Convex Optimization: Algorithms and Complexity

Sébastien Bubeck
Theory Group, Microsoft Research
sebubeck@microsoft.com

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

S. Bubeck. *Convex Optimization: Algorithms and Complexity*. Foundations and Trends® in Machine Learning, vol. 8, no. 3-4, pp. 231–357, 2015.

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

ISBN: 978-1-60198-861-4

Foundations and Trends® in Machine Learning

Volume 8, Issue 3-4, 2015

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, 2015, Volume 8, 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. 8, No. 3-4 (2015) 231–357

DOI: 10.1561/2200000050

Convex Optimization: Algorithms and Complexity

Sébastien Bubeck
Theory Group, Microsoft Research
sebubeck@microsoft.com

Contents

Abstract

Ţ This monograph presents the main complexity theorems in convex optimization and their corresponding algorithms. Starting from the fundamental theory of black-box optimization, the material progresses towards recent advances in structural optimization and stochastic optimization. Our presentation of black-box optimization, strongly influenced by Nesterov's seminal book and Nemirovski's lecture notes, includes the analysis of cutting plane methods, as well as (accelerated) gradient descent schemes. We also pay special attention to non-Euclidean settings (relevant algorithms include Frank-Wolfe, mirror descent, and dual averaging) and discuss their relevance in machine learning. We provide a gentle introduction to structural optimization with FISTA (to optimize a sum of a smooth and a simple non-smooth term), saddle-point mirror prox (Nemirovski's alternative to Nesterov's smoothing), and a concise description of interior point methods. In stochastic optimization we discuss stochastic gradient descent, mini-batches, random coordinate descent, and sublinear algorithms. We also briefly touch upon convex relaxation of combinatorial problems and the use of randomness to round solutions, as well as random walks based methods.

S. Bubeck. *Convex Optimization: Algorithms and Complexity*. Foundations and Trends® in Machine Learning, vol. 8, no. 3-4, pp. 231–357, 2015.
DOI: 10.1561/2200000050.

1

Introduction

The central objects of our study are convex functions and convex sets in $\mathbb{R}^n$.

Definition 1.1 (Convex sets and convex functions)**.** A set $\mathcal{X} \subset \mathbb{R}^n$ is said to be convex if it contains all of its segments, that is

$$\forall (x, y, \gamma) \in \mathcal{X} \times \mathcal{X} \times [0, 1],\ (1 - \gamma)x + \gamma y \in \mathcal{X}.$$

A function $f : \mathcal{X} \to \mathbb{R}$ is said to be convex if it always lies below its chords, that is

$$\forall (x, y, \gamma) \in \mathcal{X} \times \mathcal{X} \times [0, 1],\ f((1 - \gamma)x + \gamma y) \leq (1 - \gamma)f(x) + \gamma f(y).$$

We are interested in algorithms that take as input a convex set $\mathcal{X}$ and a convex function f and output an approximate minimum of f over $\mathcal{X}$. We write compactly the problem of finding the minimum of f over $\mathcal{X}$ as

$$\begin{aligned} &\text{min. } f(x) \\ &\text{s.t. } x \in \mathcal{X}. \end{aligned}$$

In the following we will make more precise how the set of constraints $\mathcal{X}$ and the objective function f are specified to the algorithm. Before that we proceed to give a few important examples of convex optimization problems in machine learning.

1.1 Some convex optimization problems in machine learning

Many fundamental convex optimization problems in machine learning take the following form:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x) + \lambda \mathcal{R}(x), \tag{1.1}$$

where the functions $f_1, \ldots, f_m, \mathcal{R}$ are convex and $\lambda \geq 0$ is a fixed parameter. The interpretation is that $f_i(x)$ represents the cost of using x on the i^{th} element of some data set, and $\mathcal{R}(x)$ is a regularization term which enforces some "simplicity" in x. We discuss now major instances of (1.1). In all cases one has a data set of the form $(w_i, y_i) \in \mathbb{R}^n \times \mathcal{Y}, i = 1, \ldots, m$ and the cost function f_i depends only on the pair (w_i, y_i). We refer to Hastie et al. [2001], Schölkopf and Smola [2002], Shalev-Shwartz and Ben-David [2014] for more details on the origin of these important problems. The mere objective of this section is to expose the reader to a few concrete convex optimization problems which are routinely solved.

In classification one has $\mathcal{Y} = \{-1, 1\}$. Taking $f_i(x) = \max(0, 1 - y_i x^\top w_i)$ (the so-called hinge loss) and $\mathcal{R}(x) = \|x\|_2^2$ one obtains the SVM problem. On the other hand taking $f_i(x) = \log(1+\exp(-y_i x^\top w_i))$ (the logistic loss) and again $\mathcal{R}(x) = \|x\|_2^2$ one obtains the (regularized) logistic regression problem.

In regression one has $\mathcal{Y} = \mathbb{R}$. Taking $f_i(x) = (x^\top w_i - y_i)^2$ and $\mathcal{R}(x) = 0$ one obtains the vanilla least-squares problem which can be rewritten in vector notation as

$$\min_{x \in \mathbb{R}^n} \|Wx - Y\|_2^2,$$

where $W \in \mathbb{R}^{m \times n}$ is the matrix with $w_i^\top$ on the i^{th} row and $Y = (y_1, \ldots, y_n)^\top$. With $\mathcal{R}(x) = \|x\|_2^2$ one obtains the ridge regression problem, while with $\mathcal{R}(x) = \|x\|_1$ this is the LASSO problem Tibshirani [1996].

Our last two examples are of a slightly different flavor. In particular the design variable x is now best viewed as a matrix, and thus we denote it by a capital letter X. The sparse inverse covariance estimation problem can be written as follows, given some empirical covariance matrix Y,

$$\begin{aligned} &\text{min. } \mathrm{Tr}(XY) - \mathrm{logdet}(X) + \lambda \|X\|_1 \\ &\text{s.t. } X \in \mathbb{R}^{n \times n}, X^\top = X, X \succeq 0. \end{aligned}$$

Intuitively the above problem is simply a regularized maximum likelihood estimator (under a Gaussian assumption).

Finally we introduce the convex version of the matrix completion problem. Here our data set consists of observations of some of the entries of an unknown matrix Y, and we want to "complete" the unobserved entries of Y in such a way that the resulting matrix is "simple" (in the sense that it has low rank). After some massaging (see Candès and Recht [2009]) the (convex) matrix completion problem can be formulated as follows:

$$\begin{aligned} &\text{min. } \mathrm{Tr}(X) \\ &\text{s.t. } X \in \mathbb{R}^{n \times n}, X^\top = X, X \succeq 0, X_{i,j} = Y_{i,j} \text{ for } (i,j) \in \Omega, \end{aligned}$$

where $\Omega \subset [n]^2$ and $(Y_{i,j})_{(i,j)\in\Omega}$ are given.

1.2 Basic properties of convexity

A basic result about convex sets that we shall use extensively is the Separation Theorem.

Theorem 1.1 (Separation Theorem)**.** Let $\mathcal{X} \subset \mathbb{R}^n$ be a closed convex set, and $x_0 \in \mathbb{R}^n \setminus \mathcal{X}$. Then, there exists $w \in \mathbb{R}^n$ and $t \in \mathbb{R}$ such that

$$w^\top x_0 < t, \text{ and } \forall x \in \mathcal{X}, w^\top x \geq t.$$

Note that if $\mathcal{X}$ is not closed then one can only guarantee that $w^\top x_0 \leq w^\top x, \forall x \in \mathcal{X}$ (and $w \neq 0$). This immediately implies the Supporting Hyperplane Theorem ($\partial\mathcal{X}$ denotes the boundary of $\mathcal{X}$, that is the closure without the interior):

Theorem 1.2 (Supporting Hyperplane Theorem)**.** Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set, and $x_0 \in \partial\mathcal{X}$. Then, there exists $w \in \mathbb{R}^n, w \neq 0$ such that

$$\forall x \in \mathcal{X}, w^\top x \geq w^\top x_0.$$

We introduce now the key notion of *subgradients*.

Definition 1.2 (Subgradients)**.** Let $\mathcal{X} \subset \mathbb{R}^n$, and $f : \mathcal{X} \to \mathbb{R}$. Then $g \in \mathbb{R}^n$ is a subgradient of f at $x \in \mathcal{X}$ if for any $y \in \mathcal{X}$ one has

$$f(x) - f(y) \leq g^\top (x - y).$$

The set of subgradients of f at x is denoted $\partial f(x)$.

To put it differently, for any $x \in \mathcal{X}$ and $g \in \partial f(x)$, f is above the linear function $y \mapsto f(x)+g^\top(y-x)$. The next result shows (essentially) that a convex functions always admit subgradients.

Proposition 1.1 (Existence of subgradients)**.** Let $\mathcal{X} \subset \mathbb{R}^n$ be convex, and $f : \mathcal{X} \to \mathbb{R}$. If $\forall x \in \mathcal{X}, \partial f(x) \neq \emptyset$ then f is convex. Conversely if f is convex then for any $x \in \mathrm{int}(\mathcal{X}), \partial f(x) \neq \emptyset$. Furthermore if f is convex and differentiable at x then $\nabla f(x) \in \partial f(x)$.

Before going to the proof we recall the definition of the epigraph of a function $f : \mathcal{X} \to \mathbb{R}$:

$$\mathrm{epi}(f) = \{(x, t) \in \mathcal{X} \times \mathbb{R} : t \geq f(x)\}.$$

It is obvious that a function is convex if and only if its epigraph is a convex set.

Proof. The first claim is almost trivial: let $g \in \partial f((1-\gamma)x + \gamma y)$, then by definition one has

$$\begin{aligned} f((1-\gamma)x + \gamma y) &\leq f(x) + \gamma g^\top (y - x), \\ f((1-\gamma)x + \gamma y) &\leq f(y) + (1-\gamma) g^\top (x - y), \end{aligned}$$

which clearly shows that f is convex by adding the two (appropriately rescaled) inequalities.

Now let us prove that a convex function f has subgradients in the interior of $\mathcal{X}$. We build a subgradient by using a supporting hyperplane to the epigraph of the function. Let $x \in \mathcal{X}$. Then clearly $(x, f(x)) \in \partial\mathrm{epi}(f)$, and $\mathrm{epi}(f)$ is a convex set. Thus by using the Supporting Hyperplane Theorem, there exists $(a, b) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$a^\top x + b f(x) \geq a^\top y + bt, \forall (y, t) \in \mathrm{epi}(f). \tag{1.2}$$

Clearly, by letting t tend to infinity, one can see that $b \leq 0$. Now let us assume that x is in the interior of $\mathcal{X}$. Then for $\varepsilon > 0$ small enough, $y = x + \varepsilon a \in \mathcal{X}$, which implies that b cannot be equal to 0 (recall that if $b = 0$ then necessarily $a \neq 0$ which allows to conclude by contradiction). Thus rewriting (1.2) for $t = f(y)$ one obtains

$$f(x) - f(y) \leq \frac{1}{|b|} a^\top (x - y).$$

Thus $a/|b| \in \partial f(x)$ which concludes the proof of the second claim.

Finally let f be a convex and differentiable function. Then by definition:

$$\begin{aligned} f(y) &\geq \frac{f((1-\gamma)x + \gamma y) - (1-\gamma) f(x)}{\gamma} \\ &= f(x) + \frac{f(x + \gamma(y - x)) - f(x)}{\gamma} \\ &\underset{\gamma \to 0}{\to} f(x) + \nabla f(x)^\top (y - x), \end{aligned}$$

which shows that $\nabla f(x) \in \partial f(x)$. $\square$

In several cases of interest the set of contraints can have an empty interior, in which case the above proposition does not yield any information. However it is easy to replace $\mathrm{int}(\mathcal{X})$ by $\mathrm{ri}(\mathcal{X})$ -the relative interior of $\mathcal{X}$- which is defined as the interior of $\mathcal{X}$ when we view it as subset of the affine subspace it generates. Other notions of convex analysis will prove to be useful in some parts of this text. In particular the notion of *closed convex functions* is convenient to exclude pathological cases: these are the convex functions with closed epigraphs. Sometimes it is also useful to consider the extension of a convex function $f : \mathcal{X} \to \mathbb{R}$ to a function from $\mathbb{R}^n$ to $\overline{\mathbb{R}}$ by setting $f(x) = +\infty$ for $x \notin \mathcal{X}$. In convex analysis one uses the term *proper convex function* to denote a convex function with values in $\mathbb{R} \cup \{+\infty\}$ such that there exists $x \in \mathbb{R}^n$ with $f(x) < +\infty$. **From now on all convex functions will be closed, and if necessary we consider also their proper extension.** We refer the reader to Rockafellar [1970] for an extensive discussion of these notions.

1.3 Why convexity?

The key to the algorithmic success in minimizing convex functions is that these functions exhibit a *local to global* phenomenon. We have already seen one instance of this in Proposition 1.1, where we showed that $\nabla f(x) \in \partial f(x)$: the gradient $\nabla f(x)$ contains a priori only local information about the function f around x while the subdifferential $\partial f(x)$ gives a global information in the form of a linear lower bound on the entire function. Another instance of this local to global phenomenon is that local minima of convex functions are in fact global minima:

Proposition 1.2 (Local minima are global minima)**.** Let f be convex. If x is a local minimum of f then x is a global minimum of f. Furthermore this happens if and only if $0 \in \partial f(x)$.

Proof. Clearly $0 \in \partial f(x)$ if and only if x is a global minimum of f. Now assume that x is local minimum of f. Then for γ small enough one has for any y,

$$f(x) \leq f((1-\gamma)x + \gamma y) \leq (1-\gamma) f(x) + \gamma f(y),$$

which implies $f(x) \leq f(y)$ and thus x is a global minimum of f. $\square$

The nice behavior of convex functions will allow for very fast algorithms to optimize them. This alone would not be sufficient to justify the importance of this class of functions (after all constant functions are pretty easy to optimize). However it turns out that surprisingly many optimization problems admit a convex (re)formulation. The excellent book Boyd and Vandenberghe [2004] describes in great details the various methods that one can employ to uncover the convex aspects of an optimization problem. We will not repeat these arguments here, but we have already seen that many famous machine learning problems (SVM, ridge regression, logistic regression, LASSO, sparse covariance estimation, and matrix completion) are formulated as convex problems.

We conclude this section with a simple extension of the optimality condition "$0 \in \partial f(x)$" to the case of constrained optimization. We state this result in the case of a differentiable function for sake of simplicity.

Proposition 1.3 (First order optimality condition). Let f be convex and $\mathcal{X}$ a closed convex set on which f is differentiable. Then

$$x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} f(x),$$

if and only if one has

$$\nabla f(x^*)^\top (x^* - y) \leq 0, \forall y \in \mathcal{X}.$$

Proof. The "if" direction is trivial by using that a gradient is also a subgradient. For the "only if" direction it suffices to note that if $\nabla f(x)^\top (y - x) < 0$, then f is locally decreasing around x on the line to y (simply consider $h(t) = f(x + t(y - x))$ and note that $h'(0) = \nabla f(x)^\top (y - x)$). $\square$

1.4 Black-box model

We now describe our first model of "input" for the objective function and the set of constraints. In the black-box model we assume that we have unlimited computational resources, the set of constraint $\mathcal{X}$ is known, and the objective function $f : \mathcal{X} \to \mathbb{R}$ is unknown but can be accessed through queries to *oracles*:

- A zeroth order oracle takes as input a point $x \in \mathcal{X}$ and outputs the value of f at x.
- A first order oracle takes as input a point $x \in \mathcal{X}$ and outputs a subgradient of f at x.

In this context we are interested in understanding the *oracle complexity* of convex optimization, that is how many queries to the oracles are necessary and sufficient to find an ε-approximate minima of a convex function. To show an upper bound on the sample complexity we need to propose an algorithm, while lower bounds are obtained by information theoretic reasoning (we need to argue that if the number of queries is "too small" then we don't have enough information about the function to identify an ε-approximate solution).

From a mathematical point of view, the strength of the black-box model is that it will allow us to derive a *complete* theory of convex optimization, in the sense that we will obtain matching upper and lower bounds on the oracle complexity for various subclasses of interesting convex functions. While the model by itself does not limit our computational resources (for instance any operation on the constraint set $\mathcal{X}$ is allowed) we will of course pay special attention to the algorithms' *computational complexity* (i.e., the number of elementary operations that the algorithm needs to do). We will also be interested in the situation where the set of constraint $\mathcal{X}$ is unknown and can only be accessed through a *separation oracle*: given $x \in \mathbb{R}^n$, it outputs either that x is in $\mathcal{X}$, or if $x \notin \mathcal{X}$ then it outputs a separating hyperplane between x and $\mathcal{X}$.

The black-box model was essentially developed in the early days of convex optimization (in the Seventies) with Nemirovski and Yudin [1983] being still an important reference for this theory (see also Nemirovski [1995]). In the recent years this model and the corresponding algorithms have regained a lot of popularity, essentially for two reasons:

- It is possible to develop algorithms with dimension-free oracle complexity which is quite attractive for optimization problems in very high dimension.

- Many algorithms developed in this model are robust to noise in the output of the oracles. This is especially interesting for stochastic optimization, and very relevant to machine learning applications. We will explore this in details in Chapter 6.

Chapter 2, Chapter 3 and Chapter 4 are dedicated to the study of the black-box model (noisy oracles are discussed in Chapter 6). We do not cover the setting where only a zeroth order oracle is available, also called derivative free optimization, and we refer to Conn et al. [2009], Audibert et al. [2011] for further references on this.

1.5 Structured optimization

The black-box model described in the previous section seems extremely wasteful for the applications we discussed in Section 1.1. Consider for instance the LASSO objective: $x \mapsto \|Wx - y\|_2^2 + \|x\|_1$. We know this function *globally*, and assuming that we can only make local queries through oracles seem like an artificial constraint for the design of algorithms. Structured optimization tries to address this observation. Ultimately one would like to take into account the global structure of both f and $\mathcal{X}$ in order to propose the most efficient optimization procedure. An extremely powerful hammer for this task are the Interior Point Methods. We will describe this technique in Chapter 5 alongside with other more recent techniques such as FISTA or Mirror Prox.

We briefly describe now two classes of optimization problems for which we will be able to exploit the structure very efficiently, these are the LPs (Linear Programs) and SDPs (Semi-Definite Programs). Ben-Tal and Nemirovski [2001] describe a more general class of Conic Programs but we will not go in that direction here.

The class LP consists of problems where $f(x) = c^\top x$ for some $c \in \mathbb{R}^n$, and $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$ for some $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

The class SDP consists of problems where the optimization variable is a symmetric matrix $X \in \mathbb{R}^{n \times n}$. Let $\mathbb{S}^n$ be the space of $n \times n$ symmetric matrices (respectively $\mathbb{S}^n_+$ is the space of positive semi-definite matrices), and let $\langle \cdot, \cdot \rangle$ be the Frobenius inner product (recall that it can be written as $\langle A, B \rangle = \mathrm{Tr}(A^\top B)$). In the class SDP the problems are of the following form: $f(x) = \langle X, C \rangle$ for some $C \in \mathbb{R}^{n \times n}$, and $\mathcal{X} = \{X \in \mathbb{S}^n_+ : \langle X, A_i \rangle \leq b_i, i \in \{1, \dots, m\}\}$ for some $A_1, \dots, A_m \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^m$. Note that the matrix completion problem described in Section 1.1 is an example of an SDP.

1.6 Overview of the results and disclaimer

The overarching aim of this monograph is to present the main complexity theorems in convex optimization and the corresponding algorithms. We focus on five major results in convex optimization which give the overall structure of the text: the existence of efficient cutting-plane methods with optimal oracle complexity (Chapter 2), a complete characterization of the relation between first order oracle complexity and curvature in the objective function (Chapter 3), first order methods beyond Euclidean spaces (Chapter 4), non-black box methods (such as interior point methods) can give a quadratic improvement in the number of iterations with respect to optimal black-box methods (Chapter 5), and finally noise robustness of first order methods (Chapter 6). Table 1.1 can be used as a quick reference to the results proved in Chapter 2 to Chapter 5, as well as some of the results of Chapter 6 (this last chapter is the most relevant to machine learning but the results are also slightly more specific which make them harder to summarize).

An important disclaimer is that the above selection leaves out methods derived from duality arguments, as well as the two most popular research avenues in convex optimization: (i) using convex optimization in non-convex settings, and (ii) practical large-scale algorithms. Entire books have been written on these topics, and new books have yet to be written on the impressive collection of new results obtained for both (i) and (ii) in the past five years.

A few of the blatant omissions regarding (i) include (a) the theory of submodular optimization (see Bach [2013]), (b) convex relaxations of combinatorial problems (a short example is given in Section 6.6), and (c) methods inspired from convex optimization for non-convex problems such as low-rank matrix factorization (see e.g. Jain et al. [2013] and references therein), neural networks optimization, etc.

With respect to (ii) the most glaring omissions include (a) heuristics (the only heuristic briefly discussed here is the non-linear conjugate gradient in Section 2.4), (b) methods for distributed systems, and (c) adaptivity to unknown parameters. Regarding (a) we refer to Nocedal and Wright [2006] where the most practical algorithms are discussed in great details (e.g., quasi-newton methods such as BFGS and L-BFGS, primal-dual interior point methods, etc.). The recent survey Boyd et al. [2011] discusses the alternating direction method of multipliers (ADMM) which is a popular method to address (b). Finally (c) is a subtle and important issue. In the entire monograph the emphasis is on presenting the algorithms and proofs in the simplest way, and thus for sake of convenience we assume that the relevant parameters describing the regularity and curvature of the objective function (Lipschitz constant, smoothness constant, strong convexity parameter) are known and can be used to tune the algorithm's own parameters. Line search is a powerful technique to replace the knowledge of these parameters and it is heavily used in practice, see again Nocedal and Wright [2006]. We observe however that from a theoretical point of view (c) is only a matter of logarithmic factors as one can always run in parallel several copies of the algorithm with different guesses for the values of the parameters[1]. Overall the attitude of this text with respect to (ii) is best summarized by a quote of Thomas Cover: "theory is the first term in the Taylor series of practice", Cover [1992].

Notation. We always denote by x^* a point in $\mathcal{X}$ such that $f(x^*) = \min_{x \in \mathcal{X}} f(x)$ (note that the optimization problem under consideration will always be clear from the context). In particular we always assume that x^* exists. For a vector $x \in \mathbb{R}^n$ we denote by $x(i)$ its i^{th} coordinate. The dual of a norm $\|\cdot\|$ (defined later) will be denoted either $\|\cdot\|_*$ or $\|\cdot\|^*$ (depending on whether the norm already comes with a subscript). Other notation are standard (e.g., I_n for the $n \times n$ identity matrix, $\succeq$ for the positive semi-definite order on matrices, etc).

[1]Note that this trick does not work in the context of Chapter 6.

f	Algorithm	Rate	# Iter	Cost/iter
non-smooth	center of gravity	$\exp\left(-\frac{t}{n}\right)$	$n\log\left(\frac{1}{\varepsilon}\right)$	1 ∇, 1 n-dim $\int$
non-smooth	ellipsoid method	$\frac{R}{r}\exp\left(-\frac{t}{n^2}\right)$	$n^2\log\left(\frac{R}{r\varepsilon}\right)$	1 ∇, mat-vec $\times$
non-smooth	Vaidya	$\frac{Rn}{r}\exp\left(-\frac{t}{n}\right)$	$n\log\left(\frac{Rn}{r\varepsilon}\right)$	1 ∇, mat-mat $\times$
quadratic	CG	exact $\exp\left(-\frac{t}{\kappa}\right)$	n $\kappa\log\left(\frac{1}{\varepsilon}\right)$	1 ∇
non-smooth, Lipschitz	PGD	$RL/\sqrt{t}$	R^2L^2/ε^2	1 ∇, 1 proj.
smooth	PGD	$\beta R^2/t$	$\beta R^2/\varepsilon$	1 ∇, 1 proj.
smooth	AGD	$\beta R^2/t^2$	$R\sqrt{\beta/\varepsilon}$	1 ∇
smooth (any norm)	FW	$\beta R^2/t$	$\beta R^2/\varepsilon$	1 ∇, 1 LP
strong. conv., Lipschitz	PGD	$L^2/(\alpha t)$	$L^2/(\alpha\varepsilon)$	1 ∇ , 1 proj.
strong. conv., smooth	PGD	$R^2\exp\left(-\frac{t}{\kappa}\right)$	$\kappa\log\left(\frac{R^2}{\varepsilon}\right)$	1 ∇ , 1 proj.
strong. conv., smooth	AGD	$R^2\exp\left(-\frac{t}{\sqrt{\kappa}}\right)$	$\sqrt{\kappa}\log\left(\frac{R^2}{\varepsilon}\right)$	1 ∇
$f+g$, f smooth, g simple	FISTA	$\beta R^2/t^2$	$R\sqrt{\beta/\varepsilon}$	1 ∇ of f Prox of g
$\max_{y\in\mathcal{Y}}\varphi(x,y)$, φ smooth	SP-MP	$\beta R^2/t$	$\beta R^2/\varepsilon$	MD on $\mathcal{X}$ MD on $\mathcal{Y}$
linear, $\mathcal{X}$ with F ν-self-conc.	IPM	$\nu\exp\left(-\frac{t}{\sqrt{\nu}}\right)$	$\sqrt{\nu}\log\left(\frac{\nu}{\varepsilon}\right)$	Newton step on F
non-smooth	SGD	$BL/\sqrt{t}$	B^2L^2/ε^2	1 stoch. ∇, 1 proj.
non-smooth, strong. conv.	SGD	$B^2/(\alpha t)$	$B^2/(\alpha\varepsilon)$	1 stoch. ∇, 1 proj.
$f=\frac{1}{m}\sum f_i$ f_i smooth strong. conv.	SVRG	–	$(m+\kappa)\log\left(\frac{1}{\varepsilon}\right)$	1 stoch. ∇

Table 1.1: Summary of the results proved in Chapter 2 to Chapter 5 and some of the results in Chapter 6.

References

A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. *Arxiv preprint arXiv:1410.0723*, 2014.

Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *Arxiv preprint arXiv:1407.1537*, 2014.

K. M. Anstreicher. Towards a practical volumetric cutting plane method for convex programming. *SIAM Journal on Optimization*, 9(1):190–206, 1998.

J.Y Audibert, S. Bubeck, and R. Munos. Bandit view on noisy optimization. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*. MIT press, 2011.

J.Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39:31–45, 2014.

F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

B. Barak. Sum of squares upper bounds, lower bounds, and open questions. Lecture Notes, 2014.

A. Beck and M. Teboulle. Mirror Descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. Society for Industrial and Applied Mathematics (SIAM), 2001.

D. Bertsimas and S. Vempala. Solving convex programs by random walks. *Journal of the ACM*, 51:540–556, 2004.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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.

S. Bubeck. Introduction to online optimization. Lecture Notes, 2011.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. *Arxiv preprint arXiv:1412.1587*, 2014.

S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Arxiv preprint arXiv:1507.02564*, 2015a.

S. Bubeck, Y.-T. Lee, and M. Singh. A geometric alternative to nesterov's accelerated gradient descent. *Arxiv preprint arXiv:1506.08187*, 2015b.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

A. Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

K. Clarkson, E. Hazan, and D. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM*, 2012.

A. Conn, K. Scheinberg, and L. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics (SIAM), 2009.

T. M. Cover. 1990 shannon lecture. *IEEE information theory society newsletter*, 42(4), 1992.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

J. C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62 (2):432–444, 1978.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

M. P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4):1326–1350, 2007.

M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18: 53–58, 1995.

B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific J. Math*, 10(4):1257–1261, 1960.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

E. Hazan. The convex optimization approach to regret minimization. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2011.

M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 427–435, 2013.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, 2013.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

L. K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, pages 608–613, 1992.

A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 121–147. MIT press, 2011a.

A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem's structure. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 149–183. MIT press, 2011b.

N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Y.-T. Lee and A. Sidford. Path finding i :solving linear programs with ÃŢ(sqrt(rank)) linear system solves. *Arxiv preprint arXiv:1312.6677*, 2013.

Y.-T. Lee, A. Sidford, and S. C.-W Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. *abs/1508.04874*, 2015.

A. Levin. On an algorithm for the minimization of convex functions. In *Soviet Mathematics Doklady*, volume 160, pages 1244–1247, 1965.

L. Lovász. Hit-and-run mixes fast. *Math. Prog.*, 86:443–461, 1998.

G. Lugosi. Comment on: ℓ_1-penalization for mixture regression models. *Test*, 19(2):259–263, 2010.

N. Maculan and G. G. de Paula. A linear-time median-finding algorithm for projecting a vector on the simplex of rn. *Operations research letters*, 8(4): 219–222, 1989.

A. Nemirovski. Orth-method for smooth convex optimization. *Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika*, 2, 1982.

A. Nemirovski. Information-based complexity of convex programming. *Lecture Notes*, 1995.

A. 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, 2004a.

A. Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes*, 2004b.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

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.

Y. Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. CORE Discussion Papers 1997019, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 1997.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004a.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2004b.

Y. Nesterov. Gradient methods for minimizing composite objective function. Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22:341–362, 2012.

Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. Society for Industrial and Applied Mathematics (SIAM), 1994.

D. Newman. Location of the maximum on unimodal surfaces. *Journal of the ACM*, 12(3):395–398, 1965.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):123–231, 2013.

A. Rakhlin. Lecture notes on online learning. 2009.

J. Renegar. *A mathematical view of interior-point methods in convex optimization*, volume 3. Siam, 2001.

P. Richtárik and M. Takác. Parallel coordinate descent methods for big data optimization. *Arxiv preprint arXiv:1212.0873*, 2012.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164:60–72, 1999.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, 2013a.

S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems (NIPS)*, 2013b.

W. Su, S. Boyd, and E. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008.

A. Tsybakov. Optimal rates of aggregation. In *Conference on Learning Theory (COLT)*, pages 303–313. 2003.

P. M. Vaidya. A new algorithm for minimizing convex functions over convex sets. In *Foundations of Computer Science, 1989., 30th Annual Symposium on*, pages 338–343, 1989.

P. M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical programming*, 73(3):291–341, 1996.

S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57 (7):2479–2493, 2009.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Arxiv preprint arXiv:1409.3257*, 2014.