Full text available at: http://dx.doi.org/10.1561/2400000036

Acceleration Methods

Full text available at: http://dx.doi.org/10.1561/2400000036

Other titles in Foundations and Trends® in Optimization

Algorithms for Verifying Deep Neural Networks
Changliu Liu, Tomer Arnon, Christopher Lazarus, Christopher Strong, Clark Barrett and Mykel J. Kochenderfer
ISBN: 978-1-68083-786-5

Distributionally Robust Learning
Ruidi Chen and Ioannis Ch. Paschalidis
ISBN: 978-1-68083-772-8

Atomic Decomposition via Polar Alignment: The Geometry of Structured Optimization
Zhenan Fan, Halyun Jeong, Yifan Sun and Michael P. Friedlander
ISBN: 978-1-68083-742-1

Optimization Methods for Financial Index Tracking: From Theory to Practice
Konstantinos Benidis, Yiyong Feng and Daniel P. Palomar
ISBN: 978-1-68083-464-2

The Many Faces of Degeneracy in Conic Optimization
Dmitriy Drusvyatskiy and Henry Wolkowicz
ISBN: 978-1-68083-390-4

Full text available at: http://dx.doi.org/10.1561/2400000036

Acceleration Methods

Alexandre d'Aspremont
CNRS & Ecole Normale Supérieure, Paris
France
aspremon@ens.fr

Damien Scieur
Samsung SAIT AI Lab & Mila, Montreal
Canada
damien.scieur@gmail.com

Adrien Taylor
INRIA & Ecole Normale Supérieure, Paris
France
adrien.taylor@inria.fr

now
the essence of knowledge
Boston — Delft

Full text available at: http://dx.doi.org/10.1561/2400000036

Foundations and Trends® in Optimization

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

A. d'Aspremont et al.. *Acceleration Methods*. Foundations and Trends® in Optimization, vol. 5, no. 1-2, pp. 1–245, 2021.

ISBN: 978-1-68083-929-6
© 2021 A. d'Aspremont et al.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1 781 871 0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Full text available at: http://dx.doi.org/10.1561/2400000036

Foundations and Trends® in Optimization
Volume 5, Issue 1-2, 2021

Editorial Board

Editors-in-Chief

Garud Iyengar
Columbia University, USA

Editors

Dimitris Bertsimas
Massachusetts Institute of Technology

John R. Birge
The University of Chicago

Robert E. Bixby
Rice University

Emmanuel Candes
Stanford University

David Donoho
Stanford University

Laurent El Ghaoui
University of California, Berkeley

Donald Goldfarb
Columbia University

Michael I. Jordan
University of California, Berkeley

Zhi-Quan (Tom) Luo
University of Minnesota, Twin Cites

George L. Nemhauser
Georgia Institute of Technology

Arkadi Nemirovski
Georgia Institute of Technology

Yurii Nesterov
HSE University

Jorge Nocedal
Northwestern University

Pablo A. Parrilo
Massachusetts Institute of Technology

Boris T. Polyak
Institute for Control Science, Moscow

Tamás Terlaky
Lehigh University

Michael J. Todd
Cornell University

Kim-Chuan Toh
National University of Singapore

John N. Tsitsiklis
Massachusetts Institute of Technology

Lieven Vandenberghe
University of California, Los Angeles

Robert J. Vanderbei
Princeton University

Stephen J. Wright
University of Wisconsin

Full text available at: http://dx.doi.org/10.1561/2400000036

Editorial Scope

Topics

Foundations and Trends® in Optimization publishes survey and tutorial articles in the following topics:

- algorithm design, analysis, and implementation (especially, on modern computing platforms
- models and modeling systems, new optimization formulations for practical problems
- applications of optimization in machine learning, statistics, and data analysis, signal and image processing, computational economics and finance, engineering design, scheduling and resource allocation, and other areas

Information for Librarians

Foundations and Trends® in Optimization, 2021, Volume 5, 4 issues. ISSN paper version 2167-3888. ISSN online version 2167-3918. Also available as a combined paper and online subscription.

Full text available at: http://dx.doi.org/10.1561/2400000036

Contents

Full text available at: http://dx.doi.org/10.1561/2400000036

Full text available at: http://dx.doi.org/10.1561/2400000036

Full text available at: http://dx.doi.org/10.1561/2400000036

Acceleration Methods

Alexandre d'Aspremont[1], Damien Scieur[2] and Adrien Taylor[3]

[1] *CNRS & Ecole Normale Supérieure, Paris, France; aspremon@ens.fr*
[2] *Samsung SAIT AI Lab & Mila, Montreal, Canada; damien.scieur@gmail.com*
[3] *INRIA & Ecole Normale Supérieure, Paris, France; adrien.taylor@inria.fr*

ABSTRACT

This monograph covers some recent advances in a range of acceleration techniques frequently used in convex optimization. We first use quadratic optimization problems to introduce two key families of methods, namely momentum and nested optimization schemes. They coincide in the quadratic case to form the *Chebyshev method*.

We discuss momentum methods in detail, starting with the seminal work of Nesterov [1] and structure convergence proofs using a few master templates, such as that for *optimized gradient methods*, which provide the key benefit of showing how momentum methods optimize convergence guarantees. We further cover proximal acceleration, at the heart of the *Catalyst* and *Accelerated Hybrid Proximal Extragradient* frameworks, using similar algorithmic patterns.

Common acceleration techniques rely directly on the knowledge of some of the regularity parameters in the problem at hand. We conclude by discussing *restart* schemes, a set of simple techniques for reaching nearly optimal convergence rates while adapting to unobserved regularity parameters.

Alexandre d'Aspremont, Damien Scieur and Adrien Taylor (2021), "Acceleration Methods", Foundations and Trends® in Optimization: Vol. 5, No. 1-2, pp 1–245. DOI: 10.1561/2400000036.
©2021 A. d'Aspremont et al.

Full text available at: http://dx.doi.org/10.1561/2400000036

1

Introduction

Optimization methods are a core component of the modern numerical toolkit. In many cases, iterative algorithms for solving convex optimization problems have reached a level of efficiency and reliability comparable to that of advanced linear algebra routines. This is largely true for medium scale-problems where interior point methods reign supreme, but less so for large-scale problems where the complexity of first-order methods is not as well understood and efficiency remains a concern.

The situation has improved markedly in recent years, driven in particular by the emergence of a number of applications in statistics, machine learning, and signal processing. Building on Nesterov's path-breaking algorithm from the 80's, several accelerated methods and numerical schemes have been developed that both improve the efficiency of optimization algorithms and refine their complexity bounds. Our objective in this monograph is to cover these recent developments using a few master templates.

The methods described in this manuscript can be arranged in roughly two categories. The first, stemming from the work of Nesterov [1], produces variants of the gradient method with accelerated worst-case

Full text available at: http://dx.doi.org/10.1561/2400000036

convergence rates that are provably optimal under classical regularity assumptions. The second uses outer iteration (a.k.a. nested) schemes to speed up convergence. In this second setting, accelerated schemes run both an inner loop and an outer loop, with the inner iterations being solved by classical optimization methods, and the outer loop containing the acceleration mechanism.

Direct acceleration techniques. Ever since the original algorithm by Nesterov [1], the acceleration phenomenon was regarded as somewhat of a mystery. While accelerated gradient methods can be seen as iteratively building a model for the function and using it to guide gradient computations, the argument is essentially algebraic and is simply an effective exploitation of regularity assumptions. This approach of collecting inequalities induced by regularity assumptions and cleverly chaining them to prove convergence was also used in e.g., [2], to produce an optimal proximal gradient method. There too, however, the proof yielded little evidence as to why the method is actually faster.

Fortunately, we are now better equipped to push the proof mechanisms much further. Recent advances in the programmatic design of optimization algorithms allow us to design and analyze algorithms by following a more principled approach. In particular, the *performance estimation approach*, pioneered by Drori and Teboulle [3], can be used to design optimal methods from scratch, selecting algorithmic parameters to optimize worst-case performance guarantees [3], [4]. Primal dual optimality conditions on the design problem then provide a blueprint for the accelerated algorithm structure and for its convergence proof.

Using this framework, acceleration is no longer a mystery: it is the main objective in the design of the algorithm. We recover the usual "soup of regularity inequalities" that forms the template of classical convergence proofs, but the optimality conditions of the design problem explicitly produce a method that optimizes the convergence guarantee. In this monograph, we cover accelerated first-order methods using this systematic template and describe a number of convergence proofs for classical variants of the accelerated gradient method, such as those of Nesterov (1983, 2003), Beck and Teboulle [2] and Tseng [6] as well as more recent ones [4].

Full text available at: http://dx.doi.org/10.1561/2400000036

Nested acceleration schemes. The second category of acceleration techniques that we cover in this monograph is composed of outer iteration schemes, in which classical optimization algorithms are used as a black-box in the inner loop and acceleration is produced by an argument in the outer loop. We describe three acceleration results of this type.

The first scheme is based on nonlinear acceleration techniques. Based on arguments dating back to [7]–[9], these techniques use a weighted average of iterates to extrapolate a better candidate solution than the last iterate. We begin by describing the Chebyshev method for solving quadratic problems, which interestingly qualifies both as a gradient method and as an outer iteration scheme. It takes its name from the use of Chebyshev polynomial coefficients to approximately minimize the gradient at the extrapolated solution. The argument can be extended to non-quadratic optimization problems provided the extrapolation procedure is regularized.

The second scheme, due to [10]–[12] relies on a conceptual accelerated proximal point algorithm, and uses classical iterative methods to approximate the proximal point in an inner loop. In particular, this framework produces accelerated gradient methods (in the same sense as Nesterov's acceleration) when the approximate proximal points are computed using linearly converging gradient-based optimization methods, taking advantage of the fact that the inner problems are always strongly convex.

Finally, we describe restart schemes. These techniques exploit regularity properties called Hölderian error bounds, which extend strong convexity properties near the optimum and hold almost generically, to improve the convergence rates of most first-order methods. The parameters of the Hölderian error bounds are usually unknown, but the restart schemes are robust: that is, they are adaptive to the Hölderian parameters and their empirical performance is excellent on problems with reasonable precision targets.

Content and organization. We present a few convergence acceleration techniques that are particularly relevant in the context of (first-order) convex optimization. Our summary includes our own points of view on

Full text available at: http://dx.doi.org/10.1561/2400000036

the topic and is focused on techniques that have received substantial attention since the early 2000's, although some of the underlying ideas are much older. We do not pretend to be exhaustive, and we are aware that valuable references might not appear below.

The sections can be read nearly independently. However, we believe the insights of some sections can benefit the understanding of others. In particular, Chebyshev acceleration (Section 2) and nonlinear acceleration (Section 3) are clearly complementary readings. Similarly, Chebyshev acceleration (Section 2) and Nesterov acceleration (Section 4), Nesterov acceleration (Section 4) and proximal acceleration (Section 5), as well as Nesterov acceleration (Section 4) and restart schemes (Section 6) certainly belong together.

Prerequisites and complementary readings. This monograph is not meant to be a general-purpose manuscript on convex optimization, for which we refer the reader to the now classical references [13]–[15]. Other directly related references are provided in the text.

We assume the reader to have a working knowledge of base linear algebra and convex analysis (such as of subdifferentials), as we do not detail the corresponding technical details while building on them. Classical references on the latter include [16]–[18].

Full text available at: http://dx.doi.org/10.1561/2400000036

threshold. We have, the following proposition directly linking the null space property and the Hölderian error bound (HEB).

Proposition 6.1. Given a coding matrix $A \in \mathbb{R}^{n \times p}$ satisfying (NSP) at order s with constant $\alpha \geq 1$, if the original signal $x_\star$ is s-sparse, then for any $x \in \mathbb{R}^p$ satisfying $Ax = b$, $x \neq x_\star$, we have

$$\|x\|_1 - \|x_\star\|_1 > \frac{\alpha - 1}{\alpha + 1} \|x - x_\star\|_1. \tag{6.29}$$

This implies signal recovery, i.e. optimality of $x_\star$ for (ℓ_1 recovery) and the Hölderian error bound (HEB) with $\mu = \frac{\alpha-1}{\alpha+1}$.

6.9 Notes and References

The optimal complexity bounds and exponential restart schemes detailed here can be traced back to [242]. Restart schemes were extensively benchmarked in the numerical toolbox TFOCS by [114], with a particular focus on compressed sensing applications. The robustness result showing that a log scale grid search produces near optimal complexity bounds is due to [116].

Restart schemes based on the gradient norm as a termination criterion also reach nearly optimal complexity bounds and adapt to strong convexity [80] or HEB parameters [252].

Hölderian error bounds for analytic functions can be traced back to the work of Lojasiewicz [253]. They were extended to much broader classes of functions by [241], [254]. Several examples of problems in signal processing where this condition holds can be found in, e.g., [248], [255]. Calculus rules for the exponent are discussed in details in, e.g., [244].

Restarting is also helpful in the stochastic setting, with [256] showing recently that stochastic algorithms with geometric step decay converge linearly on functions satisfying Hölderian error bounds. This validates a classical empirical acceleration trick, which is to restarts every few epochs after adjusting the step size (aka the learning rate in machine learning terminology).

Full text available at: http://dx.doi.org/10.1561/2400000036

Appendices

Full text available at: http://dx.doi.org/10.1561/2400000036

A

Useful Inequalities

In this appendix, we prove basic inequalities involving smooth strongly convex functions. Most of these inequalities are not used in our developments. Nevertheless, we believe they are useful for gaining intuition about smooth strongly convex of functions, as well as for comparisons with the literature.

Also note that these inequalities can be considered standard (see, e.g., [5, Theorem 2.1.5].

A.1 Smoothness and Strong Convexity in Euclidean spaces

In this section, we consider a Euclidean setting, where $\|x\|_2^2 = \langle x; x\rangle$ and $\langle .; .\rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a dot product.

The following theorem summarizes known inequalities that characterize the class of smooth convex functions. Note that these characterizations of $f \in \mathcal{F}_{0,L}$ are all equivalent assuming that $f \in \mathcal{F}_{0,\infty}$ since convexity is not implied by some of the points below. In particular, (i), (ii), (v), (vi), and (vii) do not encode the convexity of f when taken on their own, whereas (iii) and (iv) encode both smoothness and convexity.

Full text available at: http://dx.doi.org/10.1561/2400000036

Theorem A.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function. The following statements are equivalent for inclusion in $\mathcal{F}_{0,L}$.

(i) ∇f satisfies a Lipschitz condition: for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

(ii) f is upper bounded by quadratic functions: for all $x, y \in \mathbb{R}^d$,

$$f(x) \leq f(y) + \langle \nabla f(y); x - y\rangle + \frac{L}{2}\|x - y\|_2^2.$$

(iii) f satisfies, for all $x, y \in \mathbb{R}^d$,

$$f(x) \geq f(y) + \langle \nabla f(y); x - y\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

(iv) ∇f is cocoercive: for all $x, y \in \mathbb{R}^d$,

$$\langle \nabla f(x) - \nabla f(y); x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

(v) ∇f satisfies, for all $x, y \in \mathbb{R}^d$,

$$\langle \nabla f(x) - \nabla f(y); x - y\rangle \leq L\|x - y\|_2^2.$$

(vi) $\frac{L}{2}\|x\|_2^2 - f(x)$ is convex.

(vii) f satisfies, for all $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) - \lambda(1 - \lambda)\frac{L}{2}\|x - y\|_2^2.$$

Proof. We start with (i)⇒(ii). We use the first-order expansion

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x + \tau(y - x)); y - x\rangle d\tau.$$

The quadratic upper bound then follows from algebraic manipulations

Full text available at: http://dx.doi.org/10.1561/2400000036

and from upper bounding the integral term:

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x); y - x \rangle \\ &\quad + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x); y - x \rangle d\tau \\ &\leq f(x) + \langle \nabla f(x); y - x \rangle \\ &\quad + \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_2 \|y - x\|_2 d\tau \\ &\leq f(x) + \langle \nabla f(x); y - x \rangle + L\|x - y\|_2^2 \int_0^1 \tau d\tau \\ &= f(x) + \langle \nabla f(x); y - x \rangle + \frac{L}{2}\|x - y\|_2^2. \end{aligned}$$

We proceed with (ii)⇒(iii). The idea is to require the quadratic upper bound to be everywhere above the linear lower bound arising from the convexity of f. That is, for all $x, y, z \in \mathbb{R}^d$,

$$f(y) + \langle \nabla f(y); z - y \rangle \leq f(z) \leq f(x) + \langle \nabla f(x); z - x \rangle + \frac{L}{2}\|x - z\|_2^2.$$

In other words, for all $z \in \mathbb{R}^d$, we must have

$$\begin{aligned} & f(y) + \langle \nabla f(y); z - y \rangle \leq f(x) + \langle \nabla f(x); z - x \rangle + \frac{L}{2}\|x - z\|_2^2 \\ \Leftrightarrow\ & f(y) - f(x) + \langle \nabla f(y); z - y \rangle - \langle \nabla f(x); z - x \rangle - \frac{L}{2}\|x - z\|_2^2 \leq 0 \\ \Leftrightarrow\ & f(y) - f(x) + \max_{z \in \mathbb{R}^d} \langle \nabla f(y); z - y \rangle - \langle \nabla f(x); z - x \rangle - \frac{L}{2}\|x - z\|_2^2 \leq 0 \\ \Leftrightarrow\ & f(y) - f(x) + \langle \nabla f(y); x - y \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 0, \end{aligned}$$

where the last line follows from the explicit maximization on z. That is, we pick $z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y))$ and reach the desired result after base algebraic manipulations.

We continue with (iii)⇒(iv), which simply follows from adding

$$\begin{aligned} f(x) &\geq f(y) + \langle \nabla f(y); x - y \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\ f(y) &\geq f(x) + \langle \nabla f(x); y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2. \end{aligned}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

To obtain (iv)⇒(i), one can use Cauchy-Schwartz:

$$\begin{aligned}\frac{1}{L}\|\nabla f(x)-\nabla f(y)\|_2^2 \leq & \langle \nabla f(x)-\nabla f(y); x-y\rangle \\ \leq & \|\nabla f(x)-\nabla f(y)\|_2\|x-y\|_2,\end{aligned}$$

which allows us to conclude that $\|\nabla f(x)-\nabla f(y)\|_2 \leq L\|x-y\|_2$, thus reaching the final statement.

To obtain (ii)⇒(v), we simply add

$$\begin{aligned}f(x) &\leq f(y)+\langle \nabla f(y); x-y\rangle + \frac{L}{2}\|x-y\|_2^2 \\ f(y) &\leq f(x)+\langle \nabla f(x); y-x\rangle + \frac{L}{2}\|x-y\|_2^2\end{aligned}$$

and reorganize the resulting inequality.

To obtain (v)⇒(ii), we again use a first-order expansion:

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x+\tau(y-x)); y-x\rangle d\tau.$$

The quadratic upper bound then follows from algebraic manipulations and from upper bounding the integral term. (We use the intermediate variable $z_\tau = x+\tau(y-x)$ for convenience)

$$\begin{aligned}f(y) &= f(x)+\langle \nabla f(x); y-x\rangle \\ &\quad + \int_0^1 \langle \nabla f(x+\tau(y-x))-\nabla f(x); y-x\rangle d\tau \\ &= f(x)+\langle \nabla f(x); y-x\rangle + \int_0^1 \frac{1}{\tau}\langle \nabla f(z_\tau)-\nabla f(x); z_\tau - x\rangle d\tau \\ &\leq f(x)+\langle \nabla f(x); y-x\rangle + \int_0^1 \frac{L}{\tau}\|z_\tau - x\|_2^2 d\tau \\ &= f(x)+\langle \nabla f(x); y-x\rangle + L\|x-y\|_2^2 \int_0^1 \tau d\tau \\ &= f(x)+\langle \nabla f(x); y-x\rangle + \frac{L}{2}\|x-y\|_2^2.\end{aligned}$$

For the equivalence (vi)⇔(ii), simply define $h(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ (and hence $\nabla h(x) = Lx - \nabla f(x)$) and observe that for all $x, y \in \mathbb{R}^d$,

$$h(x) \geq h(y)+\langle \nabla h(y); x-y\rangle \;\Leftrightarrow\; f(x) \leq f(y)+\langle \nabla f(y); x-y\rangle + \frac{L}{2}\|x-y\|_2^2,$$

Full text available at: http://dx.doi.org/10.1561/2400000036

which follows from base algebraic manipulations.

Finally, the equivalence (vi)⇔(vii) follows the same $h(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ (and hence $\nabla h(x) = Lx - \nabla f(x)$) and the observation that for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$, we have

$$\begin{aligned} &h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) \\ &\qquad\qquad \Leftrightarrow \\ &f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\frac{L}{2}\|x-y\|_2^2, \end{aligned}$$

which follows from base algebraic manipulations. ■

To obtain the corresponding inequalities in the strongly convex case, one can rely on Fenchel conjugation between smoothness and strong convexity; see, for example, [17, Proposition 12.6]. The following inequalities are stated without proofs; they can be obtained either as direct consequences of the definitions or from Fenchel conjugation along with the statements of Theorem A.1.

Theorem A.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a closed convex proper function. The following statements are equivalent for inclusion in $\mathcal{F}_{\mu,L}$.

(i) ∇f satisfies a Lipschitz and an inverse Lipschitz condition: for all $x, y \in \mathbb{R}^d$,

$$\mu\|x-y\|_2 \leq \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2.$$

(ii) f is lower and upper bounded by quadratic functions: for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned} f(y) + \langle \nabla f(y); x-y\rangle + \frac{\mu}{2}\|x-y\|_2^2 \\ \leq f(x) \leq f(y) + \langle \nabla f(y); x-y\rangle + \frac{L}{2}\|x-y\|_2^2. \end{aligned}$$

(iii) f satisfies, for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned} f(y) + \langle \nabla f(y); x-y\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\ \leq f(x) \leq \\ f(y) + \langle \nabla f(y); x-y\rangle + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|_2^2. \end{aligned}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

(iv) ∇f satisfies, for all $x, y \in \mathbb{R}^d$,

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y); x - y\rangle \leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

(v) ∇f satisfies, for all $x, y \in \mathbb{R}^d$,

$$\mu\|x - y\|_2^2 \leq \langle \nabla f(x) - \nabla f(y); x - y\rangle \leq L\|x - y\|_2^2.$$

(vi) For all $\lambda \in [0, 1]$,

$$\lambda f(x)+(1-\lambda)f(y) - \lambda(1-\lambda)\frac{L}{2}\|x - y\|_2^2 \leq f(\lambda x + (1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x - y\|_2^2.$$

(vii) $f(x) - \frac{\mu}{2}\|x\|_2^2$ and $\frac{L}{2}\|x\|_2^2 - f(x)$ are convex and $(L - \mu)$-smooth.

Finally, we mention that the existence of an inequality that allows us to encode both smoothness and strong convexity together. This inequality is also known as an *interpolation* inequality [210], and it turns out to be particularly useful for proving worst-case guarantees.

Theorem A.3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. f is L-smooth μ-strongly convex if and only if

$$f(x) \geq f(y)+\langle \nabla f(y); x - y\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\mu}{2(1-\mu/L)}\|x - y - \frac{1}{L}(\nabla f(x) - \nabla f(y))\|_2^2. \tag{A.1}$$

Proof. ($f \in \mathcal{F}_{\mu,L} \Rightarrow$ (A.1)) The idea is to require the quadratic upper bound from smoothness to be everywhere above the quadratic lower bound arising from strong convexity. That is, for all $x, y, z \in \mathbb{R}^d$

$$f(y) + \langle \nabla f(y); z - y\rangle + \frac{\mu}{2}\|z - y\|_2^2 \leq f(z) \leq f(x) + \langle \nabla f(x); z - x\rangle + \frac{L}{2}\|x - z\|_2^2.$$

Full text available at: http://dx.doi.org/10.1561/2400000036

In other words, for all $z \in \mathbb{R}^d$, we must have

$$\begin{aligned} f(y)+\langle \nabla f(y); z-y\rangle + \frac{\mu}{2}\|z-y\|_2^2 \leq f(x) \\ + \langle \nabla f(x); z-x\rangle + \frac{L}{2}\|x-z\|_2^2 \\ \Leftrightarrow f(y) - f(x) + \langle \nabla f(y); z-y\rangle + \frac{\mu}{2}\|z-y\|_2^2 - \langle \nabla f(x); z-x\rangle \\ - \frac{L}{2}\|x-z\|_2^2 \leq 0 \\ \Leftrightarrow f(y) - f(x) + \max_{z\in\mathbb{R}^d} \left(\langle \nabla f(y); z-y\rangle + \frac{\mu}{2}\|z-y\|_2^2 \right. \\ \left. - \langle \nabla f(x); z-x\rangle - \frac{L}{2}\|x-z\|_2^2 \right) \leq 0 \end{aligned}$$

explicit maximization over z. That is, picking $z = \frac{Lx-\mu y}{L-\mu} - \frac{1}{L-\mu}(\nabla f(x) - \nabla f(y))$ allows the desired inequality to be reached by base algebraic manipulations.

((A.1)$\Rightarrow f \in \mathcal{F}_{\mu,L}$) $f \in \mathcal{F}_{0,L}$ is direct by observing that (A.1) is stronger than Theorem A.1(iii); $f \in \mathcal{F}_{\mu,L}$ is then direct by reformulating (A.1) as

$$\begin{aligned} f(x) \geq f(y)+\langle \nabla f(y); x-y\rangle + \frac{\mu}{2}\|x-y\|_2^2 \\ + \frac{1}{2L(1-\mu/L)}\|\nabla f(x) - \nabla f(y) - \mu(x-y)\|_2^2, \end{aligned}$$

which is stronger than $f(x) \geq f(y) + \langle \nabla f(y); x-y\rangle + \frac{\mu}{2}\|x-y\|_2^2$. ■

Remark A.1. It is crucial to recall that some of the inequalities above are only valid when $\mathbf{dom}\, f = \mathbb{R}^d$—in particular, this holds for Theorem A.1(iii & iv), Theorem A.2(iii&iv), and Theorem A.3. We refer to [90] for an illustration that some inequalities are not valid when restricted on some $\mathbf{dom}\, f \neq \mathbb{R}^d$. Most standard inequalities, however, do hold even in the case of restricted domains, as established in, e.g., [5]. Some other inequalities, such as Theorem A.1(iv) and Theorem A.2(iv), do hold under the additional assumption of twice continuous differentiability(see, for example, [257]).

Full text available at: http://dx.doi.org/10.1561/2400000036

A.2 Smoothness for General Norms and Restricted Sets

In this section, we show that requiring a Lipschitz condition on ∇f, on a convex set $C \subseteq \mathbb{R}^d$, implies a quadratic upper bound on f. That is, requiring that for all $x, y \in C$,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|,$$

where $\|.\|$ is some norm and $\|.\|_*$ is the corresponding dual norm, implies a quadratic upper bound $\forall x, y \in C$:

$$f(x) \leq f(y) + \langle \nabla f(y); x - y\rangle + \frac{L}{2}\|x - y\|^2.$$

Theorem A.4. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be continuously differentiable on some open convex set $C \subseteq \mathbb{R}^d$, and let it satisfy a Lipschitz condition

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|,$$

for all $x, y \in C$. Then, it holds that

$$f(x) \leq f(y) + \langle \nabla f(y); x - y\rangle + \frac{L}{2}\|x - y\|^2,$$

for all $x, y \in C$.

Proof. The desired result is obtained from a first-order expansion:

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x + \tau(y - x)); y - x\rangle d\tau.$$

The quadratic upper bound then follows from algebraic manipulations and from upper bounding the integral term

$$\begin{aligned}
f(y) &= f(x) + \langle \nabla f(x); y - x\rangle \\
&\quad + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x); y - x\rangle d\tau \\
&\leq f(x) + \langle \nabla f(x); y - x\rangle \\
&\quad + \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_* \|y - x\| d\tau \\
&\leq f(x) + \langle \nabla f(x); y - x\rangle + L\|x - y\|^2 \int_0^1 \tau d\tau \\
&= f(x) + \langle \nabla f(x); y - x\rangle + \frac{L}{2}\|x - y\|^2.
\end{aligned}$$

■

Full text available at: http://dx.doi.org/10.1561/2400000036

B

Variations on Nesterov Acceleration

B.1 Relations between Acceleration Methods

B.1.1 Optimized Gradient Method: Forms I & II

In this short section, we show that Algorithm 9 and Algorithm 10 generate the same sequence $\{y_k\}_k$. A direct consequence of this statement is that the sequences $\{x_k\}_k$ also match, as in both cases they are generated from simple gradient steps on $\{y_k\}_k$.

For this purpose we show that Algorithm 10 is a reformulation of Algorithm 9.

Proposition B.1. The sequence $\{y_k\}_k$ generated by Algorithm 9 is equal to that generated by Algorithm 10.

Proof. We first observe that the sequences are initiated the same way in both formulations of the OGM. Furthermore, consider one iteration of the OGM in form I:

$$y_k = \left(1 - \frac{1}{\theta_{k,N}}\right) x_k + \frac{1}{\theta_{k,N}} z_k.$$

Therefore, we clearly have $z_k = \theta_{k,N} y_k + (1 - \theta_{k,N}) x_k$. At the next

Full text available at: http://dx.doi.org/10.1561/2400000036

iteration, we have

$$\begin{aligned} y_{k+1} &= \left(1 - \frac{1}{\theta_{k+1,N}}\right) x_{k+1} + \frac{1}{\theta_{k+1,N}} \left(z_k - \frac{2\theta_{k,N}}{L} \nabla f(y_k)\right) \\ &= \left(1 - \frac{1}{\theta_{k+1,N}}\right) x_{k+1} \\ &\quad + \frac{1}{\theta_{k+1,N}} \left(\theta_{k,N} y_k + (1 - \theta_{k,N}) x_k - \frac{2\theta_{k,N}}{L} \nabla f(y_k)\right), \end{aligned}$$

where we substituted z_k by its equivalent expression from the previous iteration. Now, by noting that $-\frac{1}{L}\nabla f(y_k) = x_{k+1} - y_k$, we reach

$$\begin{aligned} y_{k+1} &= \frac{\theta_{k+1,N} - 1}{\theta_{k+1,N}} x_{k+1} + \frac{1}{\theta_{k+1,N}} \left((1 - \theta_{k,N}) x_k + 2\theta_{k,N} x_{k+1} - \theta_{k,N} y_k\right) \\ &= x_{k+1} + \frac{\theta_{k,N} - 1}{\theta_{k+1,N}} (x_{k+1} - x_k) + \frac{\theta_{k,N}}{\theta_{k+1,N}} (x_{k+1} - y_k), \end{aligned}$$

where we reorganized the terms to achieve the same format as in Algorithm 10. ■

B.1.2 Nesterov's Method: Forms I, II, and III

Proposition B.2. The two sequences $\{x_k\}_k$ and $\{y_k\}_k$ generated by Algorithm 11 are equal to those generated by Algorithm 12.

Proof. In order to prove the result, we use the identities $A_{k+1} = a_k^2$ as well as $A_k = \sum_{i=0}^{k-1} a_i$, and $a_{k+1}^2 = a_k^2 + a_{k+1}$.

Given that the sequences $\{x_k\}_k$ are obtained from gradient steps on y_k in both formulations, it is sufficient to prove that the sequences $\{y_k\}_k$ match. The equivalence is clear for $k = 0$, as both methods generate $y_1 = x_0 - \frac{1}{L}\nabla f(x_0)$. For $k \geq 0$, from Algorithm 11, one can write iteration k as

$$y_k = \frac{A_k}{A_{k+1}} x_k + \left(1 - \frac{A_k}{A_{k+1}}\right) z_k,$$

and hence,

$$\begin{aligned} z_k &= \frac{A_{k+1}}{A_{k+1} - A_k} y_k + \left(1 - \frac{A_{k+1}}{A_{k+1} - A_k}\right) x_k \\ &= a_k y_k + (1 - a_k)\, x_k. \end{aligned}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

Substituting this expression in that for iteration $k+1$, we reach

$$\begin{aligned} y_{k+1} =& \frac{A_{k+1}}{A_{k+2}} x_{k+1} + \frac{A_{k+2} - A_{k+1}}{A_{k+2}} \left(z_k - \frac{A_{k+1} - A_k}{L} \nabla f(y_k) \right) \\ =& \frac{a_k^2}{a_{k+1}^2} x_{k+1} + \frac{1}{a_{k+1}} \left(a_k y_k + (1 - a_k) x_k - \frac{a_k}{L} \nabla f(y_k) \right) \\ =& \frac{a_k^2}{a_{k+1}^2} x_{k+1} + \frac{1}{a_{k+1}} \left(a_k x_{k+1} + (1 - a_k) x_k \right) \\ =& x_{k+1} + \frac{a_k - 1}{a_{k+1}} (x_{k+1} - x_k), \end{aligned}$$

where we substituted the expression for z_k and used previous identities to reach the desired statement. ■

The same relationship holds with Algorithm 13, as provided by the next proposition.

Proposition B.3. The three sequences $\{z_k\}_k$, $\{x_k\}_k$ and $\{y_k\}_k$ generated by Algorithm 11 are equal to those generated by Algorithm 13.

Proof. Clearly, we have $x_0 = z_0 = y_0$ in both methods. Let us assume that the sequences match up to iteration k, that is, up to y_{k-1}, x_k, and z_k. Clearly, both y_k and z_{k+1} are computed in the same way in both methods. It remains to compare the update rules for x_{k+1}: in Algorithm 13, we have

$$\begin{aligned} x_{k+1} &= \frac{A_k}{A_{k+1}} x_k + \left(1 - \frac{A_k}{A_{k+1}} \right) z_{k+1} \\ &= y_k - \left(1 - \frac{A_k}{A_{k+1}} \right) \frac{A_{k+1} - A_k}{L} \nabla f(y_k), \end{aligned}$$

where we used the update rule for z_{k+1}. Further simplifications, along with the identity $(A_{k+1} - A_k)^2 = A_{k+1}$ allows us to arrive at

$$\begin{aligned} x_{k+1} &= y_k - \frac{(A_{k+1} - A_k)^2}{L A_{k+1}} \nabla f(y_k) \\ &= y_k - \frac{1}{L} \nabla f(y_k), \end{aligned}$$

which is clearly the same update rule as that of Algorithm 11. Hence, all sequences match and the desired statement is proved. ■

Full text available at: http://dx.doi.org/10.1561/2400000036

B.1.3 Nesterov's Accelerated Gradient Method (Strongly Convex Case): Forms I, II, and III

In this short section, we provide alternate, equivalent, formulations for Algorithm 14.

Algorithm 28 Nesterov's method, form II

Input: L-smooth μ-strongly convex function f and initial point x_0.

1: **Initialize** $z_0 = x_0$; $q = \mu/L$, $A_0 = 0$, and $A_1 = (1-q)^{-1}$.
2: **for** $k = 0, \ldots$ **do**
3: $\quad A_{k+2} = \frac{2A_{k+1}+1+\sqrt{4A_{k+1}+4qA_{k+1}^2+1}}{2(1-q)}$
4: $\quad x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
5: $\quad y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$
6: $\quad$ with $\beta_k = \frac{(A_{k+2}-A_{k+1})(A_{k+1}(1-q)-A_k-1)}{A_{k+2}(2qA_{k+1}+1)-qA_{k+1}^2}$
7: **end for**

Output: Approximate solution x_N.

Proposition B.4. The two sequences $\{x_k\}_k$ and $\{y_k\}_k$ generated by Algorithm 14 are equal to those generated by Algorithm 28.

Proof. Without loss of generality, we can consider that a third sequence z_k is present in Algorithm 28 (although it is not computed).

Obviously, we have $x_0 = z_0 = y_0$ in both methods. Let us assume that the sequences match up to iteration k, that is, up to y_k, x_k, and z_k. Clearly, x_{k+1} is computed in the same way in both methods as a gradient step from y_k, and it remains to compare the update rules for y_{k+1}. In Algorithm 14, we have

$$\begin{aligned} y_{k+1} =& x_k + (\tau_k - \tau_{k+1}(\tau_k - 1)(1 - q\delta_k))\,(z_k - x_k) \\ & - \frac{(\delta_k - 1)\tau_{k+1} + 1}{L}\nabla f(y_k), \end{aligned}$$

whereas in Algorithm 14, we have

$$y_{k+1} = x_k + (\beta_k + 1)\tau_k(z_k - x_k) - \frac{1+\beta_k}{L}\nabla f(y_k).$$

By noting that $\beta_k = \tau_{k+1}(\delta_k - 1)$, we see that the coefficients in front of $\nabla f(y_k)$ match in both expressions. It remains to check that

$$(\beta_k + 1)\tau_k - (\tau_k - \tau_{k+1}(\tau_k - 1)(1 - q\delta_k))$$

Full text available at: http://dx.doi.org/10.1561/2400000036

is identically 0 to reach the desired statement. By substituting $\beta_k = \tau_{k+1}(\delta_k - 1)$, this expression reduces to

$$\tau_{k+1}(\delta_k(\tau_k(1-q)+q)-1),$$

and we have to verify that $(\delta_k(\tau_k(1-q)+q)-1)$ is zero. Substituting and reworking this expression using the expressions for τ_k, and δ_k, we arrive at

$$\frac{\tau_k\left((A_{k+1}-A_k)^2 - A_{k+1} - qA_{k+1}^2\right)}{(A_{k+1}-A_k)(1+qA_{k+1})} = 0,$$

as we recognize that $(A_{k+1}-A_k)^2 - A_{k+1} - qA_{k+1}^2 = 0$ (which is the expression we used to select A_{k+1}). ■

Algorithm 29 Nesterov's method, form III

Input: L-smooth μ-strongly convex function f and initial point x_0.

1: **Initialize** $z_0 = x_0$ and $A_0 = 0$; $q = \mu/L$.
2: **for** $k = 0, \ldots$ **do**
3: $\quad A_{k+1} = \frac{2A_k+1+\sqrt{4A_k+4qA_k^2+1}}{2(1-q)}$
4: $\quad$ set $\tau_k = \frac{(A_{k+1}-A_k)(1+qA_k)}{A_{k+1}+2qA_kA_{k+1}-qA_k^2}$ and $\delta_k = \frac{A_{k+1}-A_k}{1+qA_{k+1}}$
5: $\quad y_k = x_k + \tau_k(z_k - x_k)$
6: $\quad z_{k+1} = (1-q\delta_k)z_k + q\delta_k y_k - \frac{\delta_k}{L}\nabla f(y_k)$
7: $\quad x_{k+1} = \frac{A_k}{A_{k+1}}x_k + (1-\frac{A_k}{A_{k+1}})z_{k+1}$
8: **end for**

Output: Approximate solution x_N.

Proposition B.5. The three sequences $\{z_k\}_k$, $\{x_k\}_k$, and $\{y_k\}_k$ generated by Algorithm 14 are equal to those generated by Algorithm 29.

Proof. Clearly, we have $x_0 = z_0 = y_0$ in both methods. Let us assume that the sequences match up to iteration k, that is, up to y_{k-1}, x_k, and z_k. Since y_k and z_{k+1} are clearly computed in the same way in both methods, we only have to verify that the update rules for x_{k+1} match. In other words, we have to verify that

$$\frac{A_k}{A_{k+1}}x_k + (1-\frac{A_k}{A_{k+1}})z_{k+1} = y_k - \frac{1}{L}\nabla f(y_k),$$

Full text available at: http://dx.doi.org/10.1561/2400000036

which, using the update rules for z_{k+1} and y_k, amounts to verifying that

$$-\frac{(A_{k+1}-A_k)^2 - A_{k+1} - qA_{k+1}^2}{LA_{k+1}(1+qA_{k+1})}\nabla f(y_k) = 0.$$

This statement is true since we recognize $(A_{k+1}-A_k)^2 - A_{k+1} - qA_{k+1}^2 = 0$ as the expression used to select A_{k+1}. ■

B.2 Conjugate Gradient Method

Historically, Nesterov's accelerated gradient method [1] was preceded by a few other methods with optimal worst-case convergence rates $O(N^{-2})$ for smooth convex minimization. However, the alternate schemes required the capability to optimize exactly over a few dimensions—plane-searches were used in [164], [258] and line-searches were used in [259]; unfortunately these references are not available in English, and we refer to [260] for related discussions.

In this vein, accelerated methods can be obtained through their links with conjugate gradients (Algorithm 30), as a by-product of the worst-case analysis. In this section, we illustrate the absolute perfection of the connection between the OGM and conjugate gradients is absolutely perfect: an identical proof (achieving the lower bound) is valid for both methods. The conjugate gradient (CG) method for solving quadratic

Algorithm 30 Conjugate gradient method

Input: L-smooth convex function f, initial point y_0, and budget N.

1: **for** $k = 0, \ldots, N-1$ **do**

2: $\quad y_{k+1} = \operatorname{argmin}_x\{f(x) : x \in y_0 + \operatorname{span}\{\nabla f(y_0), \ldots, \nabla f(y_k)\}\}$

3: **end for**

Output: Approximate solution y_N.

optimization problems is known to have an efficient form that does not require span-searches (which are in general too expensive to be of any practical interest); see, for example, [15]. Beyond quadratics, it is generally not possible to reformulate the CG method in an efficient way. However, it is possible to find other methods for which the same worst-case analysis applies, and it turns out that the OGM is one of

Full text available at: http://dx.doi.org/10.1561/2400000036

them—see [203] for details. Similarly, by slightly weakening the analysis of the CG method, one can find other methods, such as Nesterov's accelerated gradient (see Remark B.1 below for more details).

More precisely, recall the previous definition for the sequence $\{\theta_{k,N}\}_k$, defined in (4.8):

$$\theta_{k+1,N} = \begin{cases} \frac{1+\sqrt{4\theta_{k,N}^2+1}}{2} & \text{if } k \leq N-2 \\ \frac{1+\sqrt{8\theta_{k,N}^2+1}}{2} & \text{if } k = N-1. \end{cases}$$

As a result of the worst-case analysis presented below, all methods satisfying

$$\langle \nabla f(y_i); y_i - \left[\left(1 - \frac{1}{\theta_{i,N}}\right) \left(y_{i-1} - \tfrac{1}{L}\nabla f(y_{i-1})\right) + \frac{1}{\theta_{i,N}} \left(y_0 - \tfrac{2}{L} \sum_{j=0}^{i-1} \theta_{j,N} \nabla f(y_j) \right) \right] \rangle \leq 0 \tag{B.1}$$

achieve the optimal worst-case complexity of smooth convex minimization that is provided by Theorem 4.7. On the one hand, the CG ensures that this inequality holds thanks to its span-searches (which ensure the orthogonality of successive search directions); that is,

$$\begin{aligned} \langle \nabla f(y_i); y_i - y_{i-1} + \frac{1}{\theta_{i,N}}(y_{i-1} - y_0) \rangle &= 0 \\ \langle \nabla f(y_i); \nabla f(y_0) \rangle &= 0 \\ &\vdots \\ \langle \nabla f(y_i); \nabla f(y_{i-1}) \rangle &= 0. \end{aligned}$$

On the other hand, the OGM enforces this inequality by using

$$y_i = \left(1 - \frac{1}{\theta_{i,N}}\right) \left(y_{i-1} - \tfrac{1}{L}\nabla f(y_{i-1})\right) + \frac{1}{\theta_{i,N}} \left(y_0 - \tfrac{2}{L} \sum_{j=0}^{i-1} \theta_{j,N} \nabla f(y_j) \right).$$

Optimized and Conjugate Gradient Methods: Worst-case Analyses

The worst-case analysis below relies on the same potentials used for the optimized gradient method; see Theorem 4.4 and Lemma 4.5.

Full text available at: http://dx.doi.org/10.1561/2400000036

Theorem B.1. Let f be an L-smooth convex function, $N \in \mathbb{N}$ and some $x_\star \in \operatorname{argmin}_x f(x)$. The iterates of the conjugate gradient method (CG, Algorithm 30) and of all methods whose iterates are compliant with (B.1) satisfy

$$f(y_N) - f(x_\star) \leq \frac{L\|y_0 - x_\star\|_2^2}{2\theta_{N,N}^2},$$

for all $y_0 \in \mathbb{R}^d$.

Proof. The result is obtained from the same potential as that used for the OGM, obtained from further inequalities. That is, we first perform a weighted sum of the following inequalities.

- Smoothness and convexity of f between y_{k-1} and y_k with weight $\lambda_1 = 2\theta_{k-1,N}^2$:

$$\begin{aligned} 0 \geq & f(y_k) - f(y_{k-1}) + \langle \nabla f(y_k); y_{k-1} - y_k \rangle \\ & + \frac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|_2^2. \end{aligned}$$

- Smoothness and convexity of f between $x_\star$ and y_k with weight $\lambda_2 = 2\theta_{k,N}$:

$$0 \geq f(y_k) - f(x_\star) + \langle \nabla f(y_k); x_\star - y_k \rangle + \frac{1}{2L}\|\nabla f(y_k)\|_2^2.$$

- Search procedure to obtain y_k, with weight $\lambda_3 = 2\theta_{k,N}^2$:

$$0 \geq \langle \nabla f(y_k); y_k - \left[\left(1 - \frac{1}{\theta_{k,N}}\right) \left(y_{k-1} - \tfrac{1}{L}\nabla f(y_{k-1})\right) + \frac{1}{\theta_{k,N}} z_k \right] \rangle,$$

where we used $z_k := y_0 - \frac{2}{L}\sum_{j=0}^{k-1} \theta_{j,N} \nabla f(y_j)$.

Full text available at: http://dx.doi.org/10.1561/2400000036

The weighted sum is a valid inequality:

$$\begin{aligned}0 \geq & \lambda_1[f(y_k) - f(y_{k-1}) + \langle \nabla f(y_k); y_{k-1} - y_k\rangle \\ & \quad + \frac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|_2^2] \\ & + \lambda_2[f(y_k) - f(x_\star) + \langle \nabla f(y_k); x_\star - y_k\rangle + \frac{1}{2L}\|\nabla f(y_k)\|_2^2] \\ & + \lambda_3[\langle \nabla f(y_k); y_k - \left[\left(1 - \frac{1}{\theta_{k,N}}\right)\left(y_{k-1} - \tfrac{1}{L}\nabla f(y_{k-1})\right)\right. \\ & \quad \left. + \frac{1}{\theta_{k,N}} z_k\right]\rangle].\end{aligned}$$

Substituting z_{k+1}, the previous inequality can be reformulated exactly as

$$\begin{aligned}0 \geq & 2\theta_{k,N}^2\left(f(y_k) - f_\star - \frac{1}{2L}\|\nabla f(y_k)\|_2^2\right) + \frac{L}{2}\|z_{k+1} - x_\star\|_2^2 \\ & - 2\theta_{k-1,N}^2\left(f(y_{k-1}) - f_\star - \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2\right) - \frac{L}{2}\|z_k - x_\star\|_2^2 \\ & + 2\left(\theta_{k-1,N}^2 - \theta_{k,N}^2 + \theta_{k,N}\right)\left(f(y_k) - f_\star + \frac{1}{2L}\|\nabla f(y_k)\|_2^2\right) \\ & + 2\left(\theta_{k-1,N}^2 - \theta_{k,N}^2 + \theta_{k,N}\right)\langle \nabla f(y_k); y_{k-1} - \tfrac{1}{L}\nabla f(y_{k-1}) - y_k\rangle.\end{aligned}$$

We reach the desired inequality by selecting $\theta_{k,N}$ that satisfies $\theta_{k,N} \geq \theta_{k-1,N}$ and

$$\theta_{k-1,N}^2 - \theta_{k,N}^2 + \theta_{k,N} = 0,$$

thereby reaching the same potential as in Theorem 4.4.

To obtain the technical lemma that allows us to bound the final $f(y_N) - f_\star$, we follow the same steps with the following inequalities.

- Smoothness and convexity of f between y_{k-1} and y_k with weight $\lambda_1 = 2\theta_{N-1,N}^2$:

$$\begin{aligned}0 \geq & f(y_N) - f(y_{N-1}) + \langle \nabla f(y_N); y_{N-1} - y_N\rangle \\ & + \frac{1}{2L}\|\nabla f(y_N) - \nabla f(y_{N-1})\|_2^2.\end{aligned}$$

- Smoothness and convexity of f between $x_\star$ and y_k with weight $\lambda_2 = \theta_{N,N}$:

$$0 \geq f(y_N) - f(x_\star) + \langle \nabla f(y_N); x_\star - y_N\rangle + \frac{1}{2L}\|\nabla f(y_N)\|_2^2.$$

Full text available at: http://dx.doi.org/10.1561/2400000036

- Search procedure to obtain y_N, with weight $\lambda_3 = \theta_{N,N}^2$:

$$0 \geq \langle \nabla f(y_N); y_N - \left[\left(1 - \frac{1}{\theta_{N,N}}\right) \left(y_{N-1} - \tfrac{1}{L}\nabla f(y_{N-1})\right) + \frac{1}{\theta_{N,N}} z_N \right] \rangle.$$

The weighted sum can then be reformulated as:

$$\begin{aligned} 0 \geq & \theta_{N,N}^2 \left(f(y_N) - f_\star\right) + \frac{L}{2}\|z_N - \tfrac{\theta_{N,N}}{L}\nabla f(y_N) - x_\star\|_2^2 \\ & - 2\theta_{N-1,N}^2 \left(f(y_{N-1}) - f_\star - \frac{1}{2L}\|\nabla f(y_{N-1})\|_2^2\right) - \frac{L}{2}\|z_N - x_\star\|_2^2 \\ & + \left(2\theta_{N-1,N}^2 - \theta_{N,N}^2 + \theta_{N,N}\right) \left(f(y_N) - f_\star + \frac{1}{2L}\|\nabla f(y_N)\|_2^2\right) \\ & + \left(2\theta_{N-1,N}^2 - \theta_{N,N}^2 + \theta_{N,N}\right) \langle \nabla f(y_N); y_{N-1} - \tfrac{1}{L}\nabla f(y_{N-1}) - y_N \rangle, \end{aligned}$$

thus reaching the desired inequality, as in Lemma 4.5, by selecting $\theta_{N,N}$ that satisfies $\theta_{N,N} \geq \theta_{N-1,N}$ and

$$2\theta_{N-1,N}^2 - \theta_{N,N}^2 + \theta_{N,N}.$$

Hence, the potential argument from Corollary 4.6 applies as such, and we reach the desired conclusion. In other words, for all $k \in \{0, \ldots, N\}$, one can define

$$\phi_k \triangleq 2\theta_{k-1,N}^2 \left(f(y_{k-1}) - f_\star - \frac{1}{2L}\|\nabla f(y_{k-1})\|_2^2\right) + \frac{L}{2}\|z_k - x_\star\|_2^2$$

and

$$\phi_{N+1} \triangleq \theta_{N,N}^2 \left(f(y_N) - f_\star\right) + \frac{L}{2}\|z_N - \tfrac{\theta_{N,N}}{L}\nabla f(y_N) - x_\star\|_2^2$$

and reach the desired statement by chaining the inequalities:

$$\theta_{N,N}^2 (f(y_N) - f_\star) \leq \phi_{N+1} \leq \phi_N \leq \ldots \leq \phi_0 = \frac{L}{2}\|y_0 - x_\star\|_2^2.$$

■

Remark B.1. It is possible to further exploit the conjugate gradient method to design practical accelerated methods in different settings,

Full text available at: http://dx.doi.org/10.1561/2400000036

such as that of Nesterov [1]. This point of view has been exploited in [104], [260]–[262], among others. The link between the CG method and the OGM presented in this section is due to Drori and Taylor [203], though with a different presentation that does not involve the potential function.

B.3 Acceleration Without Monotone Backtracking

B.3.1 FISTA without Monotone Backtracking

In this section, we show how to incorporate backtracking strategies that may not satisfy $L_{k+1} \geq L_k$, which is important in practice. The developments are essentially the same; one possible trick is to incorporate all the knowledge about L_k in A_k. That is, we use a rescaled shape for the potential function:

$$\phi_k \triangleq B_k(f(x_k) - f_\star) + \frac{1 + \mu B_k}{2}\|z_k - x_\star\|_2^2,$$

where without the backtracking strategy, $B_k = \frac{A_k}{L}$. This seemingly cosmetic change allows ϕ_k to depend on L_k solely via B_k, and it applies to both backtracking methods presented in Section 4 (Section 4.7).

The idea used to obtain both methods below is that one can perform the same computations as in Algorithm 14, replacing A_k by $L_{k+1}B_k$ and A_{k+1} by $L_{k+1}A_{k+1}$ at iteration k. Thus, as in previous versions, only the current approximate Lipschitz constant L_{k+1} is used at iteration k: previous approximations were only used to compute B_k.

The proof follows the same lines as used for FISTA (Algorithm 4.20). In this case, f is assumed to be smooth and convex over $\mathbb{R}^d$ (i.e., it has full domain, $\mathbf{dom}\, f = \mathbb{R}^d$), and we are therefore allowed to evaluate gradients of f outside of the domain of h.

Theorem B.2. Let $f \in \mathcal{F}_{\mu,L}$ (with full domain, $\mathbf{dom}\, f = \mathbb{R}^d$), h be a closed convex proper function, $x_\star \in \operatorname{argmin}_x \{F(x) \triangleq f(x) + h(x)\}$, and $k \in \mathbb{N}$. For any $x_k, z_k \in \mathbb{R}^d$ and $B_k \geq 0$, the iterates of Algorithm 31

Full text available at: http://dx.doi.org/10.1561/2400000036

Algorithm 31 Strongly convex FISTA (general initialization of L_{k+1})

Input: An L-smooth (possibly μ-strongly) convex function f, a convex function h with proximal operator available, an initial point x_0, and an initial estimate $L_0 > \mu$.

Initialize $z_0 = x_0$, $B_0 = 0$, and some $\alpha > 1$.
for $k = 0, \ldots$ **do**
  Pick $L_{k+1} \in [L_0, L_k]$.
  loop
    set $q_{k+1} = \mu / L_{k+1}$,
    $B_{k+1} = \frac{2L_{k+1}B_k + 1 + \sqrt{4L_{k+1}B_k + 4\mu L_{k+1}B_k^2 + 1}}{2(L_{k+1} - \mu)}$
    set $\tau_k = \frac{(B_{k+1} - B_k)(1 + \mu B_k)}{(B_{k+1} + 2\mu B_k B_{k+1} - \mu B_k^2)}$ and $\delta_k = L_{k+1}\frac{B_{k+1} - B_k}{1 + \mu B_{k+1}}$
    $y_k = x_k + \tau_k(z_k - x_k)$
    $x_{k+1} = \operatorname{prox}_{h/L_{k+1}}\left(y_k - \frac{1}{L_{k+1}}\nabla f(y_k)\right)$
    $z_{k+1} = (1 - q_{k+1}\delta_k)z_k + q_{k+1}\delta_k y_k + \delta_k\left(x_{k+1} - y_k\right)$
    if (4.21) holds **then**
      break {Iterates accepted; k will be incremented.}
    else
      $L_{k+1} = \alpha L_{k+1}$ {Iterates not accepted; compute new L_{k+1}.}
    end if
  end loop
end for

Output: Approximate solution x_{k+1}.

that satisfy (4.21) also satisfy

$$B_{k+1}(F(x_{k+1}) - F_\star) + \frac{1 + \mu B_{k+1}}{2}\|z_{k+1} - x_\star\|_2^2 \leq B_k(F(x_k) - F_\star) + \frac{1 + \mu B_k}{2}\|z_k - x_\star\|_2^2,$$

with $B_{k+1} = \frac{2L_{k+1}B_k + 1 + \sqrt{4L_{k+1}B_k + 4\mu L_{k+1}B_k^2 + 1}}{2(L_{k+1} - \mu)}$.

Proof. The proof consists of a weighted sum of the following inequalities.

- Strong convexity of f between $x_\star$ and y_k with weight $\lambda_1 = B_{k+1} - B_k$:

$$f_\star \geq f(y_k) + \langle \nabla f(y_k); x_\star - y_k\rangle + \frac{\mu}{2}\|x_\star - y_k\|_2^2.$$

Full text available at: http://dx.doi.org/10.1561/2400000036

- Strong convexity of f between x_k and y_k with weight $\lambda_2 = B_k$:

$$f(x_k) \geq f(y_k) + \langle \nabla f(y_k); x_k - y_k \rangle.$$

- Smoothness of f between y_k and x_{k+1} (*descent lemma*) with weight $\lambda_3 = B_{k+1}$:

$$f(y_k) + \langle \nabla f(y_k); x_{k+1} - y_k \rangle + \frac{L_{k+1}}{2} \|x_{k+1} - y_k\|_2^2 \geq f(x_{k+1}).$$

- Convexity of h between $x_\star$ and x_{k+1} with weight $\lambda_4 = B_{k+1} - B_k$:

$$h(x_\star) \geq h(x_{k+1}) + \langle g_h(x_{k+1}); x_\star - x_{k+1} \rangle,$$

with $g_h(x_{k+1}) \in \partial h(x_{k+1})$ and $x_{k+1} = y_k - \frac{1}{L_{k+1}}(\nabla f(y_k) + g_h(x_{k+1}))$.

- Convexity of h between x_k and x_{k+1} with weight $\lambda_5 = B_k$:

$$h(x_k) \geq h(x_{k+1}) + \langle g_h(x_{k+1}); x_k - x_{k+1} \rangle.$$

We obtain the following inequality:

$$\begin{aligned}
0 \geq & \lambda_1 [f(y_k) - f_\star + \langle \nabla f(y_k); x_\star - y_k \rangle + \frac{\mu}{2} \|x_\star - y_k\|_2^2] \\
& + \lambda_2 [f(y_k) - f(x_k) + \langle \nabla f(y_k); x_k - y_k \rangle] \\
& + \lambda_3 [f(x_{k+1}) - (f(y_k) + \langle \nabla f(y_k); x_{k+1} - y_k \rangle \\
& \quad + \frac{L_{k+1}}{2} \|x_{k+1} - y_k\|_2^2)] \\
& + \lambda_4 [h(x_{k+1}) - h(x_\star) + \langle g_h(x_{k+1}); x_\star - x_{k+1} \rangle] \\
& + \lambda_5 [h(x_{k+1}) - h(x_k) + \langle g_h(x_{k+1}); x_k - x_{k+1} \rangle].
\end{aligned}$$

Substituting the y_k, x_{k+1}, and z_{k+1} with

$$\begin{aligned}
y_k &= x_k + \tau_k (z_k - x_k) \\
x_{k+1} &= y_k - \frac{1}{L_{k+1}} (\nabla f(y_k) + g_h(x_{k+1})) \\
z_{k+1} &= (1 - q_{k+1} \delta_k) z_k + q_{k+1} \delta_k y_k + \delta_k (x_{k+1} - y_k),
\end{aligned}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

after some basic but tedious algebra, yields

$$\begin{aligned}
&B_{k+1}(f(x_{k+1}) + h(x_{k+1}) - f(x_\star) - h(x_\star)) + \frac{1 + B_{k+1}\mu}{2}\|z_{k+1} - x_\star\|_2^2 \\
&\quad \le B_k(f(x_k) + h(x_k) - f(x_\star) - h(x_\star)) + \frac{1 + B_k\mu}{2}\|z_k - x_\star\|_2^2 \\
&\qquad + \frac{L_{k+1}(B_k - B_{k+1})^2 - B_{k+1} - \mu B_{k+1}^2}{1 + \mu B_{k+1}} \\
&\qquad\quad \times \frac{1}{2L_{k+1}}\|\nabla f(y_k) + g_h(x_{k+1})\|_2^2 \\
&\qquad - \frac{B_k^2(B_{k+1} - B_k)(1 + \mu B_k)(1 + \mu B_{k+1})}{\left(B_{k+1} + 2\mu B_k B_{k+1} - \mu B_k^2\right)^2}\frac{\mu}{2}\|x_k - z_k\|_2^2.
\end{aligned}$$

Then, choosing B_{k+1} such that $B_{k+1} \ge B_k$ and

$$L_{k+1}(B_k - B_{k+1})^2 - B_{k+1} - \mu B_{k+1}^2 = 0,$$

yields the desired result:

$$\begin{aligned}
&B_{k+1}(f(x_{k+1}) + h(x_{k+1}) - f(x_\star) - h(x_\star)) + \frac{1 + B_{k+1}\mu}{2}\|z_{k+1} - x_\star\|_2^2 \\
&\quad \le B_k(f(x_k) + h(x_k) - f(x_\star) - h(x_\star)) + \frac{1 + B_k\mu}{2}\|z_k - x_\star\|_2^2. \qquad \blacksquare
\end{aligned}$$

Finally, we obtain a complexity guarantee by adapting the potential argument (4.5) and by noting that B_{k+1} is a decreasing function of L_{k+1} (whose maximal value is αL, assuming $L_0 < L$; otherwise, its maximal value is L_0). The growth rate of B_k in the smooth convex setting remains unchanged (see (4.14)) since we have

$$B_{k+1} \ge \frac{\left(\frac{1}{2} + \sqrt{B_k L_{k+1}}\right)^2}{L_{k+1}},$$

and hence, $\sqrt{B_{k+1}} \ge \frac{1}{2\sqrt{L_{k+1}}} + \sqrt{B_k}$. Therefore, $B_k \ge \left(\frac{k}{2\sqrt{\ell}}\right)^2$ with $\ell = \max\{L_0, \alpha L\}$ and $L_{k+1} \le \ell$. As for the geometric rate, we similarly obtain

$$B_{k+1} \ge B_k \frac{\left(1 + \sqrt{\frac{\mu}{L_{k+1}}}\right)}{1 - \frac{\mu}{L_{k+1}}} = \frac{B_k}{1 - \sqrt{\frac{\mu}{L_{k+1}}}},$$

and therefore, $B_{k+1} \ge (1 - \sqrt{\frac{\mu}{\ell}})^{-1} B_k$.

Full text available at: http://dx.doi.org/10.1561/2400000036

Corollary B.3. Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$ (with full domain, $\mathbf{dom}\, f = \mathbb{R}^d$), h be a closed convex proper function and $x_\star \in \operatorname{argmin}_x \{F(x) \triangleq f(x)+h(x)\}$. For any $N \in \mathbb{N}$, $N \geq 1$, and $x_0 \in \mathbb{R}^d$, the output of Algorithm 31 satisfies

$$F(x_N) - F_\star \leq \min\left\{\frac{2}{N^2}, \left(1-\sqrt{\frac{\mu}{\ell}}\right)^N\right\} \ell \|x_0 - x_\star\|_2^2,$$

with $\ell = \max\{\alpha L, L_0\}$.

Proof. We assume that $L > L_0$ since otherwise, $f \in \mathcal{F}_{\mu,L_0}$ and the proof directly follows from the case without backtracking. The chained potential argument (4.5) can be used as before. Using $B_0 = 0$, we reach

$$F(x_N) - F_\star \leq \frac{\|x_0 - x_\star\|_2^2}{2B_N}.$$

Our previous bounds on B_N yields the desired result, using

$$B_1 = \frac{1}{L_{k+1} - \mu} \geq \frac{2\ell^{-1}}{1-\frac{\mu}{\ell}} = \frac{2\ell^{-1}}{\left(1-\sqrt{\frac{\mu}{\ell}}\right)\left(1+\sqrt{\frac{\mu}{\ell}}\right)} \geq \frac{\ell^{-1}}{1-\sqrt{\frac{\mu}{\ell}}},$$

and hence, $B_N \geq \ell^{-1}\left(1-\sqrt{\frac{\mu}{\ell}}\right)^{-N}$ as well as $B_k \geq \left(\frac{k}{2\sqrt{\ell}}\right)^2$. ∎

B.3.2 Another Accelerated Method without Monotone Backtracking

Just as for FISTA, we can perform the same cosmetic change to Algorithm 20 for incorporating a non-monotonic estimations of the Lipschitz constant. The proof is therefore essentially that of Algorithm 20.

Full text available at: http://dx.doi.org/10.1561/2400000036

Algorithm 32 A proximal accelerated gradient (general initialization of L_{k+1})

Input: $h \in \mathcal{F}_{0,\infty}$ with proximal operator available, $f \in \mathcal{F}_{\mu,L}(\mathbf{dom}\, h)$, an initial point $x_0 \in \mathbf{dom}\, h$, and an initial estimate $L_0 > \mu$.

Initialize $z_0 = x_0$, $A_0 = 0$, and some $\alpha > 1$.
for $k = 0, \ldots$ **do**
  Pick $L_{k+1} \in [L_0, L_k]$.
  loop
    Set $q_{k+1} = \mu / L_{k+1}$,
    $B_{k+1} = \frac{2L_{k+1}B_k + 1 + \sqrt{4L_{k+1}B_k + 4\mu L_{k+1}B_k^2 + 1}}{2(L_{k+1} - \mu)}$
    Set $\tau_k = \frac{L_{k+1}(B_{k+1} - B_k)(1 + \mu B_k)}{L_{k+1}(B_{k+1} + 2\mu B_k B_{k+1} - \mu B_k^2)}$ and $\delta_k = L_{k+1}\frac{B_{k+1} - B_k}{1 + \mu B_{k+1}}$
    $y_k = x_k + \tau_k(z_k - x_k)$
    $z_{k+1} = \mathrm{prox}_{\delta_k h / L_{k+1}}\left((1 - q_{k+1}\delta_k)z_k + q_{k+1}\delta_k y_k - \frac{\delta_k}{L_{k+1}}\nabla f(y_k)\right)$
    $x_{k+1} = \frac{A_k}{A_{k+1}}x_k + (1 - \frac{A_k}{A_{k+1}})z_{k+1}$
    if (4.21) holds **then**
      break {Iterates accepted; k will be incremented.}
    else
      $L_{k+1} = \alpha L_{k+1}$ {Iterates not accepted; compute new L_{k+1}.}
    end if
  end loop
end for

Output: An approximate solution x_{k+1}.

Full text available at: http://dx.doi.org/10.1561/2400000036

Theorem B.4. Let $h \in \mathcal{F}_{0,\infty}$, $f \in \mathcal{F}_{\mu,L}(\mathbf{dom}\, h)$, $x_\star \in \operatorname{argmin}_x \{F(x) \triangleq f(x) + h(x)\}$, and $k \in \mathbb{N}$. For any $x_k, z_k \in \mathbb{R}^d$ and $B_k \geq 0$, the iterates of Algorithm 32 that satisfy (4.21) also satisfy

$$\begin{aligned} B_{k+1}(F(x_{k+1}) - F_\star) + \frac{1+\mu B_{k+1}}{2}\|z_{k+1} - x_\star\|_2^2 \\ \leq B_k(F(x_k) - F_\star) + \frac{1+\mu B_k}{2}\|z_k - x_\star\|_2^2, \end{aligned}$$

with $B_{k+1} = \frac{2L_{k+1}B_k + 1 + \sqrt{4L_{k+1}B_k + 4\mu L_{k+1}B_k^2 + 1}}{2(L_{k+1} - \mu)}$.

Proof. First, $\{z_k\}_k$ is in $\mathbf{dom}\, h$ by construction—it is the output of a proximal/projection step. Furthermore, we have $0 \leq \frac{B_k}{B_{k+1}} \leq 1$ given that $B_{k+1} \geq B_k \geq 0$. A direct consequence is that since $z_0 = x_0 \in \mathbf{dom}\, h$, all subsequent $\{y_k\}_k$ and $\{x_k\}_k$ are also in $\mathbf{dom}\, h$ (as they are obtained from convex combinations of feasible points).

The rest of the proof consists of a weighted sum of the following inequalities (which are valid due to the feasibility of the iterates).

- Strong convexity of f between $x_\star$ and y_k with weight $\lambda_1 = B_{k+1} - B_k$:

$$f(x_\star) \geq f(y_k) + \langle \nabla f(y_k); x_\star - y_k \rangle + \frac{\mu}{2}\|x_\star - y_k\|_2^2.$$

- Convexity of f between x_k and y_k with weight $\lambda_2 = B_k$:

$$f(x_k) \geq f(y_k) + \langle \nabla f(y_k); x_k - y_k \rangle.$$

- Smoothness of f between y_k and x_{k+1} (*descent lemma*) with weight $\lambda_3 = B_{k+1}$:

$$f(y_k) + \langle \nabla f(y_k); x_{k+1} - y_k \rangle + \frac{L_{k+1}}{2}\|x_{k+1} - y_k\|_2^2 \geq f(x_{k+1}).$$

- Convexity of h between $x_\star$ and z_{k+1} with weight $\lambda_4 = B_{k+1} - B_k$:

$$h(x_\star) \geq h(z_{k+1}) + \langle g_h(z_{k+1}); x_\star - z_{k+1} \rangle,$$

with $g_h(z_{k+1}) \in \partial h(z_{k+1})$ and $z_{k+1} = (1 - q\delta_k)z_k + q\delta_k y_k - \frac{\delta_k}{L_{k+1}}(\nabla f(y_k) + g_h(z_{k+1}))$.

Full text available at: http://dx.doi.org/10.1561/2400000036

- Convexity of h between x_k and x_{k+1} with weight $\lambda_5 = B_k$:
$$h(x_k) \geq h(x_{k+1}) + \langle g_h(x_{k+1}); x_k - x_{k+1}\rangle,$$
with $g_h(x_{k+1}) \in \partial h(x_{k+1})$.
- Convexity of h between z_{k+1} and x_{k+1} with weight $\lambda_6 = B_{k+1} - B_k$:
$$h(z_{k+1}) \geq h(x_{k+1}) + \langle g_h(x_{k+1}); z_{k+1} - x_{k+1}\rangle.$$

We obtain the following inequality:

$$\begin{aligned}
0 \geq & \lambda_1[f(y_k) - f_\star + \langle \nabla f(y_k); x_\star - y_k\rangle + \frac{\mu}{2}\|x_\star - y_k\|_2^2] \\
& + \lambda_2[f(y_k) - f(x_k) + \langle \nabla f(y_k); x_k - y_k\rangle] \\
& + \lambda_3[f(x_{k+1}) - (f(y_k) + \langle \nabla f(y_k); x_{k+1} - y_k\rangle \\
& \quad + \frac{L_{k+1}}{2}\|x_{k+1} - y_k\|_2^2)] \\
& + \lambda_4[h(z_{k+1}) - h(x_\star) + \langle g_h(z_{k+1}); x_\star - z_{k+1}\rangle] \\
& + \lambda_5[h(x_{k+1}) - h(x_k) + \langle g_h(x_{k+1}); x_k - x_{k+1}\rangle] \\
& + \lambda_6[h(x_{k+1}) - h(z_{k+1}) + \langle g_h(x_{k+1}); z_{k+1} - x_{k+1}\rangle].
\end{aligned}$$

Substituting the y_k, z_{k+1}, and x_{k+1} by

$$\begin{aligned}
y_k &= x_k + \tau_k(z_k - x_k) \\
z_{k+1} &= (1 - q_{k+1}\delta_k)z_k + q_{k+1}\delta_k y_k - \frac{\delta_k}{L_{k+1}}(\nabla f(y_k) + g_h(z_{k+1})) \\
x_{k+1} &= \frac{B_k}{B_{k+1}}x_k + \left(1 - \frac{B_k}{B_{k+1}}\right) z_{k+1},
\end{aligned}$$

and algebra allows us to obtain the following reformulation:

$$\begin{aligned}
& B_{k+1}(f(x_{k+1}) + h(x_{k+1}) - f(x_\star) - h(x_\star)) + \frac{1 + \mu B_{k+1}}{2}\|z_{k+1} - x_\star\|_2^2 \\
\leq & B_k(f(x_k) + h(x_k) - f(x_\star) - h(x_\star)) + \frac{1 + \mu B_k}{2}\|z_k - x_\star\|_2^2 \\
& + \frac{(B_k - B_{k+1})^2\left(L_{k+1}(B_k - B_{k+1})^2 - B_{k+1} - \mu B_{k+1}^2\right)}{B_{k+1}(1 + \mu B_{k+1})^2} \\
& \quad \times \frac{1}{2}\|\nabla f(y_k) + g_h(z_{k+1})\|_2^2 \\
& - \frac{B_k^2(B_{k+1} - B_k)(1 + \mu B_k)(1 + \mu B_{k+1})}{\left(B_{k+1} + 2\mu B_k B_{k+1} - \mu B_k^2\right)^2}\frac{\mu}{2}\|x_k - z_k\|_2^2.
\end{aligned}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

The desired inequality follows from selecting B_{k+1} such that $B_{k+1} \geq B_k$ and

$$L_{k+1}(B_k - B_{k+1})^2 - B_{k+1} - \mu B_{k+1}^2 = 0,$$

thereby yielding

$$\begin{aligned} B_{k+1}(f(x_{k+1}) + h(x_{k+1}) - f(x_\star) - h(x_\star)) + \frac{1 + \mu B_{k+1}}{2}\|z_{k+1} - x_\star\|_2^2 \\ \leq B_k(f(x_k) + h(x_k) - f(x_\star) - h(x_\star)) + \frac{1 + B_k\mu}{2}\|z_k - x_\star\|_2^2. \end{aligned}$$

■

The final corollary follows from the same arguments as those used for Corollary B.3. It provides the final bound for Algorithm 32.

Corollary B.5. Let $h \in \mathcal{F}_{0,\infty}$, $f \in \mathcal{F}_{\mu,L}(\mathbf{dom}\, h)$, and $x_\star \in \operatorname{argmin}_x \{F(x) \triangleq f(x) + h(x)\}$. For any $N \in \mathbb{N}$, $N \geq 1$, and $x_0 \in \mathbb{R}^d$, the output of Algorithm 32 satisfies

$$F(x_N) - F_\star \leq \min\left\{\frac{2}{N^2}, \left(1 - \sqrt{\frac{\mu}{\ell}}\right)^{-N}\right\} \ell\|x_0 - x_\star\|_2^2,$$

with $\ell = \max\{\alpha L, L_0\}$.

Proof. The proof follows the same arguments as those for Corollary B.3, using the potential from Theorem B.4 and the fact that the output of the algorithm satisfies (4.21). ■

Full text available at: http://dx.doi.org/10.1561/2400000036

C

On Worst-case Analyses for First-order Methods

C.1 Principled Approaches to Worst-case Analyses

In this section, we show that obtaining convergence rates and proofs can be framed as finding feasible points to certain convex problems. More precisely, all convergence guarantees from Section 4 and Section 5 can be obtained as feasible points to certain linear matrix inequalities (LMI). As we see in what follows, this approach can be seen as a *principled* approach to worst-case analysis of first-order methods: the approach fails only when no such guarantees can be found. The purpose of this section is to provide complete examples of the LMIs for a few cases of interest: analyses of gradient and accelerated gradient methods, as well as pointers to the relevant literature. We provide a full derivation for the base case, and leave advanced ones as exercises for the reader. Notebooks for obtaining the corresponding LMIs are provided in Section C.5.

The elements of this section are largely inspired by the presentation of Taylor and Bach (2019) with elements borrowed from the presentation of Taylor, Hendrickx and Glineur (2017), which is itself largely inspired by that of Drori and Teboulle (2014). The arguments are also similar to the line of work by Lessard, Recht and Packard (2016) and follow-up works, see, e.g., [211], [213]. The latter line of works is similar in spirit to the former, but framed in control-theoretic terms, via so-called *integral*

Full text available at: http://dx.doi.org/10.1561/2400000036

quadratic constraints, popularized by Megretski and Rantzer [263].

These techniques are analogous and mostly differs in their presentation styles. Roughly speaking, they can be seen as *dual* to each others. That is, whereas the *performance estimation* viewpoint stems from the problem of computing worst-case scenarios and approaches worst-case guarantees as feasible point to the corresponding dual problems, the *integral quadratic constraint* approach directly starts from the problem of performing linear combination of inequalities, which is exactly the dual problem to that of computing worst-case scenarios. Depending on the background of the researchers involved in a work on one of those topics, things might therefore be named in different ways. We insist on the fact that those are really two facets of the same coin with only subtle differences in terms of presentations.

We choose to take the performance estimation viewpoint as using the definition of a "worst-case" allows to carefully select the most appropriate set of inequalities to be used. Informally, this advantageous construction allows certifying the approach to provide meaningful worst-case guarantees: either the approach provides a satisfying worst-case guarantee, or there exists a non-satisfying counterexample, invalidating the existence of any satisfying guarantee of the desired form.

Further discussions and a more thorough list of references are provided in Section C.5. Readability in mind, the presentation focuses on some examples of interest rather than on a general framework. We refer to [3], [201], [210] for more details.

C.2 Worst-case Analysis as Optimization/Feasibility Problems

In this section, we provide examples illustrating the type of problems that can be used for obtaining worst-case guarantees. The base idea underlying the technique is that worst-case scenarios are by definition solutions to certain optimization problems. In the context of first-order convex optimization methods, those worst-case scenarios correspond to solutions to linear semidefinite programs (SDP), which are convex; see, e.g., [264]. It nicely follows from this theory that any worst-case guarantee (i.e., any upper bound on a worst-case performance) can be formulated as a feasible point to the dual problem to that of finding

Full text available at: http://dx.doi.org/10.1561/2400000036

worst-case scenarios. Equivalently, those dual solutions correspond to appropriate weighted sums of inequalities, whose weights correspond to the values of the dual variables. Proofs from Section 4 and Section 5 correspond to such dual certificates.

Those statements are made more precise in the next sections. We begin by providing a few examples of LMIs that can be used for designing worst-case guarantees.

Preview: worst-case guarantees via LMIs. Perhaps the most basic LMI that can be presented for obtaining worst-case guarantees concerns gradient descent and its convergence in terms of distance to an optimal point. We present it for simplicity, as the corresponding LMI only involves very few variables. This LMI has also relatively simple solutions. As our target here is to present the approach, we let finding their solutions as exercises. We present the LMIs in their most *raw* forms, even without a few direct simplifications.

Note that those LMIs always involve $n(n-1)$ "dual" variables (the precise meaning of *dual* becomes clear in the sequel), where n is the number of points at which the type of guarantee under consideration requires using or specifying a function or gradient evaluation (either in the algorithm or for computing the value of the guarantee). In the following example, we need two dual variables because the guarantee only requires using two gradients of f, namely $\nabla f(x_k)$ (for expressing a gradient step $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$) and $\nabla f(x_\star)$ (for expressing optimality of $x_\star$ as $\nabla f(x_\star) = 0$).

Theorem C.1. Let $\tau \geq 0$ and $\gamma_k \in \mathbb{R}$. The inequality

$$\|x_{k+1} - x_\star\|_2^2 \leq \tau \|x_k - x_\star\|_2^2 \tag{C.1}$$

holds for all $d \in \mathbb{N}$, all $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, all $x_k, x_{k+1}, x_\star \in \mathbb{R}^d$ (such that $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ and $\nabla f(x_\star) = 0$), if and only if

$$\exists \lambda_1, \lambda_2 \geq 0 : \begin{cases} \lambda_1 = \lambda_2 \\ 0 \preceq \begin{bmatrix} \tau - 1 + \frac{\mu L(\lambda_1+\lambda_2)}{2(L-\mu)} & \gamma_k - \frac{L\lambda_1+\mu\lambda_2}{2(L-\mu)} \\ \gamma_k - \frac{L\lambda_1+\mu\lambda_2}{2(L-\mu)} & -\gamma_k^2 + \frac{\lambda_1+\lambda_2}{2(L-\mu)} \end{bmatrix}. \end{cases} \tag{C.2}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

We emphasize that the message underlying Theorem C.1 is that verifying a worst-case convergence guarantee of the form (C.1) boils down to verifying the feasibility of a certain convex problem. It is relatively straightforward to convert a feasible point of (C.2) to a proof that only consists of a weighted linear combination of inequalities, see, e.g., [265, Theorem 3.1]. The corresponding weights are the values of the multipliers (that is, in Theorem C.1, the weights are λ_1 and λ_2) as showcased in Section 4 and Section 5.

As we see in Section C.3, changing the Lyapunov, or potential, function to be verified also changes the LMI to be solved. The desired LMI can be obtained following a principled approach presented in the sequel. In particular, the following result is slightly more complicated and corresponds to verifying the potential provided by Theorem 4.2. One should note that those LMIs can be solved numerically, providing nice guides for choosing appropriate analytical weights. Symbolic computations and computer algebra software might also help.

The following LMI relies on 6 *dual variables* $\lambda_1, \dots, \lambda_6$ as it involves gradients and/or function values of $f(\cdot)$ at three points: x_k, x_{k+1}, and $x_\star$, thereby fixing $n = 3$ and hence $n(n-1) = 6$ dual variables.

Theorem C.2. Let $A_{k+1}, A_k \geq 0$ and $\gamma_k \in \mathbb{R}$. The inequality

$$A_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{L}{2}\|x_{k+1} - x_\star\|_2^2 \leq A_k(f(x_k) - f_\star) + \tfrac{L}{2}\|x_k - x_\star\|_2^2$$

holds for all $d \in \mathbb{N}$, all $f \in \mathcal{F}_L(\mathbb{R}^d)$, all $x_k, x_{k+1}, x_\star \in \mathbb{R}^d$ (such that $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ and $\nabla f(x_\star) = 0$) if and only if

$$\exists \lambda_1, \lambda_2, \dots, \lambda_6 \geq 0 : \begin{cases} 0 = A_k + \lambda_1 + \lambda_2 - \lambda_4 - \lambda_6 \\ 0 = -A_{k+1} - \lambda_2 + \lambda_3 + \lambda_4 - \lambda_5 \\ 0 \preceq \begin{bmatrix} 0 & \star & \star \\ \frac{1}{2}(\gamma_k L - \lambda_1) & \frac{\lambda_1+\lambda_2+\lambda_4+\lambda_6-\gamma_k^2 L^2-2\gamma_k L\lambda_2}{2L} & \star \\ -\frac{\lambda_3}{2} & \frac{1}{2}\left(\gamma_k(\lambda_3+\lambda_4) - \frac{\lambda_2+\lambda_4}{L}\right) & \frac{\lambda_2+\lambda_3+\lambda_4+\lambda_5}{2L} \end{bmatrix}, \end{cases}$$

(where $\star$'s denote symmetric elements in the matrix).

Remark C.1. The LMIs of this section are put in their "raw" forms, for simplicity of the presentation (which does not focus on solving those

Full text available at: http://dx.doi.org/10.1561/2400000036

LMIs analytically. Of course, a few simplifications are relatively direct: for instance, any feasible point will have $\lambda_1 = \gamma_k L$ and $\lambda_3 = 0$, as the corresponding matrix could not be positive semidefinite otherwise.

As we discuss in the sequel (see Remark C.4), it is also relatively straightforward to obtain weaker versions of those LMIs which are then only sufficient for obtaining valid worst-case guarantees. Those simplified LMIs might be simpler to solve analytically, and might therefore be advantageous in certain contexts. Brief discussions and pointers for this topic are provided in Remark C.4 and Section C.5.

A strongly convex version of Theorem C.2 is provided in Theorem C.5. It is slightly more algrebaic in its vanilla form, but allows recovering the results of Theorem 4.10 as a feasible point. Analyses of accelerated methods can be obtained in a similar way, as illustrated by the following LMI. The latter uses on 12 *dual variables* $\lambda_1, \dots, \lambda_{12}$, as it relies on evaluating gradients and/or function values of $f(\cdot)$ at four points: y_k, x_k, x_{k+1}, and $x_\star$, so $n = 4$ and hence $n(n-1) = 12$. Although this LMI might appear as a bit of a brutal approach to worst-case analysis, one might observe that many of elements of the LMI can be set to zero due to the structure of the problem.

Theorem C.3. Let $A_{k+1}, A_k \geq 0$ and $\alpha_k, \gamma_k, \tau_k \in \mathbb{R}$, and consider the iteration

$$\begin{aligned} y_k &= x_k + \tau_k(z_k - x_k) \\ x_{k+1} &= y_k - \alpha_k \nabla f(y_k) \\ z_{k+1} &= z_k - \gamma_k \nabla f(y_k). \end{aligned} \tag{C.3}$$

The inequality

$$A_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{L}{2}\|z_{k+1} - x_\star\|_2^2 \leq A_k(f(x_k) - f_\star) + \tfrac{L}{2}\|z_k - x_\star\|_2^2$$

holds for all $d \in \mathbb{N}$, all $f \in \mathcal{F}_L(\mathbb{R}^d)$, and all $x_k, x_{k+1}, z_k, z_{k+1}, x_\star \in \mathbb{R}^d$ (such that x_{k+1}, z_{k+1} are generated by (C.3) and $\nabla f(x_\star) = 0$) if and

Full text available at: http://dx.doi.org/10.1561/2400000036

only if

$$\exists \lambda_1,\lambda_2,\ldots,\lambda_{12} \geq 0 : \begin{cases} 0 = A_k + \lambda_1 + \lambda_2 - \lambda_4 - \lambda_6 - \lambda_8 + \lambda_{11} \\ 0 = -A_{k+1} - \lambda_2 + \lambda_3 + \lambda_4 - \lambda_5 - \lambda_9 + \lambda_{12} \\ 0 = \lambda_7 + \lambda_8 + \lambda_9 - \lambda_{10} - \lambda_{11} - \lambda_{12} \\ 0 \preceq \begin{bmatrix} 0 & 0 & S_{1,3} & S_{1,4} & S_{1,5} \\ 0 & 0 & S_{2,3} & S_{2,4} & S_{2,5} \\ S_{1,3} & S_{2,3} & S_{3,3} & S_{3,4} & S_{3,5} \\ S_{1,4} & S_{2,4} & S_{3,4} & S_{4,4} & S_{4,5} \\ S_{1,5} & S_{2,5} & S_{3,5} & S_{4,5} & S_{5,5} \end{bmatrix}, \end{cases}$$

with

$$\begin{aligned} S_{1,3} &= \tfrac{1}{2}(\lambda_7(\tau_k - 1) + \lambda_8\tau_k), \\ S_{1,4} &= -\tfrac{1}{2}(\lambda_1 + \tau_k(\lambda_2 + \lambda_{11})), \ S_{1,5} = \tfrac{1}{2}(\lambda_3(\tau_k - 1) + \lambda_4\tau_k), \\ S_{2,3} &= \tfrac{1}{2}(\gamma_k L - \tau_k(\lambda_7 + \lambda_8)), \\ S_{2,4} &= \tfrac{1}{2}\tau_k(\lambda_2 + \lambda_{11}), \ S_{2,5} = -\tfrac{1}{2}\tau_k(\lambda_3 + \lambda_4), \\ S_{3,3} &= \frac{\lambda_7 + \lambda_8 + \lambda_9 + \lambda_{10} + \lambda_{11} + \lambda_{12} - \gamma_k^2 L^2 - 2\alpha_k L \lambda_9}{2L}, \\ S_{3,4} &= -\frac{\alpha_k L \lambda_2 + \lambda_8 + \lambda_{11}}{2L}, \ S_{3,5} = \tfrac{1}{2}\left(\alpha_k(\lambda_3 + \lambda_4 + \lambda_{12}) - \frac{\lambda_9 + \lambda_{12}}{L}\right), \\ S_{4,4} &= \frac{\lambda_1 + \lambda_2 + \lambda_4 + \lambda_6 + \lambda_8 + \lambda_{11}}{2L}, \ S_{4,5} = -\frac{\lambda_2 + \lambda_4}{2L}, \\ S_{5,5} &= \frac{\lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_9 + \lambda_{12}}{2L}. \end{aligned}$$

C.3 Analysis of Gradient Descent via Linear Matrix Inequalities

In this section, we detail the approach to obtain LMIs such as those of Theorem C.1, Theorem C.2 and Theorem C.3. We provide full details for gradient descent. The same technique is presented in a more expeditious way for its accelerated versions afterwards.

C.3.1 Linear Convergence of Gradient Descent

We consider gradient descent for minimizing smooth strongly convex functions. For exposition purposes, we investigate a type of one-iteration

Full text available at: http://dx.doi.org/10.1561/2400000036

worst-case convergence guarantee in terms of the distance to the optimum (see Theorem C.1) for gradient descent, of the form:

$$\|x_{k+1} - x_\star\|_2^2 \leq \tau \|x_k - x_\star\|_2^2 \tag{C.4}$$

which are valid for all $d \in \mathbb{N}$, $x_k, x_{k+1}, x_\star \in \mathbb{R}^d$ and all $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$ (L-smooth μ-strongly convex function) when $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ (gradient descent) and $\nabla f(x_\star) = 0$ ($x_\star$ is optimal for f). In this context, we denote by $\tau_\star$ (we omit the dependence on γ_k, μ, and L for convenience) the smallest value τ for which (C.4) is valid. By definition, this value can be formulated as the solution to an optimization problem looking for worst-case scenarios:

$$\begin{aligned} \tau_\star \triangleq \max_{\substack{d,f \\ x_k, x_{k+1}, x_\star}} \ & \frac{\|x_{k+1} - x_\star\|_2^2}{\|x_k - x_\star\|_2^2} \\ \text{s.t. } & d \in \mathbb{N}, f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d) \\ & x_k, x_{k+1}, x_\star \in \mathbb{R}^d \\ & x_{k+1} = x_k - \gamma_k \nabla f(x_k) \\ & \nabla f(x_\star) = 0. \end{aligned} \tag{C.5}$$

As it is, this problem does not look quite practical. However, it actually admits an equivalent formulation as a linear semidefinite program. As a first step for reaching this formulation, the previous problem can be formulated in an equivalent *sampled* manner. That is, we sample f at the points where the first-order information is explicitly used:

$$\begin{aligned} \tau_\star = \max_{\substack{d \\ f_k, f_\star \\ g_k, g_\star \\ x_k, x_{k+1}, x_\star}} \ & \frac{\|x_{k+1} - x_\star\|_2^2}{\|x_k - x_\star\|_2^2} \\ \text{s.t. } & d \in \mathbb{N}, f_k, f_\star \in \mathbb{R} \\ & x_k, x_{k+1}, x_\star, g_k, g_\star \in \mathbb{R}^d \\ & \exists f \in \mathcal{F}_{\mu,L} : \begin{cases} f_k = f(x_k) \text{ and } g_k = \nabla f(x_k) \\ f_\star = f(x_\star) \text{ and } g_\star = \nabla f(x_\star) \end{cases} \\ & g_\star = 0 \\ & x_{k+1} = x_k - \gamma_k g_k, \end{aligned} \tag{C.6}$$

and f is now represented in terms of its samples at $x_\star$ and x_k.

Full text available at: http://dx.doi.org/10.1561/2400000036

A second stage in this reformulation consists of replacing the existence of a certain $f \in \mathcal{F}_{\mu,L}$ interpolating (or extending) the samples by an equivalent explicit condition provided by the following theorem.

Theorem C.4 ($\mathcal{F}_{\mu,L}$-interpolation, Theorem 4 in [210])**.** Let $L > \mu \geq 0$, I be an index set and $S = \{(x_i, g_i, f_i)\}_{i\in I} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be a set of triplets. There exists $f \in \mathcal{F}_{\mu,L}$ satisfying $f(x_i) = f_i$ and $g_i \in \partial f(x_i)$ for all $i \in I$ if and only if

$$\begin{aligned} f_i \geq f_j + \langle g_j; x_i - x_j\rangle + \frac{1}{2L}\|g_i - g_j\|_2^2 \\ + \frac{\mu}{2(1-\mu/L)}\|x_i - x_j - \tfrac{1}{L}(g_i - g_j)\|_2^2 \end{aligned} \tag{C.7}$$

holds for all $i, j \in I$.

Theorem C.4 conveniently allows replacing the existence constraint by a set of quadratic inequalities, reaching:

$$\begin{aligned} \tau_\star = \max_{\substack{d \\ f_k, f_\star \\ g_k, x_k, x_\star}} \ & \frac{\|x_k - \gamma_k g_k - x_\star\|_2^2}{\|x_k - x_\star\|_2^2} \\ \text{s.t. } & d \in \mathbb{N}, f_k, f_\star \in \mathbb{R} \\ & x_k, x_\star, g_k \in \mathbb{R}^d \\ & f_\star \geq f_k + \langle g_k; x_\star - x_k\rangle + \frac{1}{2L}\|g_k\|_2^2 \\ & \qquad + \frac{\mu}{2(1-\mu/L)}\|x_k - \frac{1}{L}g_k - x_\star\|_2^2 \\ & f_k \geq f_\star + \frac{1}{2L}\|g_k\|_2^2 \\ & \qquad + \frac{\mu}{2(1-\mu/L)}\|x_k - \frac{1}{L}g_k - x_\star\|_2^2, \end{aligned} \tag{C.8}$$

where we also substituted x_{k+1} and $g_\star$ by their respective expressions. Finally, we arrive to a first (convex) semidefinite reformulation of the problem via new variables: $G \succeq 0$ and F defined as

$$G \triangleq \begin{bmatrix} \|x_k - x_\star\|_2^2 & \langle g_k, x_k - x_\star\rangle \\ \langle g_k, x_k - x_\star\rangle & \|g_k\|_2^2 \end{bmatrix}, \quad F \triangleq f_k - f_\star.$$

Full text available at: http://dx.doi.org/10.1561/2400000036

The problem turns out to be linear in G and F:

$$\begin{aligned}\tau_\star = \max_{G,F}\ & \frac{G_{1,1}+\gamma_k^2 G_{2,2}-2\gamma_k G_{1,2}}{G_{1,1}}\\ \text{s.t. } & F\in\mathbb{R}, G\in\mathbb{S}^2\\ & G\succeq 0\\ & F+\tfrac{L\mu}{2(L-\mu)}G_{1,1}+\tfrac{1}{2(L-\mu)}G_{2,2}-\tfrac{L}{L-\mu}G_{1,2}\leq 0\\ & -F+\tfrac{L\mu}{2(L-\mu)}G_{1,1}+\tfrac{1}{2(L-\mu)}G_{2,2}-\tfrac{\mu}{L-\mu}G_{1,2}\leq 0.\end{aligned} \tag{C.9}$$

Finally, a simple homogeneity argument (for any feasible (G,F) to (C.9), the pair $(\tilde{G},\tilde{F})\triangleq(G/G_{1,1},F/G_{1,1})$ is also feasible with the same objective value, with $\tilde{G}_{1,1}=1$ so we can assume without loss of generality that $G_{1,1}=1$ without changing the optimal value of the problem—note that it is relatively straightforward to establish that the optimal solution satisfies $G_{1,1}\neq 0$) allows arriving to the equivalent:

$$\begin{aligned}\tau_\star = \max_{G,F}\ & G_{1,1}+\gamma_k^2 G_{2,2}-2\gamma_k G_{1,2}\\ \text{s.t. } & F\in\mathbb{R}, G\in\mathbb{S}^2\\ & G\succeq 0\\ & F+\tfrac{L\mu}{2(L-\mu)}G_{1,1}+\tfrac{1}{2(L-\mu)}G_{2,2}-\tfrac{L}{L-\mu}G_{1,2}\leq 0\\ & -F+\tfrac{L\mu}{2(L-\mu)}G_{1,1}+\tfrac{1}{2(L-\mu)}G_{2,2}-\tfrac{\mu}{L-\mu}G_{1,2}\leq 0\\ & G_{1,1}=1.\end{aligned} \tag{C.10}$$

For arriving to the desired LMI, it remains to dualize the problem. That is, we perform the following primal-dual associations:

$$\begin{aligned} F+\tfrac{L\mu}{2(L-\mu)}G_{1,1}+\tfrac{1}{2(L-\mu)}G_{2,2}-\tfrac{L}{L-\mu}G_{1,2}\leq 0 \quad &: \lambda_1,\\ -F+\tfrac{L\mu}{2(L-\mu)}G_{1,1}+\tfrac{1}{2(L-\mu)}G_{2,2}-\tfrac{\mu}{L-\mu}G_{1,2}\leq 0 \quad &: \lambda_2,\\ G_{1,1}=1 \quad &: \tau.\end{aligned}$$

Standard Lagrangian duality allows arriving to

$$\begin{aligned}\tau_\star = \min_{\lambda_1,\lambda_2,\tau\geq 0}\ & \tau\\ \text{s.t. } & \lambda_1=\lambda_2\\ & 0\preceq\begin{bmatrix}\tau-1+\frac{\mu L(\lambda_1+\lambda_2)}{2(L-\mu)} & \gamma_k-\frac{\lambda_1 L+\lambda_2\mu}{2(L-\mu)}\\ \gamma_k-\frac{\lambda_1 L+\lambda_2\mu}{2(L-\mu)} & -\gamma_k^2+\frac{\lambda_1+\lambda_2}{2(L-\mu)}\end{bmatrix},\end{aligned} \tag{C.11}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

where we used the fact there is no duality gap, as one can show the existence of a Slater point [13]. One such Slater point can be obtained by applying gradient descent on the function $f(x) = \frac{1}{2}x^\top \mathrm{diag}(L, \mu)x$ (i.e., $d = 2$) with $x_k = (1, 1)$. A formal statement is provided in [210, Theorem 6].

Theorem C.1 is now a direct consequence of the dual reformulation (C.11), as provided by the following proof.

Proof of Theorem C.1. (Sufficiency, $\Leftarrow$) If there exists a feasible point $(\tau, \lambda_1, \lambda_2)$ for (C.2), weak duality implies that it is an upper bound on $\tau_\star$ by construction.

(Necessity, $\Rightarrow$) For any τ such that there exists no $\lambda_1, \lambda_2 \geq 0$ for which $(\tau, \lambda_1, \lambda_2)$ is feasible for (C.2), it follows that $\tau \leq \tau_\star$, and strong duality implies that there exists a problem instance ($f \in \mathcal{F}_{\mu,L}$, $d \in \mathbb{N}$, and $x_k \in \mathbb{R}^d$) on which $\|x_{k+1} - x_\star\|_2^2 = \tau_\star \|x_k - x_\star\|_2^2 \geq \tau \|x_k - x_\star\|_2^2$. ■

Remark C.2. Following similar lines as those of this section, one can verify other types of inequalities, beyond (C.1), simply by changing the objective in (C.5). This allows obtaining the statement from Theorem C.2 and Theorem C.3.

Remark C.3. Finding analytical solutions to such LMIs (parametrized by the algorithm and problem parameters) might be challenging. For gradient descent, the solution is provided in e.g., [98, Section 4.4] and [265, Theorem 3.1]. For more complicated cases, one can rely on numerical inspiration for finding analytical solutions (or upper bounds on it).

Remark C.4. It is possible to obtain "weaker" LMIs based on other sets of inequalities (which are necessary but not sufficient for interpolation). Those LMIs are then only sufficient for finding worst-case guarantees. Those alternate LMIs might enjoy simpler analytical solutions, but this comes at the cost of loosing a priori tightness guarantees. This is in general not a problem if the worst-case guarantee is satisfying, but the subtle consequence is that those LMIs might then fail to provide a satisfying guarantee even when there exists one.

Full text available at: http://dx.doi.org/10.1561/2400000036

C.3.2 Potential Function for Gradient Descent

For formulating the LMI for verifying potential functions as those of Theorem 4.2 and Theorem 4.10, one essentially has to follow the same steps as in the previous section. The strongly convex version is a bit heavy and is provided below. In short, verifying that

$$\phi_k \triangleq A_k(f(x_k) - f_\star) + \tfrac{L+\mu A_k}{2}\|x_k - x_\star\|_2^2$$

is a potential function, that is, $\phi_{k+1} \leq \phi_k$ (for all $f \in \mathcal{F}_{\mu,L}$, $d \in \mathbb{N}$, and $x_k \in \mathbb{R}^d$), amount to verify that

$$0 \geq \max\{\phi_{k+1} - \phi_k : d \in \mathbb{N},\ f \in \mathcal{F}_{\mu,L},\ x_k, x_{k+1}, x_\star \in \mathbb{R}^d, \\ x_{k+1} = x_k - \gamma_k \nabla f(x_k),\ \text{and}\ \nabla f(x_\star) = 0\},$$

where the maximum is taken over d, f, x_k, x_{k+1} and $x_\star$. This problem can be reformulated as in Section C.3 using the same technique with more samples. More precisely, this formulation requires sampling the function f at three points (instead of two): $x_\star$, x_k, and x_{k+1}, and hence 6 dual variables are required (because 6 inequalities of the form (C.7) are used for describing the sampled version of the function f). The formal statement is provided by the following theorem, without a proof.

Theorem C.5. Let $A_{k+1}, A_k \geq 0$ and $\gamma_k \in \mathbb{R}$. The inequality

$$A_{k+1}(f(x_{k+1}) - f_\star) + \tfrac{L+\mu A_{k+1}}{2}\|x_{k+1} - x_\star\|_2^2 \\ \leq A_k(f(x_k) - f_\star) + \tfrac{L+\mu A_k}{2}\|x_k - x_\star\|_2^2$$

holds for all $d \in \mathbb{N}$, all $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, all $x_k, x_{k+1}, x_\star \in \mathbb{R}^d$ (such that $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$ and $\nabla f(x_\star) = 0$) if and only if

$$\exists \lambda_1, \lambda_2, \ldots, \lambda_6 \geq 0 : \begin{cases} 0 = A_k + \lambda_1 + \lambda_2 - \lambda_4 - \lambda_6 \\ 0 = -A_{k+1} - \lambda_2 + \lambda_3 + \lambda_4 - \lambda_5 \\ 0 \preceq \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} \\ S_{1,2} & S_{2,2} & S_{2,3} \\ S_{1,3} & S_{2,3} & S_{3,3} \end{bmatrix}, \end{cases}$$

Full text available at: http://dx.doi.org/10.1561/2400000036

with

$$S_{1,1} = \tfrac{1}{2}\mu\left(A_k - A_{k+1} + \tfrac{L(\lambda_1+\lambda_3+\lambda_5+\lambda_6)}{L-\mu}\right)$$
$$S_{1,2} = -\tfrac{\gamma_k(\mu A_{k+1}(\mu-L)+L(\mu(\lambda_3+\lambda_5+1)-L))+\lambda_6\mu+\lambda_1 L}{2(L-\mu)}$$
$$S_{1,3} = -\tfrac{\lambda_5\mu+\lambda_3 L}{2(L-\mu)}$$
$$S_{2,2} = \tfrac{\gamma_k^2(\mu A_{k+1}(\mu-L)+L(\mu(\lambda_2+\lambda_3+\lambda_4+\lambda_5+1)-L))-2\gamma_k(\lambda_4\mu+\lambda_2 L)+\lambda_1+\lambda_2+\lambda_4+\lambda_6}{2(L-\mu)}$$
$$S_{2,3} = \tfrac{\gamma_k(\mu(\lambda_2+\lambda_5)+L(\lambda_3+\lambda_4))-\lambda_2-\lambda_4}{2(L-\mu)}$$
$$S_{3,3} = \tfrac{\lambda_2+\lambda_3+\lambda_4+\lambda_5}{2(L-\mu)}.$$

Note again that a notebook is provided in Section C.5 for obtaining and verifying this LMI formulation via symbolic computations.

C.4 Accelerated Gradient Descent via Linear Matrix Inequalities

We provide the main ideas for formulating the LMI for verifying potential functions as those of Theorem 4.10 and Theorem 4.12. In short, verifying that

$$\phi_k \triangleq A_k(f(x_k) - f_\star) + \tfrac{L+\mu A_k}{2}\|z_k - x_\star\|_2^2$$

is a potential function, that is, $\phi_{k+1} \leq \phi_k$ (for all $f \in \mathcal{F}_{\mu,L}$, $d \in \mathbb{N}$, and $x_k, z_k, x_\star \in \mathbb{R}^d$, $\nabla f(x_\star) = 0$), amounts to verify that

$$0 \geq \max\big\{\phi_{k+1} - \phi_k : d \in \mathbb{N},\ f \in \mathcal{F}_{\mu,L},\ z_k, x_k, x_\star \in \mathbb{R}^d,\ \nabla f(x_\star) = 0,\\ \text{and } y_k, x_{k+1}, z_{k+1} \in \mathbb{R}^d \text{ generated by (4.17)}\big\},$$

where the maximum is taken over d, f, the iterates, as well as $x_\star$. This problem can be cast as a SDP using the same ideas as in Section C.3 with more samples, again. More precisely, this formulation requires sampling the function f at four points: $x_\star$, x_k, x_{k+1}, and y_k. The case $\mu = 0$ is covered by Theorem C.3.

C.5 Notes and References

General frameworks. The whole idea of using semidefinite programming for analyzing first-order methods dates back to [3] (more details

Full text available at: http://dx.doi.org/10.1561/2400000036

and examples in [83], [209]). The principled approach to worst-case analysis using performance estimation problems with interpolation/extension arguments was proposed in in [210], and generalized to more problem setups in [201]. The integral quadratic approach to first-order methods was proposed in [98], specifically for studying linearly converging methods (focus on strong convexity and related notions). Those two related methodologies were then further extended and linked in different setup [84], [161], [175], [206], [211], [213], [266]–[269]. Among those developments, some works performed analyses via "weaker" LMIs, based on other sets of inequalities which are necessary but not sufficient for interpolation; see, e.g., [270]. The advantage of this approach is that it is often simpler to obtain analytical solutions to some of those LMIs, at the cost of loosing tightness guarantees (which might not be a problem when the guarantee is satisfying). This is in general the case for IQC-based works. In those cases, non-tightness is usually coupled with the search for a Lyapunov function. In general, it is possible to simultaneously look for a tight guarantee and a Lyapunov/potential function, see e.g., [84], [267]. A simplified approach to performance estimation problems was implemented in the performance estimation toolbox [214, PESTO].

Designing methods using semidefinite programming. The optimized gradient method (OGM) was apparently the first method obtained by optimizing its worst-case using SDPs/LMIs. It was obtained as a solution to a convex optimization problem by Drori and Teboulle [3], which was later solved analytically by Kim and Fessler [4]. The same method was obtained through an analogy with the conjugate gradient method [203], which might serve as a strategy for designing method in various setups. Optimized methods can be developed for other criteria and setups as well. As an example, optimized methods for gradient norms $\|\nabla f(x_N)\|_2^2$ are studied in Kim and Fessler [166], [202], in the smooth convex setting. See also Section 4.6.1 and Section 4.6.2; in particular, the *Triple Momentum Method* (TMM) [95] was designed as a time-independent optimized gradient method, through Lyapunov arguments (and IQCs). See also [88], [100], [101], [271] for different ways of recovering the TMM. Optimized methods were also developed in other

Full text available at: http://dx.doi.org/10.1561/2400000036

setups, such as fixed-point iteration [206] and monotone inclusions [207] (which turned out to be a particular case of [206]).

Specific methods. The SDP/LMI approaches were taken further for studying first-order methods in a few different contexts. It was originally used for studying gradient-type methods (see, e.g., [3], [83], [98], [210]) and accelerated/fast gradient-type methods (see, e.g., [3], [83], [95], [98], [99], [175], [201], [210], [213], [272]) for convex minimization. It was used later for analyzing, among others, nonsmooth setups [203], [209], stochastic [84], [266], [268], [273], coordinate-descent [84], [274], nonconvex setups [275], [276], proximal methods [166], [200], [201], [208], splitting methods [265], [277], [278], monotone inclusions and variational inequalities [278]–[281], fixed-point iterations [206], and distributed/decentralized optimization [282], [283].

Obtaining and solving the LMIs. For solving the LMIs, standard numerical semidefinite optimization packages can be used, see, e.g., [**Yalmip**, **Article:Sedumi**, **Article:Mosek**, **toh2012implementation**]. For obtaining and verifying analytical solutions, symbolic computing might also be a great asset. For the purpose of reproducibility, we provide notebooks for obtaining the LMI formulations of this section symbolically, and for solving them numerically, at https://github.com/AdrienTaylor/AccelerationMonograph.

Full text available at: http://dx.doi.org/10.1561/2400000036

Acknowledgements

The authors would like to warmly thank Raphaël Berthier, Mathieu Barré, Aymeric Dieuleveut, Fabian Pedregosa and Baptiste Goujaud for comments on early versions of this manuscript; for spotting a few typos; and for discussions and developments related to Section 2, Section 4, and Section 5. We are also greatly indebted to Lenaïc Chizat, Laurent Condat, Jelena Diakonikolas, Alexander Gasnikov, Shuvomoy Das Gupta, Pontus Giselsson, Cristóbal Guzmán, Julien Mairal, and Irène Waldspurger for spotting a few typos and inconsistencies in the first version of the manuscript.

We further want to thank Francis Bach, Sébastien Bubeck, Radu-Alexandru Dragomir, Yoel Drori, Hadrien Hendrikx, Reza Babanezhad, Claude Brezinski, Pavel Dvurechensky, Hervé Le Ferrand, Georges Lan, Adam Ouorou, Michela Redivo-Zaglia, Simon Lacoste-Julien, Vincent Roulet, and Ernest Ryu for fruitful discussions and pointers, which largely simplified the writing and revision process of this manuscript.

AA is also extremely grateful to the French ministry of education and école Etienne Marcel for keeping school mostly open during the 2020-2021 pandemic.

AA is at the Département d'informatique de l'ENS, École normale supérieure, UMR CNRS 8548, PSL Research University, 75005 Paris, France and INRIA. AA would like to acknowledge support from the ML and Optimisation joint research initiative with the funds AXA

Full text available at: http://dx.doi.org/10.1561/2400000036

pour la Recherche and Kamet Ventures, a Google focused award, as well as funding from the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). AT is at INRIA and the Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, 75005 Paris, France. AT acknowledges support from the European Research Council (ERC grant SEQUOIA 724063).

Full text available at: http://dx.doi.org/10.1561/2400000036

References

[1] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, 2009, pp. 183–202.

[3] Y. Drori and M. Teboulle, "Performance of first-order methods for smooth convex minimization: A novel approach," *Mathematical Programming*, vol. 145, no. 1-2, 2014, pp. 451–482.

[4] D. Kim and J. A. Fessler, "Optimized first-order methods for smooth convex minimization," *Mathematical Programming*, vol. 159, no. 1-2, 2016, pp. 81–107.

[5] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Springer, 2003.

[6] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, 2008. [Online]. Available: http://www.mit.edu/~dimitrib/PTseng/papers.html.

[7] A. C. Aitken, "On Bernoulli's numerical solution of algebraic equations," *Proceedings of the Royal Society of Edinburgh*, vol. 46, 1927, pp. 289–305.

Full text available at: http://dx.doi.org/10.1561/2400000036

[8] P. Wynn, "On a device for computing the $e_m(S_n)$ transformation," *Mathematical Tables and Other Aids to Computation*, vol. 10, no. 54, 1956, pp. 91–96.

[9] E. J. Anderson and P. Nash, *Linear programming in infinite-dimensional spaces*. Chichester: Wiley, 1987.

[10] O. Güler, "New proximal point algorithms for convex minimization," *SIAM Journal on Optimization*, vol. 2, no. 4, 1992, pp. 649–664.

[11] R. D. Monteiro and B. F. Svaiter, "An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods," *SIAM Journal on Optimization*, vol. 23, no. 2, 2013, pp. 1092–1125.

[12] H. Lin, J. Mairal, and Z. Harchaoui, "A universal catalyst for first-order optimization," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[13] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[14] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[15] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[16] R. T. Rockafellar, *Convex Analysis*. Princeton.: Princeton University Press., 1970.

[17] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317. Springer Science & Business Media, 2009.

[18] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*, vol. 305. Springer science & business media, 2013.

[19] A. S. Nemirovsky and B. T. Polyak, "Iterative methods for solving linear ill-posed problems under precise information.," *ENG. CYBER.*, no. 4, 1984, pp. 50–56.

[20] D. A. Flanders and G. Shortley, "Numerical determination of fundamental modes," *Journal of Applied Physics*, vol. 21, no. 12, 1950, pp. 1326–1332.

Full text available at: http://dx.doi.org/10.1561/2400000036

[21] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*. CRC press, 2002.

[22] G. H. Golub and R. S. Varga, "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order richardson iterative methods," *Numerische Mathematik*, vol. 3, no. 1, 1961, pp. 147–156.

[23] E. Süli and D. F. Mayers, *An introduction to numerical analysis*. Cambridge university press, 2003.

[24] M. H. Gutknecht and S. Röllin, "The chebyshev iteration revisited," *Parallel Computing*, vol. 28, no. 2, 2002, pp. 263–283.

[25] G. H. Golub and R. S. Varga, "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order richardson iterative methods," *Numerische Mathematik*, vol. 3, no. 1, 1961, pp. 157–168.

[26] A. S. Nemirovsky, *Information-based complexity of convex programming*, Lecture notes, 1994.

[27] A. S. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization*. 1983.

[28] A. S. Nemirovsky, "Information-based complexity of linear operator equations," *Journal of Complexity*, vol. 8, no. 2, 1992, pp. 153–175.

[29] E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, 1952, pp. 409–435.

[30] T. A. Straeter, "On the extension of the davidon-broyden class of rank one, quasi-newton minimization methods to an infinite dimensional hilbert space with applications to optimal control problems," Tech. Rep., 1971.

[31] Y. Saad and M. H. Schultz, "Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on scientific and statistical computing*, vol. 7, no. 3, 1986, pp. 856–869.

[32] B. Fischer, "Polynomial based iteration methods for symmetric linear systems," 1996.

Full text available at: http://dx.doi.org/10.1561/2400000036

[33] F. Pedregosa and D. Scieur, "Acceleration through spectral density estimation," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[34] J. Lacotte and M. Pilanci, "Optimal randomized first-order methods for least-squares problems," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[35] D. Scieur and F. Pedregosa, "Universal asymptotic optimality of polyak momentum," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[36] D. G. Anderson, "Iterative procedures for nonlinear integral equations," *Journal of the ACM (JACM)*, vol. 12, no. 4, 1965, pp. 547–560.

[37] A. Sidi, W. F. Ford, and D. A. Smith, "Acceleration of convergence of vector sequences," *SIAM Journal on Numerical Analysis*, vol. 23, no. 1, 1986, pp. 178–196.

[38] C. C. Paige and M. A. Saunders, "Solution of sparse indefinite systems of linear equations," *SIAM journal on numerical analysis*, vol. 12, no. 4, 1975, pp. 617–629.

[39] H. F. Walker and P. Ni, "Anderson acceleration for fixed-point iterations," *SIAM Journal on Numerical Analysis*, vol. 49, no. 4, 2011, pp. 1715–1735.

[40] A. Sidi, "Efficient implementation of minimal polynomial and reduced rank extrapolation methods," *Journal of Computational and Applied Mathematics*, vol. 36, no. 3, 1991, pp. 305–337.

[41] D. Scieur, A. d'Aspremont, and F. Bach, "Regularized nonlinear acceleration," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[42] E. E. Tyrtyshnikov, "How bad are Hankel matrices?" *Numerische Mathematik*, vol. 67, no. 2, 1994, pp. 261–269.

[43] D. Scieur, E. Oyallon, A. d'Aspremont, and F. Bach, "Online regularized nonlinear acceleration," *preprint arXiv:1805.09639*, 2018.

[44] M. Barré, A. Taylor, and A. d'Aspremont, "Convergence of constrained Anderson acceleration," *preprint arXiv:2010.15482*, 2020.

Full text available at: http://dx.doi.org/10.1561/2400000036

[45] D. Scieur, F. Bach, and A. d'Aspremont, "Nonlinear acceleration of stochastic algorithms," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[46] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, 2017, pp. 83–112.

[47] 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.

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

[49] V. Mai and M. Johansson, "Anderson acceleration of proximal gradient methods," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[50] F. H. Clarke, *Optimization and nonsmooth analysis*, vol. 5. SIAM, 1990.

[51] R. Mifflin, "Semismooth and semiconvex functions in constrained optimization," *SIAM Journal on Control and Optimization*, vol. 15, no. 6, 1977, pp. 959–972.

[52] L. Qi and J. Sun, "A nonsmooth version of newton's method," *Mathematical programming*, vol. 58, no. 1-3, 1993, pp. 353–367.

[53] D. A. Smith, W. F. Ford, and A. Sidi, "Extrapolation methods for vector sequences," *SIAM review*, vol. 29, no. 2, 1987, pp. 199–233.

[54] K. Jbilou and H. Sadok, "Some results about vector extrapolation methods and related fixed-point iterations," *Journal of Computational and Applied Mathematics*, vol. 36, no. 3, 1991, pp. 385–398.

[55] C. Brezinski and M. R. Zaglia, *Extrapolation methods: theory and practice*. Elsevier, 1991.

[56] K. Jbilou and H. Sadok, "Analysis of some vector extrapolation methods for solving systems of linear equations," *Numerische Mathematik*, vol. 70, no. 1, 1995, pp. 73–89.

Full text available at: http://dx.doi.org/10.1561/2400000036

[57] K. Jbilou and H. Sadok, "Vector extrapolation methods. applications and numerical comparison," *Journal of Computational and Applied Mathematics*, vol. 122, no. 1-2, 2000, pp. 149–165.

[58] C. Brezinski, "Convergence acceleration during the 20th century," *Numerical Analysis: Historical Developments in the 20th Century*, 2001, p. 113.

[59] C. Brezinski and M. Redivo–Zaglia, "The genesis and early developments of aitken's process, shanks' transformation, the ε–algorithm, and related fixed point methods," *Numerical Algorithms*, vol. 80, no. 1, 2019, pp. 11–133.

[60] E. Gekeler, "On the solution of systems of equations by the epsilon algorithm of wynn," *Mathematics of Computation*, vol. 26, no. 118, 1972, pp. 427–436.

[61] C. Brezinski, "Sur un algorithme de résolution des systèmes non linéaires," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 272, no. A, 1971, pp. 145–148.

[62] C. Brezinski, "Application de l'ε-algorithme à la résolution des systèmes non linéaires," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 271, no. A, 1970, pp. 1174–1177.

[63] A. Toth and C. Kelley, "Convergence analysis for anderson acceleration," *SIAM Journal on Numerical Analysis*, vol. 53, no. 2, 2015, pp. 805–819.

[64] H.-R. Fang and Y. Saad, "Two classes of multisecant methods for nonlinear acceleration," *Numerical Linear Algebra with Applications*, vol. 16, no. 3, 2009, pp. 197–221.

[65] S. Cabay and L. Jackson, "A polynomial extrapolation method for finding limits and antilimits of vector sequences," *SIAM Journal on Numerical Analysis*, vol. 13, no. 5, 1976, pp. 734–752.

[66] M. Mešina, "Convergence acceleration for the iterative solution of the equations $X = AX + f$," *Computer Methods in Applied Mechanics and Engineering*, vol. 10, no. 2, 1977, pp. 165–173.

[67] R. Eddy, "Extrapolating to the limit of a vector sequence," in *Information linkage between applied mathematics and industry*, Elsevier, 1979, pp. 387–396.

Full text available at: http://dx.doi.org/10.1561/2400000036

[68] A. Sidi, "Extrapolation vs. projection methods for linear systems of equations," *Journal of Computational and Applied Mathematics*, vol. 22, no. 1, 1988, pp. 71–88.

[69] W. F. Ford and A. Sidi, "Recursive algorithms for vector extrapolation methods," *Applied numerical mathematics*, vol. 4, no. 6, 1988, pp. 477–489.

[70] A. Sidi and Y. Shapira, "Upper bounds for convergence rates of acceleration methods with initial iterations," *Numerical Algorithms*, vol. 18, no. 2, 1998, pp. 113–132.

[71] A. Sidi, "Vector extrapolation methods with applications to solution of large systems of equations and to pagerank computations," *Computers & Mathematics with Applications*, vol. 56, no. 1, 2008, pp. 1–24.

[72] A. Sidi, "Minimal polynomial and reduced rank extrapolation methods are related," *Advances in Computational Mathematics*, vol. 43, no. 1, 2017, pp. 151–170.

[73] A. Sidi, *Vector extrapolation methods with applications*. SIAM, 2017.

[74] C. Brezinski, M. Redivo-Zaglia, and Y. Saad, "Shanks sequence transformations and anderson acceleration," *SIAM Review*, vol. 60, no. 3, 2018, pp. 646–669.

[75] C. Brezinski, S. Cipolla, M. Redivo-Zaglia, and Y. Saad, "Shanks and Anderson-type acceleration techniques for systems of nonlinear equations," *preprint arXiv:2007.05716*, 2020.

[76] A. Sidi, "Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms," *SIAM Journal on Numerical Analysis*, vol. 23, no. 1, 1986, pp. 197–209.

[77] A. Sidi and J. Bridger, "Convergence and stability analyses for some vector extrapolation methods in the presence of defective iteration matrices," *Journal of Computational and Applied Mathematics*, vol. 22, no. 1, 1988, pp. 35–61.

[78] A. Sidi, "A convergence study for reduced rank extrapolation on nonlinear systems," *Numerical Algorithms*, 2019, pp. 1–26.

[79] C. Brezinski, "Généralisations de la transformation de shanks, de la table de padé et de l'ε-algorithme," *Calcolo*, vol. 12, no. 4, 1975, pp. 317–360.

Full text available at: http://dx.doi.org/10.1561/2400000036

[80] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, 2013, pp. 125–161.

[81] A. Cauchy, "Méthode générale pour la résolution des systemes d'équations simultanées," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 25, no. 1847, 1847, pp. 536–538.

[82] N. Bansal and A. Gupta, "Potential-function proofs for gradient methods," *Theory of Computing*, vol. 15, no. 1, 2019, pp. 1–32.

[83] Y. Drori, "Contributions to the complexity analysis of optimization algorithms," Ph.D. dissertation, Tel-Aviv University, 2014.

[84] A. Taylor and F. Bach, "Stochastic first-order methods: Non-asymptotic and computer-aided analyses via potential functions," in *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 2019.

[85] D. Kim and J. A. Fessler, "On the convergence analysis of the optimized gradient method," *Journal of Optimization Theory and Applications*, vol. 172, no. 1, 2017, pp. 187–205.

[86] C. Guzmán and A. S. Nemirovsky, "On lower complexity bounds for large-scale smooth convex optimization," *Journal of Complexity*, vol. 31, no. 1, 2015, pp. 1–14.

[87] Y. Drori, "The exact information-based complexity of smooth convex minimization," *Journal of Complexity*, vol. 39, 2017, pp. 1–16.

[88] Y. Drori and A. Taylor, "On the oracle complexity of smooth strongly convex minimization," *Journal of Complexity*, 2021.

[89] A. S. Nemirovsky, "On optimality of krylov's information when solving linear operator equations," *Journal of Complexity*, vol. 7, no. 2, 1991, pp. 121–130.

[90] Y. Drori, "On the properties of convex functions over open sets," *preprint arXiv:1812.02419*, 2018.

[91] M. Baes, *Estimate sequence methods: Extensions and approximations*, 2009. [Online]. Available: http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf.

[92] A. C. Wilson, B. Recht, and M. I. Jordan, "A lyapunov analysis of accelerated methods in optimization," *The Journal of Machine Learning Research (JMLR)*, vol. 22, no. 113, 2021, pp. 1–34.

Full text available at: http://dx.doi.org/10.1561/2400000036

[93] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, "Understanding the acceleration phenomenon via high-resolution differential equations," *Mathematical Programming*, 2021, pp. 1–70.

[94] A. Taylor and Y. Drori, "An optimal gradient method for smooth strongly convex minimization," *preprint arXiv:2101.09741*, 2021.

[95] B. Van Scoy, R. A. Freeman, and K. M. Lynch, "The fastest known globally convergent first-order method for minimizing strongly convex functions," *IEEE Control Systems Letters*, vol. 2, no. 1, 2017, pp. 49–54.

[96] S. Bubeck, Y. T. Lee, and M. Singh, "A geometric alternative to nesterov's accelerated gradient descent," *preprint arXiv:1506.08187*, 2015.

[97] D. Drusvyatskiy, M. Fazel, and S. Roy, "An optimal first order method based on optimal quadratic averaging," *SIAM Journal on Optimization*, vol. 28, no. 1, 2018, pp. 251–271.

[98] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, 2016, pp. 57–95.

[99] S. Cyrus, B. Hu, B. Van Scoy, and L. Lessard, "A robust accelerated optimization algorithm for strongly convex functions," in *Proceedings of the 2018 American Control Conference (ACC)*, 2018.

[100] K. Zhou, A. M.-C. So, and J. Cheng, "Boosting first-order methods by shifting objective: New schemes with faster worst case rates," in *Advances in Neural Information Processing Systems (NeuRIPS)*, 2020.

[101] L. Lessard and P. Seiler, "Direct synthesis of iterative algorithms with bounds on achievable worst-case convergence rate," in *Proceedings of the 2020 American Control Conference (ACC)*, 2020.

[102] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, 2015, pp. 231–357.

[103] S. Chen, S. Ma, and W. Liu, "Geometric descent method for convex composite minimization," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Full text available at: http://dx.doi.org/10.1561/2400000036

[104] S. Karimi and S. Vavasis, "A single potential governing convergence of conjugate gradient, accelerated gradient and geometric descent," *preprint arXiv:1712.09498*, 2017.

[105] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Transactions of the American mathematical Society*, vol. 82, no. 2, 1956, pp. 421–439.

[106] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 9, no. R2, 1975, pp. 41–76.

[107] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, 1979, pp. 964–979.

[108] G. Peyré, "The numerical tours of signal processing-advanced computational signal and image processing," *IEEE Computing in Science and Engineering*, vol. 13, no. 4, 2011, pp. 94–97.

[109] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, 2014, pp. 127–239.

[110] A. Chambolle and T. Pock, "An introduction to continuous optimization for imaging," *Acta Numerica*, vol. 25, 2016, pp. 161–319.

[111] J. A. Fessler, "Optimization methods for magnetic resonance image reconstruction: Key models and optimization algorithms," *IEEE Signal Processing Magazine*, vol. 37, no. 1, 2020, pp. 33–40.

[112] A. Goldstein, "Cauchy's method of minimization," *Numerische Mathematik*, vol. 4, no. 1, 1962, pp. 146–150.

[113] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives," *Pacific Journal of mathematics*, vol. 16, no. 1, 1966, pp. 1–3.

[114] S. R. Becker, E. J. Candès, and M. C. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical Programming Computation*, vol. 3, no. 3, 2011, pp. 165–218.

Full text available at: http://dx.doi.org/10.1561/2400000036

[115] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Foundations of computational mathematics*, vol. 15, no. 3, 2015, pp. 715–732.

[116] V. Roulet and A. d'Aspremont, "Sharpness, restart and acceleration," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[117] A. Ben-Tal and A. S. Nemirovsky, *Lectures on modern convex optimization : analysis, algorithms, and engineering applications*, ser. MPS-SIAM series on optimization. SIAM, 2001.

[118] A. Juditsky and A. S. Nemirovsky, "First order methods for nonsmooth convex large-scale optimization, i: General purpose methods," *Optimization for Machine Learning*, vol. 30, no. 9, 2011, pp. 121–148.

[119] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, 2003, pp. 167–175.

[120] Y. Nesterov and V. Shikhman, "Quasi-monotone subgradient methods for nonsmooth convex minimization," *Journal of Optimization Theory and Applications*, vol. 165, no. 3, 2015, pp. 917–940.

[121] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions on Image Processing*, vol. 18, no. 11, 2009, pp. 2419–2434.

[122] E. K. Ryu and S. Boyd, "Primer on monotone operator methods," *Appl. Comput. Math*, vol. 15, no. 1, 2016, pp. 3–43.

[123] G. Chierchia, E. Chouzenoux, P. L. Combettes, and J.-C. Pesquet, *The proximity operator repository. user's guide*, 2020. [Online]. Available: http://proximity-operator.net/download/guide.pdf.

[124] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in hilbert space," *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2, 1979, pp. 383–390.

[125] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, "Proximal splitting algorithms: A tour of recent advances, with new twists," *preprint arXiv:1912.00137*, 2019.

Full text available at: http://dx.doi.org/10.1561/2400000036

[126] K. Scheinberg, D. Goldfarb, and X. Bai, "Fast first-order methods for composite convex optimization with backtracking," *Foundations of Computational Mathematics*, vol. 14, no. 3, 2014, pp. 389–417.

[127] L. Calatroni and A. Chambolle, "Backtracking strategies for accelerated descent methods with smooth composite objectives," *SIAM Journal on Optimization*, vol. 29, no. 3, 2019, pp. 1772–1798.

[128] M. I. Florea and S. A. Vorobyov, "An accelerated composite gradient method for large-scale composite objective problems," *IEEE Transactions on Signal Processing*, vol. 67, no. 2, 2018, pp. 444–459.

[129] M. I. Florea and S. A. Vorobyov, "A generalized accelerated composite gradient method: Uniting nesterov's fast gradient method and fista," *IEEE Transactions on Signal Processing*, 2020.

[130] A. Auslender and M. Teboulle, "Interior gradient and proximal methods for convex and conic optimization," *SIAM Journal on Optimization*, vol. 16, no. 3, 2006, pp. 697–725.

[131] A. V. Gasnikov and Y. Nesterov, "Universal method for stochastic composite optimization problems," *Computational Mathematics and Mathematical Physics*, vol. 58, no. 1, 2018, pp. 48–64.

[132] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," *Mathematical Programming*, vol. 175, no. 1-2, 2019, pp. 69–107.

[133] O. Hinder, A. Sidford, and N. Sohoni, "Near-optimal methods for minimizing star-convex functions and beyond," in *Proceedings of the 33rd Conference on Learning Theory (COLT)*, 2020.

[134] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, "From error bounds to the complexity of first-order descent methods for convex functions," *Mathematical Programming*, vol. 165, no. 2, 2017, pp. 471–507.

[135] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, 2005, pp. 127–152.

Full text available at: http://dx.doi.org/10.1561/2400000036

[136] G. Lan, Z. Lu, and R. D. Monteiro, "Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming," *Mathematical Programming*, vol. 126, no. 1, 2011, pp. 1–29.

[137] J. Diakonikolas and C. Guzmán, "Complementary composite minimization, small gradients in general norms, and applications to regression problems," *preprint arXiv:2101.11041*, 2021.

[138] A. Juditsky, G. Lan, A. S. Nemirovsky, and A. Shapiro, "Stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, 2009, pp. 1574–1609.

[139] A. d'Aspremont, C. Guzman, and M. Jaggi, "Optimal affine-invariant smooth minimization algorithms," *SIAM Journal on Optimization*, vol. 28, no. 3, 2018, pp. 2384–2405.

[140] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

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

[142] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont, "Integration methods and optimization algorithms," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[143] W. Krichene, A. Bayen, and P. L. Bartlett, "Accelerated mirror descent in continuous and discrete time," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[144] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," in *Proceedings of the National Academy of Sciences*, 2016.

[145] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, "Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity," *Mathematical Programming*, vol. 168, no. 1, 2018, pp. 123–175.

Full text available at: http://dx.doi.org/10.1561/2400000036

[146] B. Sun, J. George, and S. Kia, "High-resolution modeling of the fastest first-order optimization method for strongly convex functions," in *Proceedings of the 59th Conference on Decision and Control (CDC)*, 2020.

[147] B. Shi, S. S. Du, W. Su, and M. I. Jordan, "Acceleration via symplectic discretization of high-resolution differential equations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[148] J. Diakonikolas and L. Orecchia, "The approximate duality gap technique: A unified theory of first-order methods," *SIAM Journal on Optimization*, vol. 29, no. 1, 2019, pp. 660–689.

[149] J. W. Siegel, "Accelerated first-order methods: Differential equations and lyapunov functions," *preprint arXiv:1903.05671*, 2019.

[150] J. M. Sanz Serna and K. C. Zygalakis, "The connections between Lyapunov functions for some optimization algorithms and differential equations," *SIAM Journal on Numerical Analysis*, vol. 59, no. 3, 2021, pp. 1542–1565.

[151] M. Even, R. Berthier, F. Bach, N. Flammarion, P. Gaillard, H. Hendrikx, L. Massoulié, and A. Taylor, "A continuized view on nesterov acceleration for stochastic gradient descent and randomized gossip," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[152] O. Devolder, "Stochastic first order methods in smooth convex optimization," CORE discussion paper, Tech. Rep., 2011.

[153] A. Kulunchakov and J. Mairal, "Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise," *The Journal of Machine Learning Research (JMLR)*, vol. 21, no. 155, 2020, pp. 1–52.

[154] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical programming Series B*, vol. 120, no. 1, 2009, pp. 221–259.

[155] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *The Journal of Machine Learning Research (JMLR)*, vol. 11, 2010, pp. 2543–2596.

Full text available at: http://dx.doi.org/10.1561/2400000036

[156] A. Juditsky and Y. Nesterov, "Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization," *Stochastic Systems*, vol. 4, no. 1, 2014, pp. 44–80.

[157] A. Juditsky and A. S. Nemirovsky, "First order methods for nonsmooth convex large-scale optimization, ii: Utilizing problems structure," *Optimization for Machine Learning*, vol. 30, no. 9, 2011, pp. 149–183.

[158] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, 2016, pp. 330–348.

[159] M. Teboulle, "A simplified view of first order methods for optimization," *Mathematical Programming*, vol. 170, no. 1, 2018, pp. 67–96.

[160] H. Lu, R. M. Freund, and Y. Nesterov, "Relatively smooth convex optimization by first-order methods, and applications," *SIAM Journal on Optimization*, vol. 28, no. 1, 2018, pp. 333–354.

[161] R.-A. Dragomir, A. B. Taylor, A. d'Aspremont, and J. Bolte, "Optimal complexity and certification of Bregman first-order methods," *Mathematical Programming*, 2021, pp. 1–43.

[162] F. Hanzely, P. Richtarik, and L. Xiao, "Accelerated bregman proximal gradient methods for relatively smooth convex optimization," *Computational Optimization and Applications*, vol. 79, no. 2, 2021, pp. 405–440.

[163] D. H. Gutman and J. F. Peña, "A unified framework for bregman proximal methods: Subgradient, gradient, and accelerated gradient schemes," *preprint arXiv:1812.10198*, 2018.

[164] A. S. Nemirovsky and D. B. Yudin, "Problem complexity and method efficiency in optimization.," *Willey-Interscience, New York*, 1983.

[165] Y. Nesterov, "How to make the gradients small," *Optima. Mathematical Optimization Society Newsletter*, no. 88, 2012, pp. 10–11.

[166] D. Kim and J. A. Fessler, "Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions," *Journal of Optimization Theory and Applications*, 2020.

Full text available at: http://dx.doi.org/10.1561/2400000036

[167] J. Diakonikolas and P. Wang, "Potential function-based framework for making the gradients small in convex and min-max optimization," *preprint arXiv:2101.12101*, 2021.

[168] J. Lee, C. Park, and E. K. Ryu, "A geometric structure of acceleration and its role in making gradients small fast," *preprint arXiv:2106.10439*, 2021.

[169] Y. Malitsky and K. Mishchenko, "Adaptive gradient descent without descent," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[170] A. d'Aspremont, "Smooth optimization with approximate gradient," *SIAM Journal on Optimization*, vol. 19, no. 3, 2008, pp. 1171–1183.

[171] 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 (NIPS)*, 2011.

[172] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, vol. 146, no. 1-2, 2014, pp. 37–75.

[173] O. Devolder, "Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization," Ph.D. dissertation, 2013.

[174] O. Devolder, F. Glineur, and Y. Nesterov, "Intermediate gradient methods for smooth convex problems with inexact oracle," CORE discussion paper, Tech. Rep., 2013.

[175] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "Robust accelerated gradient methods for smooth strongly convex functions," *SIAM Journal on Optimization*, vol. 30, no. 1, 2020, pp. 717–751.

[176] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, "Accelerated and inexact forward-backward algorithms," *SIAM Journal on Optimization*, vol. 23, no. 3, 2013, pp. 1607–1633.

[177] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Full text available at: http://dx.doi.org/10.1561/2400000036

[178] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, 1951, pp. 400–407.

[179] G. Lan, "Efficient methods for stochastic composite optimization," School of Industrial and Systems Engineering, Georgia Institute of Technology, Tech. Rep., 2008. [Online]. Available: http://www.optimization-online.org/DB_HTML/2008/08/2061.html.

[180] C. Hu, W. Pan, and J. Kwok, "Accelerated gradient methods for stochastic optimization and online learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[181] G. Lan, "An optimal method for stochastic composite optimization," *Mathematical Programming*, vol. 133, no. 1-2, 2012, pp. 365–397.

[182] P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle," *Journal of Optimization Theory and Applications*, vol. 171, no. 1, 2016, pp. 121–145.

[183] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "A universally optimal multistage accelerated stochastic gradient method," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[184] E. Gorbunov, M. Danilova, and A. Gasnikov, "Stochastic optimization with heavy-tailed noise via accelerated gradient clipping," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[185] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *The Journal of Machine Learning Research (JMLR)*, vol. 14, 2013, pp. 567–599.

[186] A. J. Defazio, T. S. Caetano, and J. Domke, "Finito: A faster, permutable incremental gradient method for big data problems," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

[187] J. Mairal, "Incremental majorization-minimization optimization with application to large-scale machine learning," *SIAM Journal on Optimization*, vol. 25, no. 2, 2015, pp. 829–855.

Full text available at: http://dx.doi.org/10.1561/2400000036

[188] S. Shalev-Shwartz and T. Zhang, "Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

[189] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *The Journal of Machine Learning Research (JMLR)*, vol. 18, no. 1, 2017, pp. 8194–8244.

[190] K. Zhou, F. Shang, and J. Cheng, "A simple stochastic variance reduced algorithm with fast convergence rates," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[191] K. Zhou, Q. Ding, F. Shang, J. Cheng, D. Li, and Z.-Q. Luo, "Direct acceleration of SAGA using sampled negative momentum," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[192] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, 2012, pp. 341–362.

[193] Y. T. Lee and A. Sidford, "Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems," in *54th Symposium on Foundations of Computer Science*, pp. 147–156, 2013.

[194] O. Fercoq and P. Richtárik, "Accelerated, parallel, and proximal coordinate descent," *SIAM Journal on Optimization*, vol. 25, no. 4, 2015, pp. 1997–2023.

[195] Y. Nesterov and S. U. Stich, "Efficiency of the accelerated coordinate descent method on structured optimization problems," *SIAM Journal on Optimization*, vol. 27, no. 1, 2017, pp. 110–123.

[196] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, 2006, pp. 177–205.

[197] Y. Nesterov, "Accelerating the cubic regularization of Newton's method on convex problems," *Mathematical Programming*, vol. 112, no. 1, 2008, pp. 159–181.

Full text available at: http://dx.doi.org/10.1561/2400000036

[198] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C. Uribe, "Optimal tensor methods in smooth convex and uniformly convex optimization," in *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 2019.

[199] Y. Nesterov, "Implementable tensor methods in unconstrained convex optimization," *Mathematical Programming*, 2019, pp. 1–27.

[200] D. Kim and J. A. Fessler, "Another look at the fast iterative shrinkage/thresholding algorithm (FISTA)," *SIAM Journal on Optimization*, vol. 28, no. 1, 2018, pp. 223–250.

[201] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Exact worst-case performance of first-order methods for composite convex optimization," *SIAM Journal on Optimization*, vol. 27, no. 3, 2017, pp. 1283–1313.

[202] D. Kim and J. A. Fessler, "Generalizing the optimized gradient method for smooth convex minimization," *SIAM Journal on Optimization*, vol. 28, no. 2, 2018, pp. 1920–1950.

[203] Y. Drori and A. B. Taylor, "Efficient first-order methods for convex minimization: A constructive approach," *Mathematical Programming*, vol. 184, no. 1, 2020, pp. 183–220.

[204] D. Kim and J. A. Fessler, "Adaptive restart of the optimized gradient method for convex optimization," *Journal of Optimization Theory and Applications*, vol. 178, no. 1, 2018, pp. 240–263.

[205] C. Park, J. Park, and E. K. Ryu, "Factor-$\sqrt{2}$ acceleration of accelerated gradient methods," *preprint arXiv:2102.07366*, 2021.

[206] F. Lieder, "On the convergence rate of the halpern-iteration," *Optimization Letters*, vol. 15, no. 2, 2021, pp. 405–418.

[207] D. Kim, "Accelerated proximal point method for maximally monotone operators," *Mathematical Programming*, 2021, pp. 1–31.

[208] M. Barré, A. Taylor, and F. Bach, "Principled analyses and design of first-order methods with inexact proximal operators," *preprint arXiv:2006.06041*, 2020.

[209] Y. Drori and M. Teboulle, "An optimal variant of kelley's cutting-plane method," *Mathematical Programming*, vol. 160, no. 1-2, 2016, pp. 321–351.

Full text available at: http://dx.doi.org/10.1561/2400000036

[210] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Smooth strongly convex interpolation and exact worst-case performance of first-order methods," *Mathematical Programming*, vol. 161, no. 1-2, 2017, pp. 307–345.

[211] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *SIAM Journal on Optimization*, vol. 28, no. 3, 2018, pp. 2654–2689.

[212] E. De Klerk, F. Glineur, and A. B. Taylor, "On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions," *Optimization Letters*, vol. 11, no. 7, 2017, pp. 1185–1199.

[213] B. Hu and L. Lessard, "Dissipativity theory for nesterov's accelerated method," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[214] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Performance estimation toolbox (pesto): Automated worst-case analysis of first-order optimization methods," in *Proceedings of the 56th Conference on Decision and Control (CDC)*, 2017.

[215] H. Lin, J. Mairal, and Z. Harchaoui, "Catalyst acceleration for first-order convex optimization: From theory to practice," *The Journal of Machine Learning Research (JMLR)*, vol. 18, no. 1, 2018, pp. 7854–7907.

[216] J.-J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 255, 1962, pp. 2897–2899.

[217] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, 1965, pp. 273–299.

[218] B. Martinet, "Régularisation d'inéquations variationnelles par approximations successives," *Revue Française d'Informatique et de Recherche Opérationnelle*, vol. 4, 1970, pp. 154–158.

[219] B. Martinet, "Détermination approchée d'un point fixe d'une application pseudo-contractante. cas de l'application prox.," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 274, 1972, pp. 163–165.

Full text available at: http://dx.doi.org/10.1561/2400000036

[220] C. Lemaréchal and C. Sagastizábal, "Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries," *SIAM Journal on Optimization*, vol. 7, no. 2, 1997, pp. 367–385.

[221] O. Güler, "On the convergence of the proximal point algorithm for convex minimization," *SIAM Journal on Control and Optimization*, vol. 29, no. 2, 1991, pp. 403–419.

[222] M. Barré, A. Taylor, and F. Bach, "A note on approximate accelerated forward-backward methods with absolute and relative errors, and possibly strongly convex objectives," *preprint arXiv:2106.15536*, 2021.

[223] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui, "Catalyst for gradient-based nonconvex optimization," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[224] A. Kulunchakov and J. Mairal, "A generic acceleration framework for stochastic composite optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[225] R. T. Rockafellar, "A dual approach to solving nonlinear programming problems by unconstrained optimization," *Mathematical Programming*, vol. 5, no. 1, 1973, pp. 354–373.

[226] R. T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of operations research*, vol. 1, no. 2, 1976, pp. 97–116.

[227] A. N. Iusem, "Augmented Lagrangian methods and proximal point methods for convex optimization," *Investigación Operativa*, vol. 8, no. 11-49, 1999, p. 7.

[228] J. Eckstein and P. J. Silva, "A practical relative error criterion for augmented lagrangians," *Mathematical Programming*, vol. 141, no. 1-2, 2013, pp. 319–348.

[229] J. Eckstein, "Splitting methods for monotone operators with applications to parallel optimization," Ph.D. dissertation, Massachusetts Institute of Technology, 1989.

[230] 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*, vol. 3, no. 1, 2011, pp. 1–122.

Full text available at: http://dx.doi.org/10.1561/2400000036

[231] J. Eckstein and W. Yao, "Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," *RUTCOR Research Reports*, vol. 32, no. 3, 2012.

[232] S. Salzo and S. Villa, "Inexact and accelerated proximal point algorithms," *Journal of Convex analysis*, vol. 19, no. 4, 2012, pp. 1167–1192.

[233] M. V. Solodov and B. F. Svaiter, "A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator," *Set-Valued Analysis*, vol. 7, no. 4, 1999, pp. 323–345.

[234] M. V. Solodov and B. F. Svaiter, "A hybrid projection-proximal point algorithm," *Journal of convex analysis*, vol. 6, no. 1, 1999, pp. 59–70.

[235] M. V. Solodov and B. F. Svaiter, "Error bounds for proximal point subproblems and associated inexact proximal point algorithms," *Mathematical Programming*, vol. 88, no. 2, 2000, pp. 371–389.

[236] M. V. Solodov and B. F. Svaiter, "A unified framework for some inexact proximal point algorithms," *Numerical functional analysis and optimization*, vol. 22, no. 7-8, 2001, pp. 1013–1035.

[237] A. Ivanova, D. Grishchenko, A. Gasnikov, and E. Shulgin, "Adaptive catalyst for smooth convex optimization," *preprint arXiv:1911.11271*, 2019.

[238] J. Mairal, "Cyanure: An open-source toolbox for empirical risk minimization for python, c++, and soon more," *preprint arXiv:1912.08165*, 2019.

[239] Y. Nesterov, "Inexact accelerated high-order proximal-point methods," CORE discussion paper, Tech. Rep., 2020.

[240] Y. Nesterov, "Inexact high-order proximal-point methods with auxiliary search procedure," CORE discussion paper, Tech. Rep., 2020.

[241] J. Bolte, A. Daniilidis, and A. Lewis, "The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, 2007, pp. 1205–1223.

Full text available at: http://dx.doi.org/10.1561/2400000036

[242] A. S. Nemirovsky and Y. Nesterov, "Optimal methods of smooth convex minimization," *USSR Computational Mathematics and Mathematical Physics*, vol. 25, no. 2, 1985, pp. 21–30.

[243] Y. Nesterov, "Universal gradient methods for convex optimization problems," *Mathematical Programming*, vol. 152, no. 1-2, 2015, pp. 381–404.

[244] G. Li and T. K. Pong, "Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods," *Foundations of computational mathematics*, vol. 18, no. 5, 2018, pp. 1199–1232.

[245] J.-S. Pang, "A posteriori error bounds for the linearly-constrained variational inequality problem," *Mathematics of Operations Research*, vol. 12, no. 3, 1987, pp. 474–484.

[246] Z.-Q. Luo and P. Tseng, "On the linear convergence of descent methods for convex essentially smooth minimization," *SIAM Journal on Control and Optimization*, vol. 30, no. 2, 1992, pp. 408–425.

[247] P. Tseng, "Approximation accuracy, gradient methods, and error bound for structured convex optimization," *Mathematical Programming*, vol. 125, no. 2, 2010, pp. 263–295.

[248] Z. Zhou and A. M.-C. So, "A unified approach to error bounds for structured convex optimization problems," *Mathematical Programming*, 2017, pp. 1–40.

[249] T. Kerdreux, A. d'Aspremont, and S. Pokutta, "Restarting Frank-Wolfe," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[250] S. Lacoste-Julien and M. Jaggi, "On the global linear convergence of Frank-Wolfe optimization variants," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[251] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the AMS*, vol. 22, no. 1, 2009, pp. 211–231.

[252] M. Ito and M. Fukuda, "Nearly optimal first-order methods for convex optimization under gradient norm measure: An adaptive regularization approach," *Journal of Optimization Theory and Applications*, 2021, pp. 1–35.

Full text available at: http://dx.doi.org/10.1561/2400000036

[253] S. Lojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," *Les équations aux dérivées partielles*, 1963, pp. 87–89.

[254] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, pp. 769–783, 1998.

[255] Z. Zhou, Q. Zhang, and A. M.-C. So, "L1, p-norm regularization: Error bounds and convergence rate analysis of first-order methods," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[256] D. Davis, D. Drusvyatskiy, and V. Charisopoulos, "Stochastic algorithms with geometric step decay converge linearly on sharp functions," *preprint arXiv:1907.09547*, 2019.

[257] E. De Klerk, F. Glineur, and A. B. Taylor, "Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation," *SIAM Journal on Optimization*, vol. 30, no. 3, 2020, pp. 2053–2082.

[258] A. S. Nemirovsky and D. B. Yudin, "Information-based complexity of mathematical programming (in Russian)," *Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika (the journal is translated to English as Engineering Cybernetics. Soviet J. Computer & Systems Sci.)*, vol. 1, 1983.

[259] A. S. Nemirovsky, "Orth-method for smooth convex optimization," *Izvestia AN SSSR, Transl.: Eng. Cybern. Soviet J. Comput. Syst. Sci*, vol. 2, 1982, pp. 937–947.

[260] G. Narkiss and M. Zibulevsky, *Sequential subspace optimization method for large-scale unconstrained problems*. Technion-IIT, Department of Electrical Engineering, 2005.

[261] S. Karimi and S. A. Vavasis, "A unified convergence bound for conjugate gradient and accelerated gradient," *preprint arXiv:1605.00320*, 2016.

[262] J. Diakonikolas and L. Orecchia, "Conjugate gradients and accelerated methods unified: The approximate duality gap view," *preprint arXiv:1907.00289*, 2019.

Full text available at: http://dx.doi.org/10.1561/2400000036

[263] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Transactions on Automatic Control*, vol. 42, no. 6, 1997, pp. 819–830.

[264] L. Vandenberghe and S. Boyd, "Applications of semidefinite programming," *Applied Numerical Mathematics*, vol. 29, no. 3, 1999, pp. 283–299.

[265] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Exact worst-case convergence rates of the proximal gradient method for composite convex minimization," *Journal of Optimization Theory and Applications*, vol. 178, no. 2, 2018, pp. 455–476.

[266] B. Hu, P. Seiler, and A. Rantzer, "A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints," in *Proceedings of the 30th Conference on Learning Theory (COLT)*, 2017.

[267] A. Taylor, B. Van Scoy, and L. Lessard, "Lyapunov functions for first-order methods: Tight automated convergence guarantees," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[268] B. Hu, P. Seiler, and L. Lessard, "Analysis of biased stochastic gradient descent using sequential semidefinite programs," *Mathematical Programming*, vol. 187, no. 1, 2021, pp. 383–408.

[269] E. K. Ryu and W. Yin, *Large-Scale Convex Optimization via Monotone Operators*. 2020.

[270] C. Park and Ryu, "Optimal first-order algorithms as a function of inequalities," *preprint arXiv:2110.11035*, 2021.

[271] D. Gramlich, C. Ebenbauer, and C. W. Scherer, "Convex synthesis of accelerated gradient algorithms for optimization and saddle point problems using lyapunov functions," *preprint arXiv:2006.09946*, 2020.

[272] S. Safavi, B. Joshi, G. França, and J. Bento, "An explicit convergence rate for nesterov's method from sdp," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1560–1564, 2018.

Full text available at: http://dx.doi.org/10.1561/2400000036

[273] B. Hu, S. Wright, and L. Lessard, "Dissipativity theory for accelerating stochastic variance reduction: A unified analysis of SVRG and katyusha using semidefinite programs," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[274] Z. Shi and R. Liu, "Better worst-case complexity analysis of the block coordinate descent method for large scale machine learning," in *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[275] H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani, "The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions," *Optimization Letters*, 2021.

[276] H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani, "On the rate of convergence of the difference-of-convex algorithm (DCA)," *reprint arXiv:2109.13566*, 2021.

[277] E. K. Ryu and B. C. Vũ, "Finding the forward-douglas–rachford-forward method," *Journal of Optimization Theory and Applications*, vol. 184, no. 3, 2020, pp. 858–876.

[278] E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson, "Operator splitting performance estimation: Tight contraction factors and optimal parameter selection," *SIAM Journal on Optimization*, vol. 30, no. 3, 2020, pp. 2251–2271.

[279] G. Gu and J. Yang, "On the optimal ergodic sublinear convergence rate of the relaxed proximal point algorithm for variational inequalities," *preprint arXiv:1905.06030*, 2019.

[280] G. Gu and J. Yang, "Tight sublinear convergence rate of the proximal point algorithm for maximal monotone inclusion problems," *SIAM Journal on Optimization*, vol. 30, no. 3, 2020, pp. 1905–1921.

[281] G. Zhang, X. Bao, L. Lessard, and R. Grosse, "A unified analysis of first-order methods for smooth games via integral quadratic constraints," *The Journal of Machine Learning Research (JMLR)*, vol. 22, no. 103, 2021, pp. 1–39.

Full text available at: http://dx.doi.org/10.1561/2400000036

[282] A. Sundararajan, B. Van Scoy, and L. Lessard, "Analysis and design of first-order distributed optimization algorithms over time-varying graphs," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 4, 2020, pp. 1597–1608.

[283] S. Colla and J. M. Hendrickx, "Automated worst-case performance analysis of decentralized gradient descent," in *Proceedings of the 60th Conference on Decision and Control (CDC)*, 2021.