Non-convex Optimization for Machine Learning

Prateek Jain
Microsoft Research India
prajain@microsoft.com

Purushottam Kar
IIT Kanpur
purushot@cse.iitk.ac.in

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

P. Jain and P. Kar. *Non-convex Optimization for Machine Learning*. Foundations and Trends® in Machine Learning, vol. 10, no. 3-4, pp. 142–336, 2017.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-368-3

Foundations and Trends® in Machine Learning
Volume 10, Issue 3-4, 2017

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2017, Volume 10, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in Machine Learning
Vol. 10, No. 3-4 (2017) 142–336

DOI: 10.1561/2200000058

now
the essence of knowledge

Non-convex Optimization for Machine Learning

Prateek Jain
Microsoft Research India
prajain@microsoft.com

Purushottam Kar
IIT Kanpur
purushot@cse.iitk.ac.in

Contents

Abstract

A vast majority of machine learning algorithms train their models and perform inference by solving optimization problems. In order to capture the learning and prediction problems accurately, structural constraints such as sparsity or low rank are frequently imposed or else the objective itself is designed to be a non-convex function. This is especially true of algorithms that operate in high-dimensional spaces or that train non-linear models such as tensor models and deep networks.

The freedom to express the learning problem as a non-convex optimization problem gives immense modeling power to the algorithm designer, but often such problems are NP-hard to solve. A popular workaround to this has been to relax non-convex problems to convex ones and use traditional methods to solve the (convex) *relaxed* optimization problems. However this approach may be lossy and nevertheless presents significant challenges for large scale optimization.

On the other hand, direct approaches to non-convex optimization have met with resounding success in several domains and remain the methods of choice for the practitioner, as they frequently outperform relaxation-based techniques – popular heuristics include projected gradient descent and alternating minimization. However, these are often poorly understood in terms of their convergence and other properties.

This monograph presents a selection of recent advances that bridge a long-standing gap in our understanding of these heuristics. We hope that an insight into the inner workings of these methods will allow the reader to appreciate the unique marriage of task structure and generative models that allow these heuristic techniques to (provably) succeed. The monograph will lead the reader through several widely used non-convex optimization techniques, as well as applications thereof. The goal of this monograph is to both, introduce the rich literature in this area, as well as equip the reader with the tools and techniques needed to analyze these simple procedures for non-convex problems.

P. Jain and P. Kar. *Non-convex Optimization for Machine Learning*. Foundations and Trends® in Machine Learning, vol. 10, no. 3-4, pp. 142–336, 2017.
DOI: 10.1561/2200000058.

Preface

Optimization as a field of study has permeated much of science and technology. The advent of the digital computer and a tremendous subsequent increase in our computational prowess has increased the impact of optimization in our lives. Today, tiny details such as airline schedules all the way to leaps and strides in medicine, physics and artificial intelligence, all rely on modern advances in optimization techniques.

For a large portion of this period of excitement, our energies were focused largely on convex optimization problems, given our deep understanding of the structural properties of convex sets and convex functions. However, modern applications in domains such as signal processing, bio-informatics and machine learning, are often dissatisfied with convex formulations alone since there exist non-convex formulations that better capture the problem structure. For applications in these domains, models trained using non-convex formulations often offer excellent performance and other desirable properties such as compactness and reduced prediction times.

Examples of applications that benefit from non-convex optimization techniques include gene expression analysis, recommendation systems, clustering, and outlier and anomaly detection. In order to get satisfactory solutions to these problems, that are scalable and accurate, we require a deeper understanding of non-convex optimization problems that naturally arise in these problem settings.

Such an understanding was lacking until very recently and non-convex optimization found little attention as an active area of study, being regarded as intractable. Fortunately, a long line of works have recently led areas such as computer science, signal processing, and statistics to realize that the general abhorrence to non-convex optimization problems hitherto practiced, was misled. These works demonstrated in a beautiful way, that although non-convex optimization problems do suffer from intractability in general, those that arise in *natural settings* such as machine learning and signal processing, possess additional structure that allow the intractability results to be circumvented.

The first of these works still religiously stuck to convex optimization as the method of choice, and instead, sought to show that certain classes of non-convex problems which possess suitable additional structure as offered by natural instances of those problems, could be converted to convex problems without any loss. More precisely, it was shown that the original non-convex problem and the modified convex problem possessed a common optimum and thus, the solution to the convex problem would automatically solve the non-convex problem as well! However, these approaches had a price to pay in terms of the time it took to solve these so-called *relaxed* convex problems. In several instances, these relaxed problems, although not intractable to solve, were nevertheless challenging to solve, at large scales.

It took a second wave of still more recent results to usher in provable non-convex optimization techniques which abstained from relaxations, solved the non-convex problems in their native forms, and yet seemed to offer the same quality of results as relaxation methods did. These newer results were accompanied with a newer realization that, for a range of domains such as sparse recovery, matrix completion and robust learning, these direct techniques are faster, often by an order of magnitude or more, than their relaxation-based cousins.

This monograph wishes to tell the story of this realization and the wisdom we gained from it from the point of view of machine learning and signal processing applications. The monograph will introduce the reader to a lively world of non-convex optimization problems with rich structure that can be exploited to obtain extremely scalable

solutions to these problems. Put a bit more dramatically, it will seek to show how problems that were once avoided, having been shown to be NP-hard to solve, now have solvers that operate in near-linear time, by carefully analyzing and exploiting additional task structure! It will seek to inform the reader on how to look for such structure in diverse application areas, as well as equip the reader with a sound background in fundamental tools and concepts required to analyze such problem areas and come up with newer solutions.

How to use this monograph We have made efforts to make this monograph as self-contained as possible while not losing focus of the main topic of non-convex optimization techniques. Consequently, we have devoted entire sections to present a tutorial-like treatment to basic concepts in convex analysis and optimization, as well as their non-convex counterparts. As such, this monograph can be used for a semester-length course on the basics of non-convex optimization with applications to machine learning.

On the other hand, it is also possible to cherry pick portions of the monograph, such the section on sparse recovery, or the EM algorithm, for inclusion in a broader course. Several courses such as those in machine learning, optimization, and signal processing may benefit from the inclusion of such topics. However, we advise that relevant background sections (see Figure 1) be covered beforehand.

While striving for breadth, the limits of space have constrained us from looking at some topics in much detail. Examples include the construction of design matrices that satisfy the RIP/RSC properties and pursuit style methods, but there are several others. However, for all such omissions, the bibliographic notes at the end of each section can always be consulted for references to details of the omitted topics. We have also been unable to address several application areas such as dictionary learning, advances in low-rank tensor decompositions, topic modeling and community detection in graphs but have provided pointers to prominent works in these application areas too.

The organization of this monograph is outlined below with Figure 1 presenting a suggested order of reading the various sections.

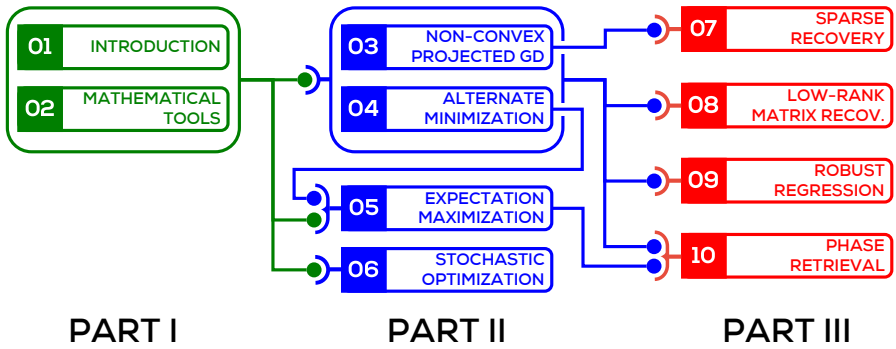


Figure 1: A schematic showing the suggested order of reading the sections. For example, concepts introduced in § 3 and 4 are helpful for § 9 but a thorough reading of § 6 is not required for the same. Similarly, we recommend reading § 5 after going through § 4 but a reader may choose to proceed to § 7 directly after reading § 3.

Part I: Introduction and Basic Tools
This part will offer an introductory note and a section exploring some basic definitions and algorithmic tools in convex optimization. These sections are recommended to readers not intimately familiar with basics of numerical optimization.

Section 1 - Introduction This section will give a more relaxed introduction to the area of non-convex optimization by discussing applications that motivate the use of non-convex formulations. The discussion will also clarify the scope of this monograph.

Section 2 - Mathematical Tools This section will set up notation and introduce some basic mathematical tools in convex optimization. This section is basically a handy repository of useful concepts and results and can be skipped by a reader familiar with them. Parts of the section may instead be referred back to, as and when needed, using the cross-referencing links in the monograph.

Part II: Non-convex Optimization Primitives
This part will equip the reader with a collection of primitives most widely used in non-convex optimization problems.

Section 3 - Non-convex Projected Gradient Descent This section will introduce the simple and intuitive projected gradient descent method in the context of non-convex optimization. Variants of this method will be used in later sections to solve problems such as sparse recovery and robust learning.

Section 4 - Alternating Minimization This section will introduce the principle of alternating minimization which is widely used in optimization problems over two or more (groups of) variables. The methods introduced in this section will be later used in later sections to solve problems such as low-rank matrix recovery, robust regression, and phase retrieval.

Section 5 - The EM Algorithm This section will introduce the EM algorithm which is a widely used optimization primitive for learning problems with latent variables. Although EM is a form of alternating minimization, given its significance, the section gives it special attention. This section will discuss some recent advances in the analysis and applications of this method and look at two case studies in learning Gaussian mixture models and mixed regression to illustrate the algorithm and its analyses.

Section 6 - Stochastic Non-convex Optimization This section will look at some recent advances in using stochastic optimization techniques for solving optimization problems with non-convex objectives. The section will also introduce the problem of tensor factorization as a case study for the algorithms being studied.

Part III - Applications

This part will take a look at four interesting applications in the areas of machine learning and signal processing and explore how the non-convex optimization techniques introduced earlier can be used to solve these problems.

Section 7 - Sparse Recovery This section will look at a very basic non-convex optimization problem, that of performing linear regression to fit a sparse model to the data. The section will discuss conditions under which it is possible to do so in polynomial time and show how the non-convex projected gradient descent method studied earlier can be used to offer provably optimal solutions. The section will also point to other techniques used to solve this problem, as well as refer to extensions and related results.

Section 8 - Low-rank Matrix Recovery This section will address the more general problem of low rank matrix recovery with specific emphasis on low-rank matrix completion. The section will gently introduce low-rank matrix recovery as a generalization of sparse linear regression that was studied in the previous section and then move on to look at matrix completion in more detail. The section will apply both the non-convex projected gradient descent and alternating minimization methods in the context of low-rank matrix recovery, analyzing simple cases and pointing to relevant literature.

Section 9 - Robust Regression This section will look at a widely studied area of machine learning, namely robust learning, from the point of view of regression. Algorithms that are robust to (adversarial) corruption in data are sought after in several areas of signal processing and learning. The section will explore how to use the projected gradient and alternating minimization techniques to solve the robust regression problem and also look at applications of robust regression to robust face recognition and robust time series analysis.

Section 10 - Phase Retrieval This section will look at some recent advances in the application of non-convex optimization to phase retrieval. This problem lies at the heart of several imaging techniques such as X-ray crystallography and electron microscopy. A lot remains to be understood about this problem and existing algorithms often struggle to cope with the retrieval problems presented in practice.

The area of non-convex optimization has considerably widened in both scope and application in recent years and newer methods and analyses are being proposed at a rapid pace. While this makes researchers working in this area extremely happy, it also makes summarizing the vast body of work in a monograph such as this, more challenging. We have striven to strike a balance between presenting results that are the best known, and presenting them in a manner accessible to a newcomer. However, in all cases, the bibliography notes at the end of each section do contain pointers to the state of the art in that area and can be referenced for follow-up readings.

Prateek Jain, Bangalore, India
Purushottam Kar, Kanpur, India
December 2, 2017

Mathematical Notation

- The set of real numbers is denoted by $\mathbb{R}$. The set of natural numbers is denoted by $\mathbb{N}$.
- The cardinality of a set S is denoted by $|S|$.
- Vectors are denoted by boldface, lower case alphabets for example, $\mathbf{x}, \mathbf{y}$. The zero vector is denoted by $\mathbf{0}$. A vector $\mathbf{x} \in \mathbb{R}^p$ will be in column format. The transpose of a vector is denoted by $\mathbf{x}^\top$. The i^{th} coordinate of a vector $\mathbf{x}$ is denoted by $\mathbf{x}_i$.
- Matrices are denoted by upper case alphabets for example, A, B. A_i denotes the i^{th} column of the matrix A and A^j denotes its j^{th} row. A_{ij} denotes the element at the i^{th} row and j^{th} column.
- For a vector $\mathbf{x} \in \mathbb{R}^p$ and a set $S \subset [p]$, the notation $\mathbf{x}_S$ denotes the vector $\mathbf{z} \in \mathbb{R}^p$ such that $\mathbf{z}_i = \mathbf{x}_i$ for $i \in S$, and $\mathbf{z}_i = 0$ otherwise. Similarly for matrices, A_S denotes the matrix B with $B_i = A_i$ for $i \in S$ and $B_i = \mathbf{0}$ for $i \neq S$. Also, A^S denotes the matrix B with $B^i = A^i$ for $i \in S$ and $B^i = \mathbf{0}^\top$ for $i \neq S$.
- The support of a vector $\mathbf{x}$ is denoted by $\text{supp}(x) := \{i : \mathbf{x}_i \neq 0\}$. A vector x is referred to as s-sparse if $|\text{supp}(x)| \leq s$.
- The canonical directions in $\mathbb{R}^p$ are denoted by $\mathbf{e}_i,\ i = 1, \ldots, p$.

- The identity matrix of order p is denoted by $I_{p\times p}$ or simply I_p. The subscript may be omitted when the order is clear from context.

- For a vector $\mathbf{x} \in \mathbb{R}^p$, the notation $\|x\|_q = \sqrt[q]{\sum_{i=1}^p |\mathbf{x}_i|^q}$ denotes its L_q norm. As special cases we define $\|\mathbf{x}\|_\infty := \max_i \ |\mathbf{x}_i|$, $\|\mathbf{x}\|_{-\infty} := \min_i \ |\mathbf{x}_i|$, and $\|\mathbf{x}\|_0 := |\mathrm{supp}(\mathbf{x})|$.

- Balls with respect to various norms are denoted as $\mathcal{B}_q(r) := \left\{\mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_q \le r\right\}$. As a special case the notation $\mathcal{B}_0(s)$ is used to denote the set of s-sparse vectors.

- For a matrix $A \in \mathbb{R}^{m\times n}$, $\sigma_1(A) \ge \sigma_2(A) \ge \ldots \ge \sigma_{\min\{m,n\}}(A)$ denote its singular values. The Frobenius norm of A is defined as $\|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\sum_i \sigma_i(A)^2}$. The nuclear norm of A is defined as $\|A\|_* := \sum_i \sigma_i(A)$.

- The trace of a square matrix $A \in \mathbb{R}^{m\times m}$ is defined as $\mathrm{tr}(A) = \sum_{i=1}^m A_{ii}$.

- The spectral norm (also referred to as the operator norm) of a matrix A is defined as $\|A\|_2 := \max_i \sigma_i(A)$.

- Random variables are denoted using upper case letters such as X, Y.

- The expectation of a random variable X is denoted by $\mathbb{E}[X]$. In cases where the distribution of X is to be made explicit, the notation $\mathbb{E}_{X\sim\mathcal{D}}[X]$, or else simply $\mathbb{E}_{\mathcal{D}}[X]$, is used.

- $\mathsf{Unif}(\mathcal{X})$ denotes the uniform distribution over a compact set $\mathcal{X}$.

- The standard *big-Oh* notation is used to describe the asymptotic behavior of functions. The *soft-Oh* notation is employed to hide poly-logarithmic factors i.e., $f = \widetilde{\mathcal{O}}(g)$ will imply $f = \mathcal{O}(g \log^c(g))$ for some absolute constant c.

Part I

Introduction and Basic Tools

1

Introduction

This section will set the stage for subsequent discussions by motivating some of the non-convex optimization problems we will be studying using real life examples, as well as setting up notation for the same.

1.1 Non-convex Optimization

The generic form of an analytic optimization problem is the following

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^p} \; & f(\mathbf{x}) \\ \text{s.t. } & \mathbf{x} \in \mathcal{C}, \end{aligned}$$

where $\mathbf{x}$ is the *variable* of the problem, $f : \mathbb{R}^p \to \mathbb{R}$ is the *objective function* of the problem, and $\mathcal{C} \subseteq \mathbb{R}^p$ is the *constraint set* of the problem. When used in a machine learning setting, the objective function allows the algorithm designer to encode proper and expected behavior for the machine learning model, such as fitting well to training data with respect to some loss function, whereas the constraint allows restrictions on the model to be encoded, for instance, restrictions on model size.

An optimization problem is said to be *convex* if the objective is a convex function, as well as the constraint set is a convex set. We refer

the reader to § 2 for formal definitions of these terms. An optimization problem that violates either one of these conditions, i.e., one that has a non-convex objective, or a non-convex constraint set, or both, is called a *non-convex* optimization problem. In this monograph, we will discuss non-convex optimization problems with non-convex objectives and convex constraints (§ 4, 5, 6, and 8), as well as problems with non-convex constraints but convex objectives (§ 3, 7, 9, 10, and 8). Such problems arise in a lot of application areas.

1.2 Motivation for Non-convex Optimization

Modern applications frequently require learning algorithms to operate in extremely high dimensional spaces. Examples include web-scale document classification problems where n-gram-based representations can have dimensionalities in the millions or more, recommendation systems with millions of items being recommended to millions of users, and signal processing tasks such as face recognition and image processing and bio-informatics tasks such as splice and gene detection, all of which present similarly high dimensional data.

Dealing with such high dimensionalities necessitates the imposition of structural constraints on the learning models being estimated from data. Such constraints are not only helpful in regularizing the learning problem, but often essential to prevent the problem from becoming ill-posed. For example, suppose we know how a user rates some items and wish to infer how this user would rate other items, possibly in order to inform future advertisement campaigns. To do so, it is essential to impose some structure on how a user's ratings for one set of items influences ratings for other kinds of items. Without such structure, it becomes impossible to infer any new user ratings. As we shall soon see, such structural constraints often turn out to be non-convex.

In other applications, the natural objective of the learning task is a non-convex function. Common examples include training deep neural networks and tensor decomposition problems. Although non-convex objectives and constraints allow us to accurately model learning problems, they often present a formidable challenge to algorithm designers.

This is because unlike convex optimization, we do not possess a handy set of tools for solving non-convex problems. Several non-convex optimization problems are known to be NP-hard to solve. The situation is made more bleak by a range of non-convex problems that are not only NP-hard to solve optimally, but NP-hard to solve approximately as well [Meka et al., 2008].

1.3 Examples of Non-Convex Optimization Problems

Below we present some areas where non-convex optimization problems arise naturally when devising learning problems.

Sparse Regression The classical problem of linear regression seeks to recover a linear model which can effectively predict a response variable as a linear function of covariates. For example, we may wish to predict the average expenditure of a household (the response) as a function of the education levels of the household members, their annual salaries and other relevant indicators (the covariates). The ability to do allows economic policy decisions to be more informed by revealing, for instance, how does education level affect expenditure.

More formally, we are provided a set of n covariate/response pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. The linear regression approach makes the modeling assumption $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \eta_i$ where $\mathbf{w}^* \in \mathbb{R}^p$ is the underlying linear model and η_i is some benign additive noise. Using the data provided $\{\mathbf{x}_i, y_i\}_{i=1,\ldots,n}$, we wish to recover back the model $\mathbf{w}^*$ as faithfully as possible.

A popular way to recover $\mathbf{w}^*$ is using the *least squares* formulation

$$\widehat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \mathbf{w} \right)^2 .$$

The linear regression problem as well as the least squares estimator, are extremely well studied and their behavior, precisely known. However, this age-old problem acquires new dimensions in situations where, either we expect only a few of the p features/covariates to be actually relevant to the problem but do not know their identity, or else are working in extremely data-starved settings i.e., $n \ll p$.

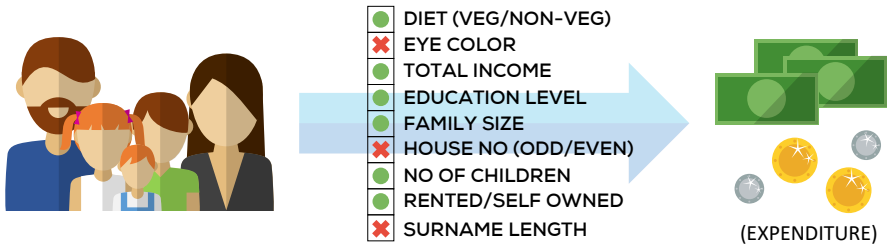


Figure 1.1: Not all available parameters and variables may be required for a prediction or learning task. Whereas the family size may significantly influence family expenditure, the eye color of family members does not directly or significantly influence it. Non-convex optimization techniques, such as sparse recovery, help discard irrelevant parameters and promote compact and accurate models.

The first problem often arises when there is an excess of covariates, several of which may be spurious or have no effect on the response. § 7 discusses several such practical examples. For now, consider the example depicted in Figure 1.1, that of expenditure prediction in a situation when the list of indicators include irrelevant ones such as whether the family lives in an odd-numbered house or not, which should arguably have no effect on expenditure. It is useful to eliminate such variables from consideration to promote consistency of the learned model.

The second problem is common in areas such as genomics and signal processing which face moderate to severe *data starvation* and the number of data points n available to estimate the model is small compared to the number of model parameters p to be estimated, i.e., $n \ll p$. Standard statistical approaches require at least $n \geq p$ data points to ensure a consistent estimation of all p model parameters and are unable to offer accurate model estimates in the face of data-starvation.

Both these problems can be handled by the *sparse recovery* approach, which seeks to fit a sparse model vector (i.e., a vector with say, no more than s non-zero entries) to the data. The least squares formulation, modified as a sparse recovery problem, is given below

$$\widehat{\mathbf{w}}_{\text{sp}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \mathbf{w}\right)^2$$
$$\text{s.t. } \mathbf{w} \in \mathcal{B}_0(s),$$

Although the objective function in the above formulation is convex, the constraint $\|\mathbf{w}\|_0 \leq s$ (equivalently $\mathbf{w} \in \mathcal{B}_0(s)$ – see list of mathematical notation at the beginning of this monograph) corresponds to a non-convex constraint set[1]. Sparse recovery effortlessly solves the twin problems of discarding irrelevant covariates and countering data-starvation since typically, only $n \geq s \log p$ (as opposed to $n \geq p$) data points are required for sparse recovery to work which drastically reduces the data requirement. Unfortunately however, sparse-recovery is an NP-hard problem [Natarajan, 1995].

Recommendation Systems Several internet search engines and e-commerce websites utilize recommendation systems to offer items to users that they would benefit from, or like, the most. The problem of recommendation encompasses benign recommendations for songs etc, all the way to critical recommendations in personalized medicine.

To be able to make accurate recommendations, we need very good estimates of how each user likes each item (song), or would benefit from it (drug). We usually have first-hand information for some user-item pairs, for instance if a user has specifically rated a song or if we have administered a particular drug on a user and seen the outcome. However, users typically rate only a handful of the hundreds of thousands of songs in any commercial catalog and it is not feasible, or even advisable, to administer every drug to a user. Thus, for the vast majority of user-item pairs, we have no direct information.

It is useful to visualize this problem as a *matrix completion* problem: for a set of m users $u_1, \ldots, u_m$ and n items $a_1, \ldots, a_n$, we have an $m \times n$ *preference matrix* $A = [A_{ij}]$ where A_{ij} encodes the preference of the i^{th} user for the j^{th} item. We are able to directly view only a small number of entries of this matrix, for example, whenever a user explicitly rates an item. However, we wish to recover the remaining entries, i.e., complete this matrix. This problem is closely linked to the *collaborative filtering* technique popular in recommendation systems.

Now, it is easy to see that unless there exists some structure in matrix, and by extension, in the way users rate items, there would be

[1] See Exercise 2.6.

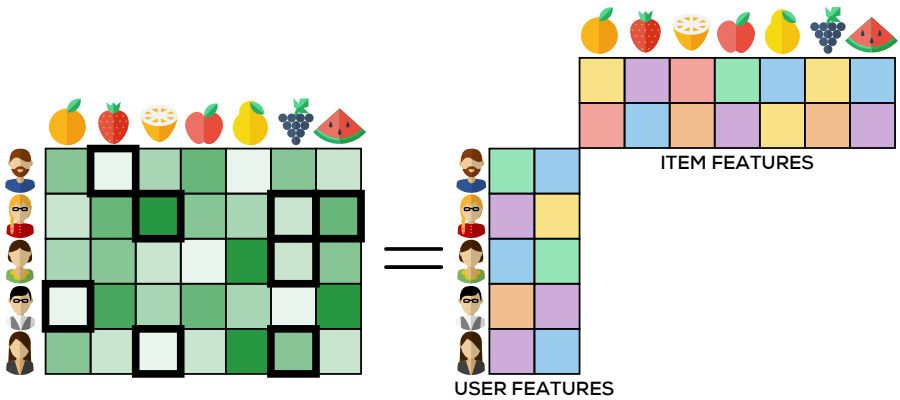


Figure 1.2: Only the entries of the ratings matrix with thick borders are observed. Notice that users rate infrequently and some items are not rated even once. Non-convex optimization techniques such as low-rank matrix completion can help recover the unobserved entries, as well as reveal hidden features that are descriptive of user and item properties, as shown on the right hand side.

no relation between the unobserved entries and the observed ones. This would result in there being no unique way to complete the matrix. Thus, it is essential to impose some structure on the matrix. A structural assumption popularly made is that of low rank: we wish to fill in the missing entries of A assuming that A is a low rank matrix. This can make the problem well-posed and have a unique solution since the additional low rank structure links the entries of the matrix together. The unobserved entries can no longer take values independently of the values observed by us. Figure 1.2 depicts this visually.

If we denote by $\Omega \subset [m] \times [n]$, the set of observed entries of A, then the low rank matrix completion problem can be written as

$$\widehat{A}_{\text{lr}} = \underset{X \in \mathbb{R}^{m \times n}}{\arg\min} \sum_{(i,j) \in \Omega} (X_{ij} - A_{ij})^2$$
$$\text{s.t. } \operatorname{rank}(X) \leq r,$$

This formulation also has a convex objective but a non-convex rank constraint[2]. This problem can be shown to be NP-hard as well. Interestingly, we can arrive at an alternate formulation by imposing the

[2]See Exercise 2.7.

low-rank constraint indirectly. It turns out that[3] assuming the ratings matrix to have rank at most r is equivalent to assuming that the matrix A can be written as $A = UV^\top$ with the matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ having at most r columns. This leads us to the following alternate formulation

$$\widehat{A}_{\text{lv}} = \operatorname*{arg\,min}_{\substack{U\in\mathbb{R}^{m\times r}\\ V\in\mathbb{R}^{n\times r}}} \sum_{(i,j)\in\Omega} \left(U_i^\top V_j - A_{ij}\right)^2.$$

There are no constraints in the formulation. However, the formulation requires joint optimization over a pair of variables (U, V) instead of a single variable. More importantly, it can be shown[4] that the objective function is non-convex in (U, V).

It is curious to note that the matrices U and V can be seen as encoding r-dimensional descriptions of users and items respectively. More precisely, for every user $i \in [m]$, we can think of the vector $U^i \in \mathbb{R}^r$ (i.e., the i-th row of the matrix U) as describing user i, and for every item $j \in [n]$, use the row vector $V^j \in \mathbb{R}^r$ to describe the item j in vectoral form. The rating given by user i to item j can now be seen to be $A_{ij} \approx \langle U^i, V^j \rangle$. Thus, recovering the rank r matrix A also gives us a bunch of r-dimensional latent vectors describing the users and items. These latent vectors can be extremely valuable in themselves as they can help us in understanding user behavior and item popularity, as well as be used in "content"-based recommendation systems which can effectively utilize item and user features.

The above examples, and several others from machine learning, such as low-rank tensor decomposition, training deep networks, and training structured models, demonstrate the utility of non-convex optimization in naturally modeling learning tasks. However, most of these formulations are NP-hard to solve exactly, and sometimes even approximately. In the following discussion, we will briefly introduce a few approaches, classical as well as contemporary, that are used in solving such non-convex optimization problems.

[3] See Exercise 3.3.
[4] See Exercise 4.1.

1.4 The Convex Relaxation Approach

Faced with the challenge of non-convexity, and the associated NP-hardness, a traditional workaround in literature has been to modify the problem formulation itself so that existing tools can be readily applied. This is often done by *relaxing* the problem so that it becomes a convex optimization problem. Since this allows familiar algorithmic techniques to be applied, the so-called *convex relaxation* approach has been widely studied. For instance, there exist relaxed, convex problem formulations for both the recommendation system and the sparse regression problems. For sparse linear regression, the relaxation approach gives us the popular LASSO formulation.

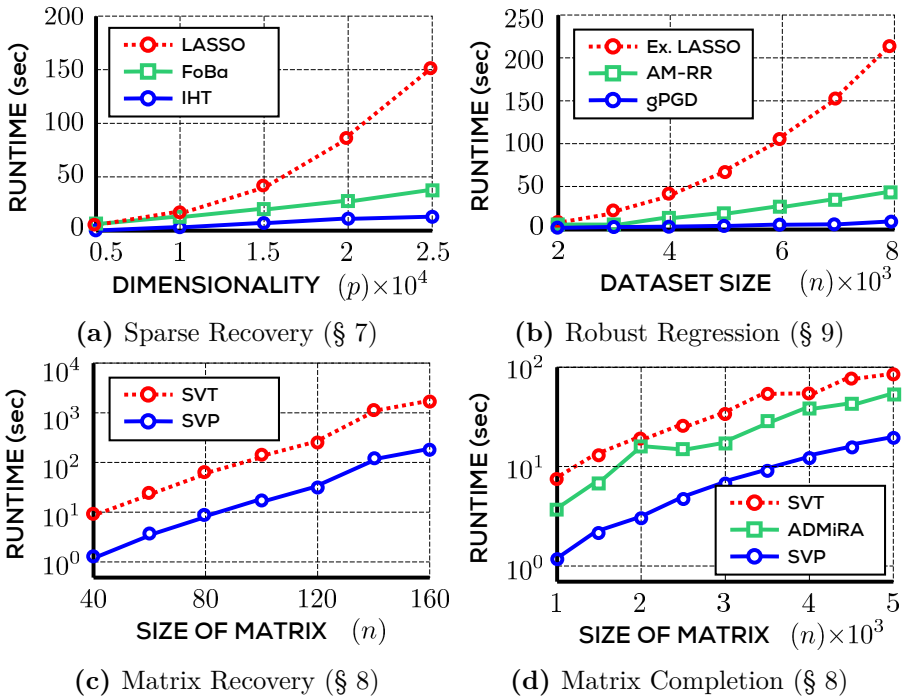
Now, in general, such modifications change the problem drastically, and the solutions of the relaxed formulation can be poor solutions to the original problem. However, it is known that if the problem possesses certain nice structure, then under careful relaxation, these distortions, formally referred to as a"relaxation gap", are absent, i.e., solutions to the relaxed problem would be optimal for the original non-convex problem as well.

Although a popular and successful approach, this still has limitations, the most prominent of them being scalability. Although the relaxed convex optimization problems are solvable in polynomial time, it is often challenging to solve them *efficiently* for large-scale problems.

1.5 The Non-Convex Optimization Approach

Interestingly, in recent years, a new wisdom has permeated the fields of machine learning and signal processing, one that advises not to relax the non-convex problems and instead solve them directly. This approach has often been dubbed the *non-convex optimization* approach owing to its goal of optimizing non-convex formulations directly.

Techniques frequently used in non-convex optimization approaches include simple and efficient primitives such as projected gradient descent, alternating minimization, the expectation-maximization algorithm, stochastic optimization, and variants thereof. These are very fast in practice and remain favorites of practitioners.



(a) Sparse Recovery (§ 7) **(b)** Robust Regression (§ 9)

(c) Matrix Recovery (§ 8) **(d)** Matrix Completion (§ 8)

Figure 1.3: An empirical comparison of run-times offered by various approaches to four different non-convex optimization problems. LASSO, extended LASSO, SVT are relaxation-based methods whereas IHT, gPGD, FoBa, AM-RR, SVP, ADMiRA are non-convex methods. In all cases, non-convex optimization techniques offer routines that are faster, often by an order of magnitude or more, than relaxation-based methods. Note that Figures 1.3c and 1.3d, employ a y-axis at logarithmic scale. The details of the methods are present in the sections linked with the respective figures.

At first glance, however, these efforts seem doomed to fail, given to the aforementioned NP-hardness results. However, in a series of deep and illuminating results, it has been repeatedly revealed that if the problem possesses nice structure, then not only do relaxation approaches succeed, but non-convex optimization algorithms do too. In such nice cases, non-convex approaches are able to only avoid NP-hardness, but actually offer provably optimal solutions. In fact, in practice, they often handsomely outperform relaxation-based approaches in terms of speed and scalability. Figure 1.3 illustrates this for some applications that we will investigate more deeply in later sections.

Very interestingly, it turns out that problem structures that allow non-convex approaches to avoid NP-hardness results, are very similar to those that allow their convex relaxation counterparts to avoid distortions and a large relaxation gap! Thus, it seems that if the problems possess nice structure, convex relaxation-based approaches, as well as non-convex techniques, both succeed. However, non-convex techniques usually offer more scalable solutions.

1.6 Organization and Scope

Our goal of this monograph is to present basic tools, both algorithmic and analytic, that are commonly used in the design and analysis of non-convex optimization algorithms, as well as present results which best represent the non-convex optimization philosophy. The presentation should enthuse, as well as equip, the interested reader and allow further readings, independent investigations, and applications of these techniques in diverse areas.

Given this broad aim, we shall appropriately restrict the number of areas we cover in this monograph, as well as the depth in which we cover each area. For instance, the literature abounds in results that seek to perform optimizations with more and more complex structures being imposed - from sparse recovery to low rank matrix recovery to low rank tensor recovery. However, we shall restrict ourselves from venturing too far into these progressions. Similarly, within the problem of sparse recovery, there exist results for recovery in the simple least squares setting, the more involved setting of sparse M-estimation, as well as the still more involved setting of sparse M-estimation in the presence of outliers. Whereas we will cover sparse least squares estimation in depth, we will refrain from delving too deeply into the more involved sparse M-estimation problems.

That being said, the entire presentation will be self contained and accessible to anyone with a basic background in algebra and probability theory. Moreover, the bibliographic notes given at the end of the sections will give pointers that should enable the reader to explore the state of the art not covered in this monograph.

References

Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *SIAM Journal of Optimization*, 26(4):2775–2799, 2016.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding Approximate Local Minima Faster than Gradient Descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2017.

Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, pages 81–102, 2016.

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.

Andreas Andresen and Vladimir Spokoiny. Convergence of an Alternating Maximization Procedure. *Journal of Machine Learning Research*, 17:1–53, 2016.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, 2014.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Anima Anandkumar. Reinforcement Learning of POMDPs using Spectral Methods. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, 2016.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical Guarantees for the EM Algorithm: From Population to Sample-based Analysis. *Annals of Statistics*, 45(1):77–120, 2017.

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.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3rd edition, 2016.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust Regression via Hard Thresholding. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent Robust Regression. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

Srinadh Bhojanapalli and Prateek Jain. Universal Matrix Completion. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Thomas Blumensath. Sampling and Reconstructing Signals From a Union of Linear Subspaces. *IEEE Transactions on Information Theory*, 57(7): 4660–4671, 2011.

Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, and Denka Kutzarova. Explicit constructions of RIP matrices and related problems. *Duke Mathematical Journal*, 159(1):145–185, 2011.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Alon Brutzkus and Amir Globerson. Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Sebastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(34):231–357, 2015.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal of Optimization*, 20(4):1956–1982, 2010.

Emmanuel Candès and Terence Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Emmanuel J. Candès. The Restricted Isometry Property and Its Implications for Compressed Sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

Emmanuel J. Candès and Xiaodong Li. Solving Quadratic Equations via PhaseLift When There Are About as Many Equations as Unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

Emmanuel J. Candès and Benjamin Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6): 717–772, 2009.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56 (5):2053–2080, 2009.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable Signal Recovery from Incomplete and Inaccurate Measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

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.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Rick Chartrand. Exact Reconstruction of Sparse Signals via Nonconvex Minimization. *IEEE Information Processing Letters*, 14(10):707–710, 2007.

Caihua Chen and Bingsheng He. Matrix Completion via an Alternating Direction Method. *IMA Journal of Numerical Analysis*, 32(1):227–245, 2012.

Laming Chen and Yuantao Gu. Local and global optimality of LP minimization for sparse recovery. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Yin Chen and Arnak S. Dalalyan. Fused sparsity and robust estimation for linear models with unknown variance. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust Sparse Regression under Adversarial Corruption. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Matrix Completion with Column Manipulation: Near-Optimal Sample-Robustness-Rank Tradeoffs. *IEEE Transactions on Information Theory*, 62(1):503–526, 2016.

Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly-optimal Robust Matrix Completion. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Anna Choromanska, Mikael Hena, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The Loss Surfaces of Multilayer Networks. In *Proceedings of the 18th International Conference on Arti cial Intelligence and Statistics (AISTATS)*, 2015.

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed Sensing and Best k-term Approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2009.

Christophe Croux and Kristel Joossens. Robust Estimation of the Vector Autoregressive Model by a Least Trimmed Squares Procedure. In *Proceedings in Computational Statistics (COMPSTAT)*, 2008.

Yann N. Dauphin, Razvan Pascanu, Çaglar Gülçehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2933–2941, 2014.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

David L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction. *Proceedings of the National Academy of Sciences USA*, 106(45):18914–18919, 2009.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications in system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Simon Foucart. A Note on Guaranteed Sparse Recovery via ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

Simon Foucart. Hard Thresholding Pursuit: an Algorithm for Compressive Sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

Rosalind Franklin and Raymond G. Gosling. Evidence for 2-Chain Helix in Crystalline Structure of Sodium Deoxyribonucleate. *Nature*, 172:156–157, 1953a.

Rosalind Franklin and Raymond G. Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171:740–741, 1953b.

Rahul Garg and Rohit Khandekar. Gradient Descent with Sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

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 (COLT)*, pages 797–842, 2015.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix Completion has No Spurious Local Minimum. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

R. W. Gerchberg and W. Owen Saxton. A Practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures. *Optik*, 35 (2):237–246, 1972.

Surbhi Goel and Adam Klivans. Learning Depth-Three Neural Networks in Polynomial Time. arXiv:1709.06010v1 [cs.DS], 2017.

Donald Goldfarb and Shiqian Ma. Convergence of Fixed-Point Continuation Algorithms for Matrix Rank Minimization. *Foundations of Computational Mathematics*, 11(2):183–210, 2011.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in Mathematical Sciences. The John Hopkins University Press, 3rd edition, 1996.

Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sparse and Spurious: Dictionary Learning With Noise and Outliers. *IEEE Transaction on Information Theory*, 61(11):6298–6319, 2015.

Moritz Hardt. Understanding Alternating Minimization for Matrix Completion. In *Proceedings of the 55th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2014.

Moritz Hardt and Mary Wootters. Fast Matrix Completion Without the Condition Number. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, 2014.

Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, 2014.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Number 143 in Monographs on Statistics and Applied Probability. The CRC Press, 2016.

Ishay Haviv and Oded Regev. The Restricted Isometry Property of Subsampled Fourier Matrices. In Bo'az Klartag and Emanuel Milman, editors, *Geometric Aspects of Functional Analysis*, volume 2169 of *Lecture Notes in Mathematics*, pages 163–179. Springer, Cham, 2017.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 2009.

Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Sparse Phase Retrieval: Convex Algorithms and Limitations. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2013.

Prateek Jain and Praneeth Netrapalli. Fast Exact Matrix Completion with Finite Samples. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, 2015.

Prateek Jain and Ambuj Tewari. Alternating Minimization for Regression Problems with Vector-valued Outputs. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

Prateek Jain, Raghu Meka, and Inderjit Dhillon. Guaranteed Rank Minimization via Singular Value Projections. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

Prateek Jain, Ambuj Tewari, and Inderjit S. Dhillon. Orthogonal Matching Pursuit with Replacement. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank Matrix Completion using Alternating Minimization. In *Proceedings of the 45th annual ACM Symposium on Theory of Computing (STOC)*, pages 665–674, 2013.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On Iterative Hard Thresholding Methods for High-dimensional M-Estimation. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

Ali Jalali, Christopher C Johnson, and Pradeep D Ravikumar. On Learning Discrete Graphical Models using Greedy Methods. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1935–1943, 2011.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to Escape Saddle Points Efficiently. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1724–1732, 2017.

Boris Sergeevich Kashin. The diameters of octahedra. *Uspekhi Matematicheskikh Nauk*, 30(4(184)):251–252, 1975.

Raghunandan H. Keshavan. *Efficient Algorithms for Collaborative Filtering*. Ph.D. Thesis, Stanford University, 2012.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion from a Few Entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for. Recommender Systems. *IEEE Computer*, 42(8):30–37, 2009.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient Descent Only Converges to Minimizers. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, pages 1246–1257, 2016.

Kiryung Lee and Yoram Bresler. ADMiRA: Atomic Decomposition for Minimum Rank Approximation. *IEEE Transactions on Information Theory*, 56 (9):4402–4416, 2010.

Yuanzhi Li and Yang Yuan. Convergence Analysis of Two-layer Neural Networks with ReLU Activation. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Amand A. Lucas. A-DNA and B-DNA: Comparing Their Historical X-ray Fiber Diffraction Images. *Journal of Chemical Education*, 85(5):737–743, 2008.

Zhi-Quan Luo and Paul Tseng. On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yoha. *Robust Statistics: Theory and Methods*. John Wiley, 2006.

R. Douglas Martin and Judy Zeh. Robust Generalized M-estimates for Autoregressive Parameters: Small-sample Behavior and Applications. Technical Report 214, University of Washington, 1978.

Raghu Meka, Prateek Jain, Constantine Caramanis, and Inderjit Dhillon. Rank Minimization via Online Learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Deanna Needell and Joel A. Tropp. CoSaMP: Iterative Signal Recovery from Incomplete and Inaccurate Samples. *Applied and Computational Harmonic Analysis*, 26:301–321, 2008.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, 2012.

Jelani Nelson, Eric Price, and Mary Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer-Academic, 2003.

Yurii Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal of Optimization*, 22(2):341–362, 2012.

Yurii Nesterov and B.T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase Retrieval using Alternating Minimization. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

Nam H. Nguyen and Trac D. Tran. Exact recoverability from dense corrupted observations via L1 minimization. *IEEE Transactions on Information Theory*, 59(4):2036–2058, 2013a.

Nam H Nguyen and Trac D Tran. Robust Lasso With Missing and Grossly Corrupted Observations. *IEEE Transaction on Information Theory*, 59(4): 2036–2058, 2013b.

Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp Time-Data Tradeoffs for Linear Inverse Problems. arXiv:1507.04793 [cs.IT], 2015.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

Benjamin Recht. A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed Minimum Rank Solutions to Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.

Sashank Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alexander J. Smola. Fast Stochastic Methods for Nonsmooth Nonconvex Optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Peter J. Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987.

Ankan Saha and Ambuj Tewari. On the Non-asymptotic Convergence of Cyclic Coordinate Descent Methods. *SIAM Journal on Optimization*, 23 (1):576–601, 2013.

Hanie Sedghi and Anima Anandkumar. Training Input-Output Recurrent Neural Networks through Spectral Methods. arXiv:1603.00954 [CS.LG], 2016.

Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.

Yiyuan She and Art B. Owen. Outlier Detection Using Nonconvex Penalized Regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

Daniel A. Spielman, Huan Wang, and John Wright. Exact Recovery of Sparsely-Used Dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors. *Optimization for Machine Learning*. The MIT Press, 2011.

Norbert Stockinger and Rudolf Dutter. Robust time series analysis: A survey. *Kybernetika*, 23(7):1–3, 1987.

Ju Sun, Qing Qu, and John Wright. When Are Nonconvex Problems Not Scary? arXiv:1510.06096 [math.OC], 2015.

Ruoyu Sun and Zhi-Quan Lu. Guaranteed Matrix Completion via Non-convex Factorization. In *Proceedings of the 56th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2015.

Ambuj Tewari, Pradeep Ravikumar, and Inderjit S. Dhillon. Greedy Algorithms for Structurally Constrained High Dimensional Problems. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Joel A. Tropp and Anna C. Gilbert. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, Dec. 2007. ISSN 0018-9448.

Meng Wang, Weiyu Xu, and Ao Tang. On the Performance of Sparse Recovery Via ℓ_p-Minimization ($0 \leq p \leq 1$). *IEEE Transactions on Information Theory*, 57(11):7255–7278, 2011.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

Karen H.S. Wilson, Sarah E. Eckenrode, Quan-Zhen Li, Qing-Guo Ruan, Ping Yang, Jing-Da Shi, Abdoreza Davoodi-Semiromi, Richard A. McIndoe, Byron P. Croker, and Jin-Xiong She. Microarray Analysis of Gene Expression in the Kidneys of New- and Post-Onset Diabetic NOD Mice. *Diabetes*, 52 (8):2151–2159, 2003.

John Wright, Alan Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

Stephen J Wright and Jorge Nocedal. *Numerical Optimization*, volume 2. Springer New York, 1999.

C.-F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Allen Y. Yang, Zihan Zhou, Arvind Ganesh Balasubramanian, S Shankar Sastry, and Yi Ma. Fast ℓ_1-Minimization Algorithms for Robust Face Recognition. *IEEE Transactions on Image Processing*, 22(8):3234–3246, 2013.

Fanny Yang, Sivaraman Balakrishnan, and Martin J. Wainwright. Statistical and computational guarantees for the Baum-Welch algorithm. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015.

Xinyang Yi and Constantine Caramanis. Regularized EM Algorithms: A Unified Framework and Statistical Guarantees. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating Minimization for Mixed Linear Regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Ya-Xiang Yuan. Recent advances in trust region algorithms. *Mathematical Programming*, 151(1):249–281, 2015.

Tong Zhang. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Transactions on Information Theory*, 57: 4689–4708, 2011.

Yuchen Zhang, Percy Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proceedings of the 30th Conference on Learning Theory*, 2017.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery Guarantees for One-hidden-layer Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale Parallel Collaborative Filtering for the Netflix Prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management (AAIM)*, 2008.