Yoshiko Kojima, Robijanto Soetedjo, and Reza Shadmehr. Encoding of action by the purkinje cells of the cerebellum. *Nature*, 526(7573):439, 2015. 11

[HKSS18] David J Herzfeld, Yoshiko Kojima, Robijanto Soetedjo, and Reza Shadmehr. Encoding of error and learning to correct that error by the purkinje cells of the cerebellum. *Nature neuroscience*, 21(5):736, 2018. 11

[HKV19] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors. *Automatic Machine Learning: Methods, Systems, Challenges*. Springer, 2019. 537

[HKW11] Geoffrey E. Hinton, A. Krizhevsky, and S. Wang. Transforming auto-encoders. In *ICANN*, 2011. 573

[HL67] R. R. Hocking and R. N. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967. 17

[HL13] Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013. 240, 297

[HL16] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML?16, page 1263?1271. JMLR.org, 2016. 362

[HLVDMW17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 578

[HLWY19] Jiang Hu, Xin Liu, Zaiwen Wen, and Yaxiang Yuan. A brief introduction to manifold optimization. *arXiv preprint arXiv:1906.05450*, 2019. 271, 302

[HMH00] L. Hubert, J. Meulman, and W. Heiser. Two purposes for matrix factorization: a historical appraisal. *SIAM Review*, 42(1):68–82, 2000. 20

[HMW13] Teiko Heinosaari, Luca Mazzarella, and Michael M. Wolf. Quantum tomography under prior information. *Communications in Mathematical Physics*, 318(2):355–374, 2013. 277

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999. 141

[Hof04] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22(1):89–115, 2004. 141, 200

[Hol07] J.K. Holmes. *Spread Spectrum Systems for GNSS and Wireless Communications*. Artech House, 2007. 458

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933. 18, 20, 145

[HRRS86] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust Statistics - The Approach Based on Influence Functions*. Wiley, New York, NY, 1986. 15

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009. 16, 31

[HTT09] Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, October 2009. x

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall, CRC, 2015. xiii, 574

[Hub81] P. Huber. *Robust statistics*. John Wiley & Sons, 1981. 15

[Hub92] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992. 286

[HV17] Benjamin D Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339, 2017. 299

[HXP20] Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2020. 537

[HY38] A. S. Householder and G. Young. Matrix approximation and latent roots. *America Math. Mon.*, 45:165–171, 1938. 20

[HY01] M.H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of American Statistical Association*, 96:746–774, 2001. 17

[HY12] Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012. 341, 343, 363

[HYP+15] T. Haque, Rabia Tugce Yazicigil, Kyle Jung-Lin Pan, J. Wright, and Peter R. Kinget. Theory and design of a quadrature analog-to-information converter for energy-efficient wideband spectrum sensing. *IEEE Trans. Circuits Syst. I*, 62(2):527–535, Feb 2015. 450, 453

[HYZ08] E. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for $\ell^1$-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008. 232, 471

[HZ00] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2000. 488, 527

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 537

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 536, 537, 547, 552, 554

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 537, 543

[JCM12] K. Jia, T.-H. Chan, and Y. Ma. Robust and practical face recognition via structured sparsity. In *Proceedings of the European Conference on Computer Vision*, 2012. 239

[JEH15] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*, 2015. 277, 280

[JEH16] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. STFT phase retrieval: Uniqueness guarantees and recovery algorithms. *IEEE Journal of selected topics in signal processing*, 10(4):770–781, 2016. 278

[JEH17] Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *Optical Compressive Imaging*, pages 263–296, 2017. 267

[JG15] Tyler B. Johnson and Carlos Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *ICML*, 2015. 363

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv:1806.07572*, June 2018. 569

[JGKA19] Majid Janzamin, Rong Ge, Jean Kossaifi, and Anima Anandkumar. Spectral learning on matrices and tensors. *Foundations and Trends in Machine Learning*, 12(5-6):393–536, 2019. 297

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *34th International Conference on Machine Learning, ICML 2017*, pages 2727–2752. International Machine Learning Society (IMLS), 2017. 271, 301

[JK+17] Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017. 265, 269, 271, 272, 299

[JMFU17] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. 537

[JNJ18] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085, 2018. 271, 301, 411, 418, 470

[JNRS10] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553, 2010. 416

[JO80] K. Jittorntrum and M. Osborne. Strong uniqueness and second order convergence in nonlinear discrete approximation. *Numerische Mathematik*, 34:439–455, 1980. 517

[Jol86] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, NY, 1986. 20, 196, 504

[Jol02] I. Jollife. *Principal Component Analysis*. Springer-Verlag, 2nd edition, 2002. 196

[Jor74] M.C. Jordan. Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliqués*, 19:35–54, 1874. 20

[Jor97] Michael I. Jordan. Serial order: A parallel distributed processing approach. In *Advances in Psychology*, volume 121, pages 471–495, 1997. 3

[Jor03] M. Jordan. *An Introduction to Probabilistic Graphical Models*. unpublished, 2003. 8

[JSA15] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015. 299

[JSZK15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. 573

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013. 361

[JZB+16] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in neural information processing systems*, pages 4116–4124, 2016. 298

[Kar72] Richard M Karp. *Reducibility among Combinatorial Problems*. Springer, 1972. 52

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016. 271, 287

[KB09] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 240, 261, 297

[KB14] Diederik P. Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 537, 626

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015. 331, 627

[KBB16] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Adaptive averaging in accelerated descent dynamics. In *Advances in Neural Information Processing Systems*, pages 2991–2999, 2016. 331, 627

[KBRW19] Michael R Kellman, Emrah Bostan, Nicole A Repina, and Laura Waller. Physics-based learned design: optimized coded-illumination for quantitative phase imaging. *IEEE Transactions on Computational Imaging*, 5(3):344–353, 2019. 278

[KCYT05] J. Kim, J. Choi, J. Yi, and M. Turk. Effective representation using ICA for face recognition robust to local distortion and partial occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1977–1981, 2005. 479, 482

[KDSA+15] Rajiv Kumar, Curt Da Silva, Okan Akalin, Aleksandr Y Aravkin, Hassan Mansour, Benjamin Recht, and Felix J Herrmann. Efficient matrix completion for seismic data reconstruction. *Geophysics*, 80(5):V97–V114, 2015. 285

[KGV83] S. Kirkpatrick, C.D. Gelett, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:621–630, 1983. 403

[KH92] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992. 536

[KH96] Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *Signal Processing Magazine, IEEE*, 13(3):43–64, May 1996. 464

[KKP+18] Anastasios Kyrillidis, Amir Kalev, Dohyung Park, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable compressed sensing quantum state tomography via non-convex methods. *npj Quantum Information*, 4(1):1–7, 2018. 271

[KM89] S. Kontogiorgis and R. Meyer. A variable-penalty alternating direction method for convex optimization. *Mathematical Programming*, 83:29–53, 1989. 363

[KMK97] D. C. Knill, P. Mamassian, and D. Kersten. The geometry of shadows. *Journal of Optical Society of America A*, 14(12):3216–3232, 1997. 492

[KMR14] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the Restricted Isometry Property. *Communications on Pure and Applied Mathematics*, 67(11), 2014. 105

[Koo31] B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proc. National Academy of Science, USA*, 17:315–318, 1931. 2

[Kor09] Yehuda Koren. The Bellkor solution to the Netflix grand prize. 2009. 285

[KPCC15] Zhao Kang, Chong Peng, Jie Cheng, and Qiang Cheng. Logdet rank minimization with application to subspace clustering. *Computational Intelligence and Neuroscience*, 2015, 2015. 542, 574

[KQC+19] Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pages 2055–2110, 2019. 271, 298

[Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images. *online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf*, 2009. 547

[Kru77] Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977. 49

[KS12] Irwin Kra and Santiago R Simanca. On circulant matrices. *Notices of the American Mathematical Society*, 59:368–377, 2012. 556, 597

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 24, 533, 537, 561

[Kuč95] Luděk Kučera. Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57(2?3):193?212, February 1995. 198

[Kus87] H.J. Kushner. Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo. *SIAM Journal Applied Mathematics*, 47:165–189, 1987. 403

[KW92] J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992. 387, 388

[KZLW19] Han-Wen Kuo, Yuqian Zhang, Yenson Lau, and John Wright. Geometry and symmetry in short-and-sparse deconvolution. In *International Conference on Machine Learning (ICML)*, June 2019. 275, 287, 295, 465, 466, 467, 471

[Lan67] H. J. Landau. Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Math.*, 117:37–52, 1967. 446

[Lan01] S. Lang. *Fundamentals of Differential Geometry*. Springer-Verlag, 2001. 412

[Lap74] P. Laplace. Memoire sur la probabilite des causes par les evenemens. *Memoires de Mthematique et de Physique, Presentes a l'Academie Royale des Sciences par divers Savans & lus dans ses Assemblees, Tome Sixieme*, pages 621–656, 1774. 13, 132

[LB95a] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. In *The handbook of brain theory and neural networks*. MIT Press, 1995. 268, 561

[LB95b] MH Loke and RD Barker. Least-squares deconvolution of apparent resistivity pseudosections. *Geophysics*, 60(6):1682–1690, 1995. 464

[LB00] A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118, 2000. 475, 480

[LB18] Yanjun Li and Yoram Bresler. Global geometry of multichannel sparse blind deconvolution on the sphere. *arXiv preprint arXiv:1805.10437*, 2018. 275, 287, 296

[LB19] Yanjun Li and Yoram Bresler. Multichannel sparse blind deconvolution on the sphere. *IEEE Transactions on Information Theory*, 65(11):7415–7436, 2019. 557, 558

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 537

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 24, 533

[LBOM12] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012. 537

[LCBD20] Tianlin Liu, Anadi Chaman, David Belius, and Ivan Dokmani. Interpreting U-Nets via task-driven multiscale dictionary learning. *arXiv:2011.12815*, 2020. 537

[LCD+19] Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man Cho So. Nonsmooth optimization over Stiefel manifold: Riemannian subgradient methods. *arXiv preprint arXiv:1911.05047*, 2019. 288, 301, 302

[LCWM09] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical report, arXiv:1009.5055, 2009. 233

[LCWY19] Jialin Liu, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. ALISTA: Analytic weights are as good as learned weights in LISTA. In *International Conference on Learning Representations*, 2019. 537

[LDP07] M. Lustig, D. Donoho, and J. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. 7, 66, 439, 442

[LDSP08] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing MRI. *Signal Processing Magazine, IEEE*, 25(2):72–82, 2008. 66, 425, 437, 442

[LeC98] Yann LeCun. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998. 563, 579

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon, Mathematical Surveys and Monographs 89*. American Mathematical Society, Providence, RI, 2001. 81, 637

[Leg05] A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Nineteenth Century Collections Online (NCCO): Science, Technology, and Medicine: 1780-1925. F. Didot, 1805. 13

[Lev44] K. Levenberg. A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944. 418

[LGW+09] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2009. 233

[LH16] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 537

[LHGT04] L. Li, W. Huang, I. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004. 206

[LHZC01] S. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2001. 479, 482

[Li13] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013. 230

[Lib12] Daniel Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2012. 576

[Liu11] Y. K. Liu. Universal low-rank matrix recovery from Pauli measurements. *Proceedings of NIPS*, 2011. 161

[LJ16] Simon Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *eprint arXiv:1607.00345*, 07 2016. 351, 353

[LJB+95] Yann LeCun, L.D. Jackel, Leon Bottou, Corinna Cortes, J. S. Denker, Harris Drucker, I. Guyon, U.A. Muller, Eduard Sackinger, Patrice Simard, and V. Vapnik. Learning algorithms for classification: A comparison on handwritten digit recognition. In J.H. Oh, C. Kwon, and S. Cho, editors, *Neural networks*, pages 261–276. World Scientific, 1995. 561

[LK81] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Workshop*, 1981. 516

[LKB18] B. Lusch, J. N. Kutz, and S. L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. 2

[LKG+19] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. PlaneRCNN: 3D plane detection and reconstruction from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4445–4454, 2019. 531

[LLA+19] Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019. 282, 285

[LLB16] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016. 287

[LLT18] B. M. Lake, N. Lawrence, and J. Tenenbaum. The emergence of organizing structure in conceptual representation. *Cognitive science*, 42 Suppl 3:809–832, 2018. 11

[LLY+13] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1):171–184, Jan 2013. 234, 367

[LM79] P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16:964–979, 1979. 363

[LM16] Ke Li and Jitendra Malik. Fast k-nearest neighbour search via dynamic continuous indexing. In *Proceedings of International Conference on Machine Learning*, 2016. 96

[LM18] Gilad Lerman and Tyler Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018. 286

[LMH15] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in neural information processing systems*, pages 3384–3392, 2015. 361

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018. 538

[LNC+11] Quoc V Le, Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, and Andrew Y Ng. On optimization methods for deep learning. In *ICML*, 2011. 537

[Log65] B. Logan. *Properties of High-Pass Signals*. PhD thesis, Columbia University, 1965. 132

[LPP+17] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017. 271

[LPP+19] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1-2):311–337, 2019. 271, 301

[LPW+17] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pages 6231–6239, 2017. 536

[LQK+19] Yenson Lau, Qing Qu, Han-Wen Kuo, Pengcheng Zhou, Yuqian Zhang, and John Wright. Short-and-sparse deconvolution – a geometric approach. *arXiv preprint arXiv:1908.10959*, 2019. 275, 287, 294, 297

[LQMS18] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. OLE: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118, 2018. 545, 546

[LRZM12] X. Liang, X. Ren, Z. Zhang, and Y. Ma. Repairing sparse low-rank texture. In *Proceedings of the European Conference on Computer Vision*, 2012. 511, 531

[LS17] Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017. 275, 287

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016. 271, 301

[LSPJ18] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. *arXiv preprint arXiv:1808.07217*, 2018. 537

[Lus13] M. Lustig. Compressed sensing MRI resources. `http://www.eecs.berkeley.edu/~mlustig/CS.html`, 2013. 37, 442

[LV09] Z. Liu and L. Vandenberghe. Semidefinite programming methods for system realization and identification. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 4676–4681, Dec 2009. 2

[LV10] Zhang. Liu and Lieven. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2010. 2

[LVB+93] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993. 479

[LWDF11] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding blind deconvolution algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2354–2367, 2011. 462

[LXB18] Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *arXiv preprint arXiv:1809.11083*, 2018. 287

[LZMCSV20] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020. 286, 302

[LZT18] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2018. 285

[MA89a] R. Monteiro and I. Adler. Interior path following primal-dual algorithms. Part I: Linear programming. *Mathematical Programming*, 44:27–41, 1989. 27

[MA89b] R. Monteiro and I. Adler. Interior path following primal-dual algorithms. Part II: Convex quadratic programming. *Mathematical Programming*, 44:43–66, 1989. 27

[Mar63] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11:431–441, 1963. 418

[Mar02] A. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):748–763, 2002. 475, 480

[Mar14] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014. 537

[Mat02] J. Matousek. *Lectures on Discrete Geometry*. Springer, 2002. 21, 26, 45, 81

[MAV18] Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. *arXiv preprint arXiv:1806.09777*, 2018. 537

[MB98] A. Martinez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center, Universitat Autonoma de Barcelona, Barcelona, Spain, 1998. 475

[MC89] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 576

[McC83] S. McCormick. *A combinatorial approach to some sparse matrix problems*. PhD thesis, Stanford University, 1983. 67

[MDHW07] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9), 2007. 304, 541, 542, 544, 574

[MDL+09] A. Mirzaei, H. Darabi, J.C. Leete, Xinyu Chen, K. Juan, and A. Yazdi. Analysis and optimization of current-driven passive mixers in narrowband direct-conversion receivers. *IEEE J. Solid-State Circuits*, 44(10):2678–2688, Oct 2009. 457

[ME10] M. Mishali and Yonina C. Eldar. From Theory to Practice: Sub-Nyquist Sampling of Sparse Wideband Analog Signals. *IEEE J. Sel. Topics Signal Process.*, 4(2):375–391, 2010. 7, 23, 447, 450

[ME11] M. Mishali and Yonina C. Eldar. Wideband Spectrum Sensing at Sub-Nyquist Rates. *IEEE Signal Processing Magazine*, 28(4):102–135, 2011. 447

[Meg89] N. Megiddo. Pathways to the optimal set in linear programming. In *Progress in Mathematical Programming: Interior-Point and Related Methods*, pages 131–158, 1989. 27

[MES08] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *Image Processing, IEEE Transactions on*, 17(1):53–69, 2008. 40, 41, 66, 512

[MG15] James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417, 2015. 537

[MHI10] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi. Median photometric stereo as applied to the Segonko tumulus and museum objects. *International Journal on Computer Vision*, 86(2):229–242, 2010. 502, 504, 505

[MHSD15] Najib J. Majaj, Ha Hong, Ethan A. Solomon, and James J. DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015. 577

[MHWG13] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013. 262, 265

[MIJ+02] Jianwei Miao, Tetsuya Ishikawa, Bart Johnson, Erik H Anderson, Barry Lai, and Keith O Hodgson. High resolution 3D X-ray diffraction microscopy. *Physical Review Letters*, 89(8):088303, 2002. 277

[Mil63] John Willard Milnor. *Morse theory*, volume 1. Princeton University Press, 1963. 271, 279

[Mil90] R. P. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America A*, 7(3):394–411, Mar 1990. 277

[MIS07] Yasuhiro Mukaigawa, Yasunori Ishii, and Takeshi Shakunaga. Analysis of photometric factors based on photometric linearization. *Journal of Optical Society of America A*, 24(10):3326–3334, 2007. 504

[MJU17] Michael T McCann, Kyong Hwan Jin, and Michael Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017. 537

[MK87] Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987. 269

[MKD06] Joseph F Murray and Kenneth Kreutz-Delgado. Learning sparse overcomplete codes for images. *Journal of VLSI signal processing systems for signal, image and video technology*, 45(1-2):97–110, 2006. 292

[MKKY18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 537

[ML18] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018. 537

[MLE19] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *arXiv preprint arXiv:1912.10557*, 2019. 537, 554, 561

[MM18] Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018. 299

[MMMO17] Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto I Oliveira. Solving SDPs for synchronization and MaxCut problems via the Grothendieck inequality. *arXiv preprint arXiv:1703.08729*, 2017. 287

[MMMS01] Y. Mukaigawa, H. Miyaki, S. Mihashi, and T. Shakunaga. Photometric image-based rendering for image generation in arbitrary illumination. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 652–659, 2001. 504

[Mor62] J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.*, 255:2897–2899, 1962. 362

[Mor78] Jorge J. Moré. The Levenberg-Marquardt algorithm: Implementation and theory. In G. A. Watson, editor, *Numerical Analysis*, pages 105–116, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. 418

[MP43] W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *the Bulletin of Mathematical Biology*, 5:115–133, 1943. 24

[MP97] M. Mesbahi and G. P. Papavassilopoulos. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Transactions on Automatic Control*, 42(2):239–243, Feb 1997. 188

[MS81] F. MacWilliams and N. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam, Netherlands, 1981. 478

[MSG18] Mathurin Massias, Joseph Salmon, and Alexandre Gramfort. Celer: a fast solver for the lasso with dual extrapolation. In *ICML*, 2018. 363

[MSKS04] Y. Ma, S. Soatto, J. Košecká, and S. Sastry. *An Invitation to 3-D Vision: From Images to Models*. Springer-Verlag, New York, 2004. 4, 488, 489, 520, 527

[MSM07] Meena Mahajan and Jayalal Sarma M.N. On the complexity of matrix rank and rigidity. In Volker Diekert, Mikhail V. Volkov, and Andrei Voronkov, editors, *Computer Science – Theory and Applications*, pages 269–280, Berlin, Heidelberg, 2007. Springer. 197

[MWCC18] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3345–3354. PMLR, 10–15 Jul 2018. 29, 280, 574

[MYW+10] K. Min, L. Yang, J. Wright, L. Wu, X. Hua, and Y. Ma. Compact projection: Simple and efficient near neighbor search with practical memory requirements. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3477–3484, June 2010. 96, 97

[MYZC08] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty. An efficient algorithm for compressed MR imaging using total variation and wavelets. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2008. 438, 439, 442

[MZ93] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. 357, 362

[MZWM10] Kerui Min, Zhengdong Zhang, John Wright, and Yi Ma. Decomposing background topics from keywords by principal component pursuit. In *CIKM*, 2010. 141, 199

[MZYM11] H. Mobahi, Z. Zhou, A. Yang, and Y. Ma. Holistic 3D reconstruction of urban structures from low-rank textures. In *ICCV Workshop on 3D Representation and Recognition*, 2011. 531

[N+18] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018. 309

[Nat95] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, 24(2):227–243, 1995. 67

[NDEG13] S. Nam, M.E. Davies, M. Elad, and R. Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30 – 56, 2013. 559

[Nem95] A. Nemirovski. *Information-Based Complexity for Convex Programming*. Lecture Notes, 1995. 311, 324, 613, 629

[Nem04] Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. 364

[Nem07] A. Nemirovski. *Efficient Methods for Convex Optimization*. Lecture Notes, 2007. 27, 311, 613, 629

[Nes83] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983. 27, 233, 324, 363

[Nes00] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000. 269

[Nes03] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2003. 27, 312, 324, 330, 364, 609, 613, 625, 629

[Nes05] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005. 233

[Nes07] Y. Nesterov. Gradient methods for minimizing composite objective function. *ECORE Discussion Paper*, 2007. 233

[NFGS15] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 811–819, Cambridge, MA, USA, 2015. MIT Press. 363

[NIGM18] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018. 536

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. 271, 300, 378, 379, 382

[NT09] D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009. 360, 362

[NW06] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, New York, 2rd edition, 2006. 398, 593

[NWMS18] Elias Nehme, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica*, 5(4):458–464, 2018. 537

[OF96a] B. Olshausen and D. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 1996. 11

[OF96b] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996. 11, 577

[OF97] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997. 10, 11, 474

[OF04] B. Olshausen and D. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14:481–487, 2004. 10

[OH10] Samet Oymak and Babak Hassibi. New null space results and recovery thresholds for matrix rank minimization. *arXiv preprint arXiv:1011.6326*, 2010. 264

[OJB+20] Gregory Ongie, Ajil Jalal, Christopher A Metzler Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 2020. 537

[OJF+15] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015. 261, 265

[OJFH13] Samet Oymak, Amin Jalali, Maryam Fazel, and Babak Hassibi. Noisy estimation of simultaneously structured models: Limitations of convex relaxation. In *IEEE Conference on Decision and Control*, pages 6019–6024, 12 2013. 261, 265

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 546

[OSB99] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing*. Prentice Hall, 1999. 2, 5, 6, 446

[OX18] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *arXiv preprint arXiv:1808.02901*, 2018. 349, 364

[Pat34] Arthur Lindo Patterson. A Fourier series method for the determination of the components of interatomic distances in crystals. *Physical Review*, 46(5):372, 1934. 267

[Pat44] A Lindo Patterson. Ambiguities in the X-ray analysis of crystal structures. *Physical Review*, 65(5-6):195, 1944. 267

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901. 18, 20, 21, 145

[Pea00] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 1st edition, 2000. 8

[Pfe18] Franz Pfeiffer. X-ray ptychography. *Nature Photonics*, 12(1):9–17, 2018. 278

[PGW+12] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012. 286, 486, 532

[PHD20] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *arXiv preprint arXiv:2008.08186*, 2020. 540, 547

[Pho75] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of ACM*, 18(6):311–317, 1975. 493

[PKCS16] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016. 285

[Pla72] R. L. Plackett. Studies in the history of probability and statistics. XXIX the discovery of the method of least squares. *Biometrika*, 59(2):239–251, 1972. 13

[PLVH20] Ambar Pal, Connor Lane, René Vidal, and Benjamin D Haeffele. On the regularization properties of structured dropout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7671–7679, 2020. 537

[PMS94] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 84–91, 1994. 474, 480

[Pol64] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comput. Math. & Math. Phys.*, 4(5):1 – 17, 1964. 325, 331, 470, 626

[Pow69] M. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969. 333, 363, 631

[Pre98] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998. 537

[PRE16] Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *Journal of Machine Learning Research*, 18, 07 2016. 561

[PRK93] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with application to wavelet decomposition. In *Asilomar Conference on Signals, Systems and Computer*, 1993. 358, 362

[PRSE18] Vardan Papyan, Yaniv Romano, Jeremias Sulam, and Michael Elad. Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. *IEEE Signal Processing Magazine*, 35(4):72–89, 2018. 537

[PS16] Thomas Pock and Shoham Sabach. Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016. 470

[PSG+16] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016. 294

[PSM82] R. Pickholtz, D. Schilling, and L. Milstein. Theory of spread-spectrum communications–A tutorial. *IEEE Trans. Commun.*, 30(5):855–884, 1982. 457

[PSV77] G. C. Papanicolaou, D. Stroock, and S. R. S. Varadhan. Martingale approach to some limit theorems. In *Proceedings of Duke Turbulence Conference in Statistical Mechanics, Dynamical Systems, (ed. D. Ruelle), Duke Univ. Math. Series*, volume 3, 1977. 401

[PTRV98] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: a probabilistic analysis. In *Proceedings of the seventeenth ACM symposium on Principles of database systems*, pages 159–168, 1998. 199

[PV08] Paolo Prandoni and Martin Vetterli. *Signal Processing for Communications*. EPFL Press, 2008. 5

[QLZ19] Qing Qu, Xiao Li, and Zhihui Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. In *Advances in Neural Information Processing Systems*, pages 4017–4028, 2019. 275, 287, 296, 301, 472, 473, 557, 558

[QSW14] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, pages 3401–3409, 2014. 288

[Qui86] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986. 544

[QYW+20] Haozhi Qi, Chong You, Xiaolong Wang, Yi Ma, and Jitendra Malik. Deep isometric learning for visual recognition. In *Proceedings of the International Conference on International Conference on Machine Learning*, 2020. 537, 543

[QZC19] Wei Qian, Yuqian Zhang, and Yudong Chen. Global convergence of least squares EM for demixing two log-concave densities. In *Advances in Neural Information Processing Systems*, pages 4795–4803, 2019. 271, 298

[QZC20] Wei Qian, Yuqian Zhang, and Yudong Chen. Structures of spurious local minima in $k$-means. *arXiv prepring arXiv:2002.06694*, 2020. 271, 298

[QZEW17] Qing Qu, Yuqian Zhang, Yonina Eldar, and John Wright. Convolutional phase retrieval. In *Advances in Neural Information Processing Systems*, pages 6086–6096, 2017. 280

[QZL+19] Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, and Zhihui Zhu. Analysis of the optimization landscapes for overcomplete representation learning. *arXiv preprint arXiv:1912.02427*, 2019. 275, 287, 294, 296, 297, 298, 301

[QZL+20a] Qing Qu, Yuexiang Zhai, Xiao Li, Yuqian Zhang, and Zhihui Zhu. Analysis of the optimization landscapes for overcomplete representation learning. In *International Conference on Learning Representations*, 2020. 472, 473

[QZL+20b] Qing Qu, Zhihui Zhu, Xiao Li, Manolis C Tsakiris, John Wright, and René Vidal. Finding the sparsest vectors in a subspace: Theory, algo-

rithms, and applications. *arXiv preprint arXiv:2001.06970*, 2020. 275, 288, 299

[Rau09] Holger Rauhut. Circulant and Toeplitz matrices in compressed sensing. *arXiv preprint arXiv:0902.4394*, 2009. 105

[Raz98] B. Razavi. *RF Microelectronics*. Prentice Hall, 1998. 457

[Raz01] B. Razavi. *Design of Analog CMOS Integrated Circuits*. Mc-Graw Hill, 2001. 457

[RB16] Ernest K. Ryu and Stephen Boyd. A primer on monotone operator methods: Survey. *Appl. Comput. Math.*, 15(1):3–43, 2016. 363

[RBZ06] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature methods*, 3(10):793, 2006. 462

[RC93] Tamás Rapcsák and Tibor Csendes. Nonlinear coordinate transformations for unconstrained optimization II. theoretical background. *Journal of Global Optimization*, 3(3):359–375, 1993. 300

[RE14] R. Rubinstein and M. Elad. Dictionary learning for analysis-synthesis thresholding. *IEEE Transactions on Signal Processing*, 62(22):5962–5972, 2014. 559

[Rec10] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 2010. 188

[Rei65] H. Reichenbach. *Philosophic foundations of quantum mechanics*. University of California Press, 1965. 277

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 536

[RFP10] B. Recht, M. Fazel, and P. Parillo. Guaranteed minimum rank solution of matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. 24, 152, 188, 232, 284, 285

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. 567, 576

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978. 17

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. 361

[Rob93] W. Harrison Robert. Phase problem in crystallography. *Journal of the Optical Society of America A*, 10(5):1046–1055, 1993. 277

[Roc73] R Tyrell Rockafellar. The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications*, 12(6):555–562, 1973. 333, 363, 631

[Ros58] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958. 24

[RS05] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719, 2005. 285

[RT96] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996. 403

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008. 102, 103, 104

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. 302

[RW18] Clément W. Royer and Stephen J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018. 388, 391, 399, 418

[Sas83] S. Sastry. The effects of small noise on implicitly defined nonlinear dynamical systems. *IEEE Transactions on Circuits and Systems*, 30(9):651–663, 1983. 400, 403

[Sas99] Shankar Sastry. *Nonlinear Systems: Analysis, Stability, and Control*. Springer, 1999. 1, 2

[SBC14] Weijie Su, Stephen Boyd, and Emmanuel Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014. 331, 627

[SBOR06] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. xi, 475

[SBRL19] Maziar Sanjabi, Sina Baharlouei, Meisam Razaviyayn, and Jason D Lee. When does non-orthogonal tensor decomposition have no spurious local minima? *arXiv preprint arXiv:1911.09815*, 2019. 298

[SC19] Laixi Shi and Yuejie Chi. Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *arXiv preprint arXiv:1911.11167*, 2019. 296, 301

[SDC03] J.-L. Starck, D. L. Donoho, and E. J. Candès. Astronomical image representation by the curvelet transform. *Astronom. Astrophys.*, 398(2):785–800, 2003. 257

[SDLF+17] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017. 297

[SEC+15] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. 267, 275, 277, 280

[SED05] J.-L. Starck, M. Elad, and D. L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Processing*, 14(10):1570–1582, 2005. 257

[Ser06] T. Serre. *Learning a dictionary of shape-components in visual cortex: Comparison with neurons, humans and machines.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2006. 474

[SFF19] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 7276–7286, 2019. 301

[SFH17] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *CoRR*, abs/1710.09829, 2017. 553

[SGHK03] Christoph Stosiek, Olga Garaschuk, Knut Holthoff, and Arthur Konnerth. In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences*, 100(12):7319–7324, 2003. 462

[SGS15] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015. 536, 537, 578

[Sha92] A. Shashua. Geometry and photometry in 3D visual recognition. *Ph.D dissertation, Department of Brain and Cognitive Science, MIT*, 1992. 504

[She94] Jonathan R Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, Carnegie Mellon University, USA, 1994. 398, 399, 593

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 301, 305, 537

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018. 538, 574

[Sho85] N. Shor. *Minimization Methods for Non-Differentiable Functions.* Springer-Verlag, 1985. 61

[Sil80] W. M. Silver. Determining shape and reflectance using multiple images. *Master's thesis, MIT*, 1980. 489

[Sim50] Thomas Simpson. *Doctrine and Application of Fluxions.* J. Nourse, London, 1750. 375

[SJL18] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018. 271, 299

[SJM19] Chaobing Song, Yong Jiang, and Yi Ma. Towards unified acceleration of high-order algorithms under Hölder continuity and uniform convexity. Technical report, (preprint) arXiv:1906.00582, 2019. 331

[SJM20a] Chaobing Song, Yong Jiang, and Yi Ma. Breaking the $O(1/\varepsilon)$ optimal rate for a class of minimax problems. *arXiv preprint arXiv:2003.11758*, 2020. 364

[SJM20b] Chaobing Song, Yong Jiang, and Yi Ma. Stochastic variance reduction via accelerated dual averaging for finite-sum optimization. *arXiv preprint arXiv:2006.10281*, 2020. 362

[SK93] William R Softky and Christof Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience*, 13(1):334–350, 1993. 577

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 536

[SLLB17] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, 2017. 537

[SLX+16] Jian Sun, Huibin Li, Zongben Xu, et al. Deep ADMM-Net for compressive sensing MRI. In *Advances in neural information processing systems*, pages 10–18, 2016. 537

[SMB10] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010. 537

[SMM+17] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 553, 554

[SNT20] Xiaoxia Sun, Nasser M Nasrabadi, and Trac D Tran. Supervised deep sparse coding networks for image classification. *IEEE Transactions on Image Processing*, 29:405–418, 2020. 537, 554

[SPM02] Jean-Luc Starck, E Pantin, and F Murtagh. Deconvolution in astronomy: A review. *Publications of the Astronomical Society of the Pacific*, 114(800):1051, 2002. 294

[SPRE18] Jeremias Sulam, Vardan Papyan, Yaniv Romano, and Michael Elad. Multilayer convolutional sparse modeling: Pursuit and dictionary learning. *IEEE Transactions on Signal Processing*, 66(15):4090–4104, 2018. 537, 561

[SQW15] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015. 28, 29, 265, 269, 271, 272, 301

[SQW17a] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and geometric picture. *IEEE Transactions on Information Theory*, 63(2), 2017. 275, 287, 288, 292

[SQW17b] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017. 275, 287, 288, 292

[SQW18] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. 275, 279

[SS86] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7(4):1307–1330, 1986. 15

[SS17] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017. 299

[SSL06] F. Sanja, D. Skocaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):337–350, 2006. 475, 482

[STDV18] Forrest Sheldon, Fabio L Traversa, and Massimiliano Di Ventra. Taming a non-convex landscape with dynamical long-range order: memcomputing the Ising spin-glass. *arXiv preprint arXiv:1810.03712*, 2018. 271

[Sto09] M. Stojnic. Various thresholds for $\ell_1$-optimization in compressed sensing. *arxiv.org/abs/0907.3666*, 2009. 264

[Sun19a] Ju Sun. Provable nonconvex methods/algorithms. https://sunju.org/research/nonconvex/, 2019. 265, 269, 271, 272, 299

[Sun19b] Ruoyu Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019. 271, 299

[SW08] R. Schneider and W. Weil. *Stochastic and Integral Geometry*. Springer, 2008. 248, 251, 252

[SWW12] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, 2012. 288, 300

[SXZ+20] Yifei Shen, Ye Xue, Jun Zhang, Khaled B. Letaief, and Vincent Lau. Complete dictionary learning via $\ell^p$-norm maximization. *arXiv preprint arXiv:2002.10043*, 2020. 288

[SY07] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007. 285

[SYJS05] Jian Sun, Lu Yuan, Jiaya Jia, and Heung-Yeung Shum. Image completion with structure propagation. *ACM Trans. Graph.*, 24(3):861–868, 2005. 513, 514

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 536

[Tal95] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathematiques de l'I.H.E.S.*, 81:73–205, 1995. 637

[TBF+12] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74(2):245–266, 2012. 363

[TG07] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. 358, 362, 452

[Tho04] Colin J Thompson. Inequality with applications in statistical mechanics. *Journal of Mathematical Physics*, 6(11):1812–1813, 2004. 638

[Tib96] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996. 18, 107, 363

[TM01] D. Taubman and M. Marcellin. *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, Norwell, MA, 2001. 40

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1991. 482

[Tro10] J. Tropp. Beyond Nyquist: efficient sampling of sparse bandlimited signals. *IEEE Transactions on Information Theory*, 56(1):520–544, 2010. 7, 23, 447

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. 640

[Tuc66] L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966. 261

[TV18] Manolis C Tsakiris and René Vidal. Dual principal component pursuit. *The Journal of Machine Learning Research*, 19(1):684–732, 2018. 287

[TW15] Lei Tian and Laura Waller. 3D intensity and phase imaging from light field measurements in an LED array microscope. *optica*, 2(2):104–111, 2015. 278

[TY09] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. preprint, 2009. [Online]. Available: http://math.nus.edu.sg/~matys/apg.pdf, 2009. 233

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 537

[Val77] Leslie G. Valiant. Graph-theoretic arguments in low-level complexity. *Lecture Notes in Computer Science*, 53:162–176, 1977. 197

[Van16] S. Van De Geer. *Estimation and Testing Under Sparsity*. Springer, 2016. xiii

[VBGS17] Rene Vidal, Joan Bruna, Raja Giryes, and Stefano Soatto. Mathematics of deep learning. *arXiv preprint arXiv:1712.04741*, 2017. 299

[VdM96] P. Van Overschee and B. de Moor. *Subspace Identification for Linear Systems*. Kluwer Academic, 1996. 2

[Ver08] R. Vershynin. Spectral norms of products of random and deterministic matrices. Online preprint, Available: http://arxiv.org/abs/0812.2432, 2008. 634

[Ver18] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. 25, 634

[VK95] M. Vetterli and J. Kovačević. *Wavelets and subband coding*. Prentice Hall, 1995. 40

[VMS16] Rene Vidal, Yi Ma, and S. S. Sastry. *Generalized Principal Component Analysis*. Springer Publishing Company, Incorporated, 1st edition, 2016. xix, 20, 200, 366, 540, 541, 574

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 60006010, Red Hook, NY, USA, 2017. Curran Associates Inc. 575

[Wai09a] M. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009. 264

[Wai09b] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming. *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009. 132

[Wai19] M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. xiii, 25, 634

[Wal63] Adriaan Walther. The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49, 1963. 277

[Wal91] G. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991. 40

[WB18] T. Wiatowski and H. Blcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, 64(3):1845–1866, 2018. 561

[WBYM21] Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021. 576

[WDCB05] Michael B Wakin, David L Donoho, Hyeokho Choi, and Richard G Baraniuk. The multiscale structure of non-differentiable image manifolds. In *Proceedings of SPIE, the International Society for Optical Engineering*, pages 59141B–1, 2005. 263, 555

[WdM15] Irène Waldspurger, Alexandre d`Aspremont, and Stéphane Mallat. Phase recovery, MaxCut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015. 280

[WGE17] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017. 280

[WGMM13] J. Wright, A. Garnesh, K. Min, and Y. Ma. Compressive principal component pursuit. *IMA Journal on Information and Inference*, 2(1):32–68, 2013. 229

[WGS+10] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision*, 2010. 138, 285, 504

[WH18] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 537

[WJ08] Martin Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 01 2008. 8

[WLY+15] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 537

[WM10] J. Wright and Y. Ma. Dense error correction via $\ell^1$-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010. 15, 485, 486, 574

[WMM+10] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010. 41, 291

[WNF08] S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008. 363

[WNF09] Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009. 471

[Woo80] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980. 137, 489

[WPPA16] Scott Wisdom, Thomas Powers, James Pitton, and Les Atlas. Interpretable recurrent neural networks using sequential sparse recovery. *ArXiv*, abs/1611.07252, 2016. 561

[Wri87] S. Wright. *Primal-Dual Interior Point Methods*. SIAM, 1987. 27

[Wri97] G. Wright. Magnetic resonance imaging. *IEEE Signal Processing Magazine*, 14(1):56–66, 1997. 426

[WTL+08] John Wright, Yangyu Tao, Zhouchen Lin, Yi Ma, and Heung-Yeung Shum. Classification via minimum incremental coding length (MICL). In *Advances in Neural Information Processing Systems*, pages 1633–1640, 2008. 542, 574

[Wun12] Henning Wunderlich. On a theorem of Razborov. *Computational Complexity*, 21(3):431–477, 2012. 197

[WWG+09] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Yi Ma. Toward a practical face recognition: Robust pose and illumination via sparse representation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2009. 486, 532

[WWG+12] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Towards a practical face recognition system: Robust alignment and illumination via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(2):372–386, 2012. 486, 532

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. 331, 627

[WX20] Denny Wu and J. Xu. On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. *ArXiv*, abs/2006.05800, 2020. 551

[WYD20] Kaizheng Wang, Yuling Yan, and Mateo Diaz. Efficient clustering for stretched mixtures: Landscape and optimality. *arXiv preprint arXiv:2003.09960*, 2020. 271

[WYG+09] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210 – 227, 2009. 23, 42, 66, 474, 485

[WYYZ08] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008. 443

[WZL+14] Jie Wang, Jiayu Zhou, Jun Liu, Peter Wonka, and Jieping Ye. A safe screening rule for sparse logistic regression. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, pages 1053–1061, Cambridge, MA, USA, 2014. MIT Press. 363

[XBSD+18] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pages 5393–5402, 2018. 537

[XCS10] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010. 286

[XCS12] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, May 2012. 223, 238

[XGD+17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. 552, 554

[XHM16] Ji Xu, Daniel Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*, 2016. 298

[Xu17] Yangyang Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM Journal on Optimization*, 27(3):1459–1484, 2017. 341, 346, 363

[XWCL15] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 536

[XZ13] Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013. 471

[XZ14] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014. 362

[Yau74] Shing-Tung Yau. Non-existence of continuous convex functions on certain Riemannian manifolds. *Mathematische Annalen*, 207(4):269–270, 1974. 300

[YCY+20] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chao-Bing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *NeurIPS*, 2020. 545

[YDZ+15] Li-Hao Yeh, Jonathan Dong, Jingshan Zhong, Lei Tian, Michael Chen, Gongguo Tang, Mahdi Soltanolkotabi, and Laura Waller. Experimental robustness of Fourier ptychography phase retrieval algorithms. *Optics express*, 23(26):33214–33240, 2015. 278

[YHK+16] Rabia Tugce Yazicigil, Tanbir Haque, Manoj Kumar, Jeffrey Yuan, John Wright, and Peter R. Kinget. A compressed sampling time-segmented quadrature analog-to-information converter that exploits adaptive thresholding and virtual extension of physical hardware for rapid interferer detection. In *IEEE Intern. Solid-State Circuits Conference*, 2016. 459, 460

[YHW+15] Rabia Tugce Yazicigil, Tanbir Haque, Michael R. Whalen, Jeffrey Yuan, John Wright, and Peter R. Kinget. Wideband rapid interferer detector exploiting compressed sampling with a quadrature analog-to-information converter. In *IEEE Intern. Solid-State Circuits Conference*, pages 3047–3064, December 2015. 453, 455, 456, 459, 460

[YMO13] Yi Yang, Jianwei Ma, and Stanley Osher. Seismic data reconstruction via matrix completion. *Inverse Problems & Imaging*, 7(4):1379, 2013. 285

[YOGD08] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for $\ell_1$-minimization with applications to compressed sensing. *SIAM Journal on Imaging Science*, 1(1):143–168, 2008. 232, 336

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. 537

[YWHM08] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2008. 66

[YWHM10] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010. 41, 292

[YY09] X. Yuan and J. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *Pacific Journal of Optimization*, 1 2009. 233

[YYY+20] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning (ICML)*, 2020. 551

[YZQM20] Chong You, Zhihui Zhu, Qing Qu, and Yi Ma. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. *arXiv preprint arXiv:2006.08857*, 2020. 538, 574

[YZY10] J. Yang, Y. Zhang, and W. Yin. A fast alternating direction method for TVL1-L2 signal reconstruction from partial Fourier data. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):288, 2010. 440, 442

[ZB18] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018. 287

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 533, 574

[ZCPR03] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfield. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003. 475

[ZGLM12] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma. TILT: Transform-invariant low-rank textures. *International Journal of Computer Vision (IJCV)*, 99(1):1–24, 2012. 516, 520, 521, 531, 532

[Zha00] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 525

[ZKW18] Yuqian Zhang, Han-Wen Kuo, and John Wright. Structured local minima in sparse blind deconvolution. In *Advances in Neural Information Processing Systems 31*, pages 2328–2337, 2018. 275, 287, 295, 466

[ZL17] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *online* https://arxiv.org/abs/1611.01578, 2017. 537

[ZLC17] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1980–2022, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR. 405

[ZLGM10] Z. Zhang, X. Liang, A. Ganesh, and Y. Ma. TILT: Transform invariant low-rank textures. In *Proceedings of Asian Conference on Computer Vision*, 2010. 516, 520, 531, 532

[ZLK+17] Yuqian Zhang, Yenson Lau, Han-Wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4381–4389. IEEE, 2017. 287, 295, 297, 465, 466, 471

[ZLM11] Z. Zhang, X. Liang, and Y. Ma. Unwrapping low-rank textures on generalized cylindrical surfaces. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011. 515, 522, 524, 525, 531

[ZLM20] Yichao Zhou, Shichen Liu, and Yi Ma. Learning to detect 3D reflection symmetry for single-view reconstruction. *arXiv preprint arXiv:2006.10042*, 2020. 531

[ZLTW18] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018. 285

[ZM20] Zhengyuan Zhou and Yi Ma. Comments on efficient singular value thresholding computation. *arXiv preprint arXiv:2011.06710*, 2020. 317

[ZMKW13] Y. Zhang, C. Mu, H. Kuo, and J. Wright. Toward guaranteed illumination models for non-convex objects. In *2013 IEEE International Conference on Computer Vision*, pages 937–944, 2013. 199, 493, 504

[ZMM11] Z. Zhang, Y. Matsushita, and Y. Ma. Camera calibration with lens distortion from low-rank textures. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011. 515, 530, 531

[ZMZM20] Yuexiang Zhai, Hermish Mehta, Zhengyuan Zhou, and Yuliang Ma. Understanding $\ell^4$-based dictionary learning: Interpretation, stability, and robustness. In *ICLR*, 2020. 293

[ZQHM19] Yichao Zhou, Haozhi Qi, Jingwei Huang, and Yi Ma. NeurVPS: Neural vanishing point scanning via conic convolution. In *NeurIPS*, 2019. 531

[ZQM19] Yichao Zhou, Haozhi Qi, and Yi Ma. End-to-end wireframe parsing. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 962–971, 10 2019. 531

[ZQZ+19] Yichao Zhou, Haozhi Qi, Yuexiang Zhai, Qi Qiang Sun, Zhili Chen, Li-Yi Wei, and Yi Ma. Learning to reconstruct 3D Manhattan wireframes from a single image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7697–7706, 2019. 531

[ZWJ14] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014. 67

[ZWMM09] Z. Zhou, A. Wagner, H. Mobahi, and Y. Ma. Face recognition with contiguous occlusion using Markov random fields. In *Proceedings of International Conference on Computer Vision*, 2009. 205, 486, 513

[ZWR+18] Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, Rene Vidal, and Manolis Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. In *Advances in Neural Information Processing Systems*, pages 2171–2181, 2018. 287, 302

[ZYL+20] Yuexiang Zhai, Zitong Yang, Zhenyu Liao, John Wright, and Yi Ma. Complete dictionary learning via $\ell^4$-norm maximization over the orthogonal group. *Journal of Machine Learning Research*, 2020. 287, 288, 293, 302, 414, 415

[ZYZY14] Xiaowei Zhou, Can Yang, Hongyu Zhao, and Weichuan Yu. Low-rank modeling and its applications in image analysis. *ACM Computing Surveys (CSUR)*, 47(2):1–33, 2014. 285

[ZZWM14] Xiaoqin Zhang, Zhengyuan Zhou, Di Wang, and Yi Ma. Hybrid singular value thresholding for tensor completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-14)*, 2014. 317

# List of Symbols

| | |
|---|---|
| $\mathbb{R}$ | The real numbers. |
| $\mathbb{C}$ | The complex numbers. |
| $\mathsf{i} = \sqrt{-1}$ | The unit imaginary number as a solution to $x^2 + 1 = 0$. |
| $\mathbb{R}^n, \mathbb{C}^n$ | The $n$-dimensional real or complex space. |
| $\mathbb{R}^{m \times n}, \mathbb{C}^{m \times n}$ | The space of $m \times n$ real or complex matrices. |
| $\mathbb{S}^{n-1}$ | A unit sphere in $\mathbb{R}^n$. |
| $\mathbb{G}$ | A general (matrix) group. |
| $[k]$ | The set $\{1, \ldots, k\}$. |
| $\mathsf{I}$ | A subset of indices usually indicating the support of a sparse vector. |
| $\Omega$ | A subset of indices for entries of a matrix. |
| $\mathsf{S}$ | A subspace. |
| $\mathsf{O}(n)$ | The orthogonal group. |
| $\mathsf{GL}(n)$ | The general linear group. |
| $\mathsf{SL}(n)$ | The special linear group. |
| $\mathsf{SP}(n)$ | The sign permutation group. |
| $a, b, c, x, y, A, B, C$ | Scalars. |
| $C_1, C_2, \ldots$ | Large constants. |
| $c_1, c_2, \ldots$ | Small constants. |
| $\boldsymbol{x}, \boldsymbol{y}$ | Vectors, always represented as columns. |
| $\operatorname{supp}(\boldsymbol{x})$ | For $\boldsymbol{x} \in \mathbb{R}^n$, the indices of the nonzero entries, $\subseteq [n]$. |
| $\operatorname{sign}(\boldsymbol{x})$ | The signs of a vector $\boldsymbol{x} \in \mathbb{R}^n$, in $\{-1, 0, 1\}^n$. |
| $\boldsymbol{X}, \boldsymbol{Y}$ | Matrices. |
| $\boldsymbol{L}, \boldsymbol{S}$ | $\boldsymbol{L}$ indicates a low-rank matrix, and $\boldsymbol{S}$ a sparse matrix. |
| $\boldsymbol{\mathcal{X}}$ | Tensors (of order $> 2$). |
| $\boldsymbol{A} \succeq \boldsymbol{B}$ | The semidefinite order, i.e., $\boldsymbol{A} - \boldsymbol{B}$ is semidefinite. |
| $\boldsymbol{A} \succ \boldsymbol{B}$ | Strict semidefinite order, i.e., $\boldsymbol{A} - \boldsymbol{B}$ is positive definite. |
| $\mathcal{S}_+^n$ | The cone of symmetric positive semidefinite matrices of size $n \times n$. |
| $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ | The standard basis vectors for $\mathbb{R}^m$. |
| $\boldsymbol{E}_{i,j}$ | The standard basis vectors for the space of matrices $\mathbb{R}^{m \times n}$. |
| $\mathbf{0}$ | The zero vector or matrix, depending on context. |
| $\mathbf{1}$ | The all ones vector or matrix, depending on context. |
| $\boldsymbol{I}$ | The identity matrix. |

| | |
|---|---|
| $\boldsymbol{a}^*$, $\boldsymbol{A}^*$ | The (conjugate) transpose of a vector $\boldsymbol{a}$ or a matrix $\boldsymbol{A}$. |
| $\boldsymbol{A}^{-1}$ | The inverse of a nonsingular matrix $\boldsymbol{A}$. |
| $\boldsymbol{A}^\dagger$ | The pseudoinverse of an arbitrary matrix $\boldsymbol{A}$. |
| $\operatorname{null}(\boldsymbol{A})$ | The null space of $\boldsymbol{A}$. |
| $\operatorname{range}(\boldsymbol{A})$ | The range (column space) of $\boldsymbol{A}$. |
| $\operatorname{range}(\boldsymbol{A}^*)$ | The row space of $\boldsymbol{A}$. |
| $\boldsymbol{X}_{i,j}$ | The $(i,j)$ element of matrix $\boldsymbol{X}$. Where possible, use $i$ for the first index, $j$ for the second index. |
| $\boldsymbol{X}_{\mathsf{I},\mathsf{J}}$ | For $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, the square submatrix index by $\mathsf{I} \subseteq [m]$, $\mathsf{J} \subseteq [n]$. |
| $\boldsymbol{X}_{*,\mathsf{J}}$ | Shorthand for the column submatrix indexed by $\mathsf{J}$. |
| $\boldsymbol{X}_{\mathsf{I},*}$ | Shorthand for the row submatrix indexed by $\mathsf{I}$. |
| $\boldsymbol{P}_{\mathsf{I}}$ | Abuse of notation for the projection (matrix) of a vector onto the coordinate subspace indexed by $\mathsf{I}$. |
| $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ | The singular value decomposition of $\boldsymbol{A}$. Prefer the "compact" form. If $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\operatorname{rank}(\boldsymbol{A}) = r$, $\boldsymbol{U} \in \mathbb{R}^{m\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$. |
| $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ | The eigenvector decomposition of a symmetric matrix $\boldsymbol{P} \in \mathbb{R}^{m\times m}$. Here, $\boldsymbol{\Lambda}$ is diagonal, and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ with $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$. |
| $[\boldsymbol{x}]_k$ | A best $k$-term approximation to $\boldsymbol{x}$. |
| $\operatorname{soft}(\cdot,\tau)$ | Entry-wise soft thresholding operator on a scalar, vector or a matrix, with a threshold $\tau \geq 0$. |
| $\mathcal{S}_\tau(\cdot)$ | A shorthand for the entry-wise soft thresholding operator, with the threshold $\tau$. |
| $\mathcal{D}_\tau(\boldsymbol{A})$ | The soft thresholding operator on the singular values of the matrix $\boldsymbol{A}$, with the threshold $\tau$. |
| $\boldsymbol{a} \circledast \boldsymbol{x}$ | The convolution of two signals $\boldsymbol{a}$ and $\boldsymbol{x}$. When both are of finite length, it can represent either circulant convolution or truncated one, depending on the context. |
| $\|\boldsymbol{x}\|_p$ | The vector $\ell^p$ norm |
| $\|\boldsymbol{X}\|$ | The $\ell^2$ operator norm, $\sigma_1(\boldsymbol{X})$. |
| $\|\boldsymbol{X}\|_F$ | The Frobenius norm. |
| $\|\boldsymbol{X}\|_*$ | The nuclear norm. |
| $\|\mathcal{A}\|_{V\to W}$ | The operator norm of $\mathcal{A}$, as an operator from normed space $V$ to normed space $W$. |
| $\|\boldsymbol{X}\|_{\ell^1\to\ell^p}$ | The $\ell^1 \to \ell^p$ operator norm, $\max_j \|\boldsymbol{X}\boldsymbol{e}_j\|_p$. |
| $\|\boldsymbol{X}\|_{\ell^2\to\ell^\infty}$ | The $\ell^2 \to \ell^\infty$ operator norm, $\max_i \|\boldsymbol{e}_i^*\boldsymbol{X}\|_2$. |
| $\|\cdot\|_\diamond^*$ | The dual norm of $\|\cdot\|_\diamond$. |
| $\|\boldsymbol{X}\|_{\ell^1\to\ell^2}^*$ | The dual norm of the $\ell^1 \to \ell^2$ operator norm, $\sum_j \|\boldsymbol{X}\boldsymbol{e}_j\|_2$. |
| $O(n)$ | "Big-O" means upper bounded by $C \cdot n$ for some constant $C$. |
| $\Omega(n)$ | "Big-Omega" means lower bounded by $C \cdot n$ for some constant $C$. |

| | |
|---|---|
| $\Theta(n)$ | "Big-Theta" means lower bounded by $c \cdot n$ for some constant $c$ and upper bounded by $C \cdot n$ for some constant $C > c$. |
| $o(n)$ | "little-o" means ultimately smaller than $n$. |
| $\partial f(\boldsymbol{x})$ | Subdifferential of a function $f(\cdot)$ at $\boldsymbol{x}$. |
| $\nabla f(\boldsymbol{x})$ | The gradient of a differentiable function $f$ at $\boldsymbol{x}$. |
| $\nabla^2 f(\boldsymbol{x})$ | The Hessian of a twice-differentiable function $f$ at $\boldsymbol{x}$. |
| $\mathcal{A}, \mathcal{B}, \mathcal{P}$ | General linear maps. These act on elements of their domain via square brackets, e.g., $\mathcal{A}[\boldsymbol{X}]$. |
| $\mathcal{P}_{\mathsf{S}}$ | Orthonormal projector onto a subspace of a vector space. |
| $\mathcal{P}_{\Omega}$ | The projection operator of a matrix onto the coordinate subspace indexed by $\Omega$. |
| $\min \ (x+1)^2$ | Unconstrained minimization. |
| $\max \ -(x+1)^2$ | Unconstrained maximization. |
| $\min \ f(\boldsymbol{x})$ subject to $h(\boldsymbol{x}) \leq 0$. | Constrained minimization. |
| $\boldsymbol{x}_{\text{true}}$, $\boldsymbol{X}_{\text{true}}$ | Ground truth solutions. |
| $\boldsymbol{x}_o$, $\boldsymbol{X}_o$ | Shorthand for ground truth solutions and objective for any algorithm. |
| $\boldsymbol{x}_0, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}$ | Initial point, and estimates at the $k$-th and the $(k+1)$-th iteration of an algorithm. |
| $\{\boldsymbol{x}_i\}$ | A sequence of (vector) iterates in optimization or a set of samples in statistics. |
| $\boldsymbol{X}_0, \boldsymbol{X}_k, \boldsymbol{X}_{k+1}$ | Initial point, and estimates at the $k$-th and the $(k+1)$-th iteration of an algorithm. |
| $\{\boldsymbol{X}_i\}$ | A sequence of (matrix) iterates in optimization or a set of samples in statistics. |
| $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{X}}$ | Estimated approximate solutions (to an estimation or optimization problem). |
| $\boldsymbol{x}_\star$, $\boldsymbol{X}_\star$ | Converged solutions of an iterative algorithm. |
| $\hat{\boldsymbol{x}} \in \arg\min \ f(\boldsymbol{x})$. | Set of minimizers of a function $f(\cdot)$. |
| $\boldsymbol{x}_\star = \arg\min \ f(\boldsymbol{x})$. | Shorthand when the minimizer of $f(\cdot)$ is unique. |
| $\mathbb{P}\left[ X > t \right] < \exp(-t^2/2)$ | Probability. |
| $\mathbb{P}\left[ X > 1 \mid X < 2 \right] = 0$ | Conditional probability. |
| $\mathbb{E}\left[\cdot\right]$ | Expectation. |
| $\mathbb{E}\left[\cdot \mid \cdot\right]$ | Conditional expectation. |
| $\mathbb{1}_{x \leq 3}$ | Indicator for an event. |
| $\boldsymbol{e}$, $\boldsymbol{E}$ | A gross error vector or matrix. |
| $\boldsymbol{z}$, $\boldsymbol{Z}$ | A vector or matrix of noise. |
| $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | The Gaussian or normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. |
| $\text{Ber}(\rho)$ | The Bernoulli distribution with the probability $\rho \in [0, 1]$. |

# Index