Adaptive Algorithms for Optimization Beyond Lipschitz Requirements

Kimon Antonakopoulos

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques et Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Kimon ANTONAKOPOULOS

Thèse dirigée par **Panayotis MERTIKOPOULOS**
et codirigée par **Elena Veronica BELMEGA**, CY Cergy Paris Université

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Algorithmes adaptatifs pour l'optimisation au-delà des conditions de Lipschitz

Adaptive Algorithms for Optimization Beyond Lipschitz Requirements

Thèse soutenue publiquement le **12 janvier 2022**,
devant le jury composé de :

Monsieur PANAYOTIS MERTIKOPOULOS
Chargé de recherche HDR, CNRS DELEGATION ALPES, Directeur de thèse
Monsieur MARC TEBOULLE
Professeur, University of Tel Aviv, Rapporteur
Monsieur WALID HACHEM
Directeur de recherche, CNRS ILE-DE-FRANCE VILLEJUIF, Rapporteur
Madame NIAO HE
Maître de conférences HDR, Ecole Polytechnique Fédérale de Zurich, Examinatrice
Monsieur JERÔME MALICK
Directeur de recherche, CNRS DELEGATION ALPES, Président
Monsieur ROBERTO COMINETTI
Professeur, Universidad Adolfo Ibanez, Examinateur
Madame ELENA VERONICA BELMEGA
Maître de conférences HDR, UNIVERSITE DE CERGY-PONTOISE, Co-directrice de thèse

ADAPTIVE METHODS FOR OPTIMIZATION WITHOUT LIPSCHITZ REQUIREMENTS

KIMON ANTONAKOPOULOS

A thesis submitted for the degree of
Doctor of Philosophy

SUPERVISORY COMMITTEE

P. MERTIKOPOULOS	CNRS & Université Grenoble Alpes
E. V. BELMEGA	ETIS / ENSEA - CY Cergy Paris Université

REVIEWERS & EXAMINERS

W. HACHEM	CNRS & Université Gustave Eiffel
M. TEBOULLE	Tel Aviv University
R. COMINETTI	Universidad Adolfo Ibañez
N. HE	ETH Zürich
J. MALICK	CNRS & Université Grenoble Alpes

Université Grenoble Alpes
École doctorale MSTII
Spécialité: *Informatique et Mathématiques Appliquées*
Grenoble, April 12, 2022

So there you sit. And how much blood was shed
That you might sit there. Do such stories bore you ?

Well, don't forget that others sat before you
who later sat on people. Keep your head!

Your science will be valueless, you'll find
And learning will be sterile, if inviting
Unless you pledge your intellect to fighting
Against all enemies of all mankind.

Never forget that men like you got hurt
That you might sit here, not the other lot.

And now don't shut your eyes, and don't desert
But learn to learn, and try to learn for what.

— Bertolt Brecht, *To The Students Of The Workers' And Peasants' Faculty*

ABSTRACT

Several important problems in learning theory and data science involve high-dimensional optimization objectives that transcend the Lipschitz regularity conditions that are standard in the field. This absence of Lipschitz regularity – smoothness or continuity – poses significant challenges to the convergence analysis of most optimization algorithms and, in many cases, it requires the introduction of novel analytical and algorithmic tools. In this thesis, we aim to partially fill this gap via the design and analysis of universal first-order methods in two general optimization frameworks: (*a*) online convex optimization (which contains as special cases deterministic and stochastic convex optimization problems); and (*b*) abstract variational inequalities (which contain as special cases min-max problems and games) both without global Lipschitz continuity/smoothness conditions.

In this "NoLips" setting, we take a geometric approach – Riemannian, Finslerian, or Bregman-based – that allows us to handle vector fields and functions whose norm or variation becomes infinite at the boundary of the problem's domain. Using these non-Euclidean surrogates for Lipschitz continuity and smoothness, we propose a range of adaptive first-order methods that concurrently achieve order-optimal convergence rates in different problem classes, without any prior knowledge of the class or the problem's (relative) smoothness parameters. These methods are based on a suitable mirror descent or mirror-prox template (for convex minimization and monotone variational inequalities respectively), and they revolve around adaptive step-size policies that exploit the geometry of the gradient data observed at earlier iterations to perform more informative (extra-)gradient steps in later ones. Our results do not always coincide with what one would expect in standard Lipschitz problems, and serve to further highlight the differences between the "Lispschitz" and "NoLips" frameworks.

RESUME

PLUSIEURS problèmes importants issus de l'apprentissage statistique et de la science des données concernen des objectifs d'optimisation à très haute dimension qui vont au delà des hypothèses de régularité de Lipschitziennes. L'absence de régularité de Lipschitz – continuité ou lissitude – pose des défis importants à l'analyse de convergence des algorithmes existants d'optimisation et nécessite souvent de nouveaux outils analytiques et algorithmiques pour être traitée de manière efficace. Dans cette thèse, nous visons à combler partiellement cette lacune en proposant de nouvelles méthodes universelles du premier ordre dans deux cadres généraux : (a) l'optimisation convexe en ligne (qui contient comme cas particuliers les problèmes d'optimisation convexe déterministes et stochastiques) ; et (b) les inégalités variationnelles (qui contiennent comme cas particulier les problèmes de point-selle et les jeux).

Nous étudions ces deux problèmes génériques dans un "NoLips" et nous adoptons une approche géométrique - Riemannienne ou Bregmanienne - qui nous permet de traiter des champs vectoriels et/ou des fonctions dont la norme ou la variation explose vers le bord du domaine du problème. En utilisant ces substituts non-euclidiens, nous proposons des nouvelles méthodes adaptatives du premier ordre qui atteignent simultanément des taux de convergence optimaux dans différentes classes de problèmes, sans aucune connaissance préalable des paramètres de régularité du problème. Nos méthodes sont basées sur un template "mirror descent" ou "mirror-prox" (pour les problèmes de minimisation convexe et les inégalités variationnelles monotones respectivement), et elles se basent sur des politiques adaptatives de pas qui exploitent l'historique des gradients observés afin d'effectuer des pas de gradient mieux adaptés à la régularité du problème. Nos résultats ne coïncident pas toujours avec ce que l'on attendrait dans le cadre Lipschitz, et ainsi apportent une intuition très utile pour comprendre les différences fondamentales entre les problèmes "Lispschitz" et "NoLips".

ACKNOWLEDGMENTS

First of all, I would like to warmly thank my advisor Panayotis Mertikopoulos, without him this thesis (literally) wouldn't have never been written. During the many ups and downs of this journey, he was always the voice of reason in this madness, helping me out to focus on the important at any given moment. Whenever I despaired with a proof or a calculation (and happened very often), Panayotis was always there to lend a helping hand and a word of support; he would always identify the essence and the correct path to proceed. But, above all else, Panayotis has been a friend, a wise friend who gave advice (very) carefully but freely, and on whose experience and kindness I could always draw when needed. His door was always open to discuss and to empathize my desperation at the (endlessly) many dead ends. For these reasons (and many more) I am truly indebted to Panayotis and I can only hope that it does not end with this thesis.

Moreover, I would like to warmly thank my co-advisor Veronica Belmega who did also her fair part to make the write-up of this document possible. She literally was the enthusiastic voice which, even at the most difficult times, made always the glass to seem half full; and for this I truly grateful.

Of course, this document would never have existed without the great effort and diligence of all the members of my committee. Therefore, I would first like to thank professors Marc Teboulle and Walid Hachem for their extremely detailed reviews (especially under these tight time constraints) which helped to improve the rigour of this document significantly. Moreover, I thank each member of my committee, Roberto Cominetti, Niao He and Jerome Malick for the fruitful discussion during my defence and of course for their time and effort to read this thesis.

Last but not least, I would like to thank all the people who put up with me during this long journey: my parents without their sacrifices nothing would have been possible, my friends (Eleni, Giannis) for their support and my girlfriend Archontia for tolerating me all these years.

FINANCIAL SUPPORT. The author of this thesis gratefully acknowledges financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grants ORACLESS (ANR-16-CE33-0004) and ALIAS (ANR-19-CE48-0018-01). This research was also supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Kurd-L	Kurdyka–Łojasiewicz
AdaNorm	adaptive inverse-norm-squared
AdaGrad	adaptive gradient descent
AcceleGrad	adaptive accelerated gradient
AdaProx	adaptive mirror-prox
AMP	adaptive mirror prox
DualX	dual extrapolation
MDS	martingale difference sequence
UniXGrad	universal extra-gradient
AdaMir	adaptive mirror descent
EGD	entropic gradient descent
EG	extra-gradient
UniXGrad	universal extra-gradient
UPGD	universal primal gradient descent

KL	Kullback-Leibler
GAN	generative adversarial network
GD	gradient descent
LMD	lazy mirror descent
PIP	Poisson inverse problems
IGA	improved interior gradient algorithm
FISTA	fast iterative shrinkage-thresholding algorithm
i.i.d.	independent and identically distributed
l.s.c.	lower semi-continuous
MD	mirror descent
NE	Nash equilibrium
OCO	online convex optimization
OMD	online mirror descent
PR	proportional response
RHS	right-hand side
SFO	stochastic first-order oracle
SP	saddle-point
VI	variational inequality
MP	mirror-prox
BL	Bach–Levy
GMP	generalized mirror-prox
GRAAL	golden ratio algorithm

PHD PUBLICATIONS

Some of the material presented in this thesis has appeared – or is set to appear – in the following publications:

1. K. Antonakopoulos and P. Mertikopoulos, "*Universal methods for variational inequalities with divergent operators*," *in preparation*, 2021.
2. K. Antonakopoulos, T. Pethick, A. Kavis, P. Mertikopoulos and V. Cevher, "*Sifting through the noise: Universal first-order methods for stochastic variational inequalities*,"*in NeurIPS 2021: Proceedings of the 35th International Conference on Neural Processing Information Systems*, 2021.
3. K. Antonakopoulos and P. Mertikopoulos, "*Adaptive First-Order Methods Revisited: Convex Minimization without Lipschitz Requirements*," *in NeurIPS 2021: Proceedings of the 35th International Conference on Neural Processing Information Systems*, 2021.
4. D.-Q. Vu, K. Antonakopoulos and P. Mertikopoulos, "*Fast Routing in an Uncertain World: Adaptive Learning in Congestion Games via Exponential Weights*," *in NeurIPS 2021: Proceedings of the 35th International Conference on Neural Processing Information Systems*, 2021.
5. Y.-G. Hsieh, K. Antonakopoulos and P. Mertikopoulos,"*Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Equilibrium*," *in COLT 2021: Proceedings of 34th Annual Conference on Learning Theory*, 2021.
6. K. Antonakopoulos, E.V. Belmega and P. Mertikopoulos, "*Adaptive Extra-Gradient Methods for Min-Max Optimization and Games*," *in ICLR 2021: Proceedings of the 9th International Conference on Learning Representations*, 2021.
7. K. Antonakopoulos, E.V. Belmega and P. Mertikopoulos, "*Online and Stochastic Optimization Beyond Lipschitz Continuity: A Riemannian Approach*," *in ICLR 2020: Proceedings of the 8th International Conference on Learning Representations*, 2020.
8. K. Antonakopoulos, E.V. Belmega and P. Mertikopoulos, "*An Adaptive Mirror-Prox Method for Variational Inequalities with Singular Operators*," *in NeurIPS 2019: Proceedings of the 33rd International Conference on Neural Processing Information Systems*, 2019.

1

INTRODUCTION

The rise of machine learning protocols has reaffirmed the interest in the theory of optimization problems. To that end, two important settings stand out, that of *online convex optimization* and (monotone) *variational inequality problems*.

The first framework refers to a scenario where the optimizer faces a (possibly adversarial) sequence of time-varying loss functions f_t, $t = 1, 2, \ldots$, one at a time – for instance, when drawing different sample points from a large training set [31, 105]. Specifically, if the optimizer faces a sequence of G-Lipschitz convex losses, the incurred min-max regret, a standard performance criterion that will be discussed in detail later, is $\Omega(GT^{1/2})$ after T rounds and this bound can be achieved by inexpensive first-order methods – such as online mirror descent and its variants [31, 105, 106, 122].

This setting properly includes (static) convex minimization problems, but the situation in this case changes dramatically. The analysis of static minimization problems typically revolves around two main regularity conditions for the problem at hand: (*a*) *Lipschitz continuity of the problem's objective function* and/or (*b*) *Lipschitz continuity of its gradient (also referred to as Lipschitz smoothness).* Depending on which of these conditions holds, the lower bounds for first-order methods with perfect gradient input are $\Theta(1/\sqrt{T})$ and $\Theta(1/T^2)$ after T gradient queries, and they are achieved by gradient descent and Nesterov's fast gradient algorithm respectively [88, 89]. By contrast, if the optimizer only has access to stochastic gradients (as is often the case in machine learning and distributed control), the corresponding lower bound is $\Theta(1/\sqrt{T})$ for both problem classes [30, 86, 89].

On the other hand, the surge of recent breakthroughs in generative adversarial networks (GANs) [46], robust reinforcement learning [97], and other adversarial learning models [73] has sparked renewed interest in the theory of min-max optimization problems and games. In this broad setting, it has become empirically clear that, ceteris paribus, the simultaneous training of two (or more) antagonistic models faces drastically new challenges relative to the training of a single one. Perhaps the most prominent of these challenges is the appearance of cycles and recurrent (or even chaotic) behavior in min-max games. This has been studied extensively in the context of learning in bilinear games, in both continuous [41, 80, 96] and discrete time [37, 43, 44, 81], and the methods proposed to overcome recurrence typically focus on mitigating the rotational component of min-max games.

The method with the richest history in this context is the extra-gradient (EG) algorithm of Korpelevich (1976) and its variants. The EG algorithm exploits the

Lipschitz smoothness of the problem and, if coupled with a Polyak–Ruppert averaging scheme, it achieves an $\mathcal{O}(1/T)$ rate of convergence in smooth, convex-concave min-max problems [85]. This rate is known to be tight [84, 95] but, in order to achieve it, the original method requires the problem's Lipschitz constant to be known in advance. If the problem is not Lipschitz smooth (or the algorithm is run with a vanishing step-size schedule), the method's rate of convergence drops to $\mathcal{O}(1/\sqrt{T})$.

From the above, one may directly observe that from a practical perspective the challenging part in order to apply the respective optimal solution method to the problem at hand is to able to identify which regularity condition and/or oracle feedback she has at hand. Therefore, a question that naturally arises in this context is the following:

Is it possible to design methods that simultaneously achieve optimal convergence rates without any prior knowledge of the problem's regularity features ?

The positive answer to the above question gives rise to the so-called *adaptive methods*. In its general context, adaptivity of a method may refer to (at least) two different things:

1. Automatic adjustment to the function's regularity parameters within a fixed problem class (Lipschitz continuous, Lipschitz smooth, etc.).
2. Interpolation of convergence rates between different problem classes (e.g., $\mathcal{O}(1/\sqrt{T})$ for non-smooth vs. $\mathcal{O}(1/T)$ or $\mathcal{O}(1/T^2)$ for smooth, etc.).

In what follows we treat both questions in tandem.

1.1 RELATED WORK

MINIMIZATION PROBLEMS. There is an extensive corpus of literature concerning the convex minimization framework. To name out the methods of [54] and [65] successfully interpolate between the stochastic and smooth deterministic regimes achieving a $\mathcal{O}(1/\sqrt{T})$ convergence rate for the former and an $\mathcal{O}(1/T^2)$ rate for the latter; however, their interpolation guarantees require prior knowledge of the function's smoothness parameter. More recently, [92] proposed a method that adjusts automatically to the Lipschitz (or Hölder) modulus of the function based on line-search queries of the objective[1]; in the Lipschitz smooth case, the method of Nesterov [92] attains an accelerated rate of convergence of the. order $\mathcal{O}(1/T^2)$. However, in order to establish an implementable stopping criterion, said method requires as an input parameter an estimate of the distance between the algorithm's initial state to the problem's solution set (i.e., this upper bound should be known to the optimizer a priori).

Such an estimate is difficult to come by in problems with unbounded domains, so the performance of the method is unclear in this case.

By contrast, the AcceleGrad method of [67] and the more recent UnixGrad algorithm of [60] successfully interpolate between the $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T^2)$

1 Non smooth and smooth objectives are included as extreme cases for the Hölderian exponent $q = 0, 1$.

rates for the Lipschitz continuous and/or stochastic settings and smooth regimes respectively without requiring a line search – but the boundedness caveat is still present. Finally, beyond the minimization framework, [14] proposed a universal mirror-prox method for solving (stochastic) variational inequalities, with or without smoothness requirements. When applied to function minimization, the algorithm of [14] interpolates between the $\mathcal{O}(1/\sqrt{T})$ and an unaccelerated $\mathcal{O}(1/T)$ rate. What makes this particularly interesting for our purposes is that this scheme does not require vanishing gradients near a minimizer.

VARIATIONAL INEQUALITIES. There have been several works focusing on the convergence guarantees of the original EG / mirror-prox (MP) template. We review the most relevant of these works below. In unconstrained problems with an operator that is locally Lipschitz continuous (but not necessarily globally so), the golden ratio algorithm (GRAAL) of [75] achieves convergence without requiring prior knowledge of the problem's Lipschitz parameter. However, GRAAL provides no rate guarantees for non-smooth problems – and hence, a fortiori, no interpolation guarantees either. By contrast, such guarantees are provided in problems with a bounded domain by the generalized mirror-prox (GMP) algorithm of [109] under the umbrella of Hölder continuity.

Another method that simultaneously achieves an $\mathcal{O}(1/\sqrt{T})$ rate in non-smooth problems and an $\mathcal{O}(1/T)$ rate in smooth ones is the recent algorithm of Bach and Levy (2019). The BL algorithm employs an adaptive, AdaGrad-like step-size policy which allows the method to interpolate between the two regimes – and this, even with noisy gradient feedback. On the negative side, the BL algorithm requires a bounded domain with a (Bregman) diameter that is known in advance; as a result, its theoretical guarantees do not apply to unbounded problems.

BEYOND LIPSCHITZ REGULARITY. Despite the fact that the (Euclidean-based) Lipschitz regularity conditions appear quite generic there exists a whole set of real life situations where both of these conditions fail, either because the loss profile of the problem grows too rapidly (e.g., as in support vector machines or GAN models with Kullback-Leibler losses), or because the problem exhibits singularities near the boundary of the feasible region (e.g., as in resource allocation and inverse problems). A prominent example that will serve as motivation for the NoLips setting is that of Poisson Inverse Problems. We examine this in detail below.

Example 1.1 (Poisson Inverse Problems)**.** Poisson inverse problem (PIP) arise in various practical problems stemming from image sciences and machine learning problems. Informally, this consists of two components: a matrix $A \in \mathbb{R}^{m\times n}$ which models the experimental protocol and a vector $b \in \mathbb{R}^m_+$ represents the measurements made by the optimizer. With all this in hand, the objective would be to recover the signal or image $x \in \mathbb{R}^n_+$ from the noisy measurements b such that:

$$Ax \simeq b \tag{1.1}$$

A natural measure that evaluates the proximity of these two vectors is that of the Kullback–Leibler (KL) divergence. Namely, we are facing the following convex minimization problem:

$$\begin{aligned} &\text{minimize} \quad d(b, Ax) = \sum_{i=1}^{m} \left[b_i \log \frac{b_i}{(Ax)_i} + (Ax)_i - b_i \right] \\ &\text{subject to} \quad x \in \mathbb{R}^n_+ \end{aligned} \tag{1.2}$$

As one may recognize the above minimization objective is neither Lipschitz contiuous nor smooth due to the singular behaviour of the logarithm near the origin.

The above schemes all rely intrinsically on Lipschitz/Hölder continuity and/or smoothness. Achieving convergence beyond the Lipschitz framework has been the focal point of a recent strand in the literature, starting with the work of [19] and the concurrent paper of [72]. More recent works have provided different extensions to non-convex [25] and stochastic optimization [48], including a tentative path towards acceleration [49]; however, these methods are neither universal nor adaptive.

In more detail Bauschke et al. [19] introduced a "Lipschitz-like" smoothness condition for convex minimization problems and used it to establish a $\mathcal{O}(1/T)$ value convergence rate for mirror descent methods (as opposed to mirror-prox). Always in the context of loss minimization problems, Bolte et al. [25] subsequently extended the results of Bauschke et al. [19] to non-convex problems that satisfy the Kurdyka–Łojasiewicz (KL) inequality, while Lu et al. [72] considered functions that are also relatively strongly convex and showed that mirror descent achieves a geometric convergence rate in this context.

The condition of Bauschke et al. [19] is remarkably simple as it only posits that the problem's loss function f is such that :

$$\beta h - f \ \text{ is convex} \tag{RS}$$

for some reference Bregman function h and some $\beta > 0$. A straightforward extension of this condition to an operator setting would be to require the monotonicity of $\beta \nabla h - A$, where A is the operator defining the variational inequality under study. However, the cornerstone of this "Lipschitz-like" condition is a descent lemma which does not carry over to variational inequalities, so it does not seem possible to extend the analysis of Bauschke et al. [19] to an operator setting at least not directly.

Insofar as Lipschitz continuity of the objective is concerned, Lu [71] also considered a "relative continuity" condition for loss minimization problems positing that

$$\|\nabla f(x)\| \leq G \inf_{x'} \sqrt{2D(x', x)} / \|x' - x\| \tag{1.3}$$

(where f is the problem's objective and D is the Bregman divergence of h). Written this way, the condition of Lu [71] can also be extended to an operator setting, but this would provide a surrogate for operator *boundedness*, not Lipschitz continuity (since $A = \nabla f$ in minimization problems). Extending the above definition Zhou et al. [119] proposed a similar notion, i.e.,

$$\langle \nabla f(x), x - x' \rangle \leq G \sqrt{2D(x', x)} \tag{RC}$$

and applied it for the context of online convex optimization problems. Finally, in Teboulle [111] the notion of $W[h]$-continuity is proposed by singling out particular properties of Bregman divergences; formally, given an appropriate regularizer h an operator A is called to be $W[h]$- continuous:

$$t\langle A(x), x - x'\rangle - D(x', x) \leq \frac{t^2}{2}G^2 \quad \text{for all } x' \in \operatorname{dom} h,\ x \in \operatorname{dom} \partial h. \tag{W}$$

In the sequel, we shall introduce an alternative way that will allow us to extend the Lipschitz continuity conditions in a unified manner for both minimization and (VI) problems.

1.2 MAIN OBJECTIVE AND CONTRIBUTIONS OF THIS THESIS

In view of the above, the objective of this thesis is twofold:

1. Introduce novel regularity conditions, which are able to include variational inequality problems whose associated operator exhibits a "singular" behaviour.
2. Bridge the gap between the development of general Lipschitz continuity conditions on the one hand and the lack of respective adaptive methods on the other.

Tackling each objective separately, we begin by introducing two novel classes of operators. In particular inspired by the idea that Lipschitz continuity is first and foremost a metric space property we use the notion of local norms extensively as a primal geometrical tool in order to capture finer geometrical aspects of the problem. More precisely, in contrast to the traditional setting, local norms dependent on the point where it is evaluated, i.e., we have a continuous assignment $\|\cdot\|_x$ for all $x \in \mathcal{X}$. This in turn defines the associated dual norm in the standard way, i.e., for all $w \in \mathcal{V}^*$,

$$\|w\|_{x,*} = \max\{\langle w, x'\rangle : \|x'\|_x = 1\} \tag{1.4}$$

Armed with this geometry-aware local norm machinery we revisit the Euclidean based regularity conditions. In particular, we define two new operator classes that of metrical boundedness and metrical smoothness (see [6, 8]). Formally, an operator A is called *metrically bounded* when:

$$\|A(x)\|_{x,*} \leq G \tag{MB}$$

and *metrically smooth* whenever the following inequality holds:

$$\|A(x) - A(x')\|_{x,*} \leq \beta\|x - x'\|_{x'} \tag{MS}$$

In this context, the adaptivity results evolve throughout this thesis gradually. More precisely, our contributions can be summarized as follows:

- We begin gently by investigating online convex optimization (OCO) problems by recovering optimal regret minimization upper bounds under (MB).
- We proceed by taking a closer look at static/ stochastic convex minimization problems. More precisely we establish optimal interpolation guarantees

for both stochastic and/or deterministic oracle feedback under the blanket assumptions of (RS) and (RC).

- In the last part of this thesis, we focus on the generic framework of variational inequalities. To that end, for this setting we provide convergence rates starting from non-adaptive to adaptive to the "Lipschitz"-like modulus and finally regime-agnostic order optimal interpolation guarantees for both deterministic and stochastic (VI)'s under (MB) and/or (MS).

In what follows we present the content of each chapter in a more detailed manner.

1.3 DIAGRAMMATIC OUTLINE

This thesis consists of two parts. In Part I the general theoretical setup is presented, while Part II examines the particular algorithmic guarantees achieved in each setting. We now provide a quick overview of the content of each chapter individually.

Preliminaries
- Chapter 2 contains the main ingredients of this thesis; the particular problem set-ups along with the state of the art first order methods and the respective convergence rate guarantees. An important part of this chapter is devoted to the pivotal role that Lipschitz continuity plays in all these optimization scenarios.

NoLips
- Chapter 3 introduces and examines in detail the NoLips conditions discussed above. In doing, we distinguish our presentation for the different optmization frameworks.

Bregman Methods
- Chapter 4 provides concrete definitions of the main algorithmic schemes which will be of interest throughout the sequel. More precisely, we start with the basic mathematical toolkit of Bregman divergences which serves as the key ingredient for generalizing the standard Euclidean based projection operators. Based on this machinery, we describe a set of Bregman driven iterative methods for both optimization scenarios.

Online Convex Optimization
- Chapter 5 Motivated by applications to machine learning and imaging science, we study a class of online and stochastic optimization problems with loss functions that are not Lipschitz continuous; in particular, the loss functions encountered by the optimizer could exhibit gradient singularities or be singular themselves. Drawing on tools and techniques from Finsler geometry, we examine the (MB) continuity condition which is tailored to the singularity landscape of the problem's loss functions. In this way, we are able to tackle cases beyond the Lipschitz framework provided by a global norm, and we derive optimal regret bounds and last iterate convergence results through the use of regularized learning methods (such as online mirror descent).

Convex Optimization
- Chapter 6 We propose a new family of adaptive first-order methods for a class of convex minimization problems that may fail to be Lipschitz continuous or smooth in the standard sense. Specifically, we consider problems that are continuous or smooth relative to a reference Bregman function – as opposed to a global, ambient norm (Euclidean or otherwise). In this setting, the application of existing order-optimal adaptive methods – like UnixGrad or AcceleGrad– is not possible, especially in the presence of randomness

and uncertainty. The proposed method, adaptive mirror descent (AdaMir), aims to close this gap by concurrently achieving min-max optimal rates in problems that are relatively continuous or smooth, including stochastic ones.

Variational Inequalities

- Chapter 7 We present a new family of min-max optimization algorithms that automatically exploit the geometry of the gradient data observed at earlier iterations to perform more informative extra-gradient steps in later ones.

  Thanks to this adaptation mechanism, our proposed method, adaptive mirror-prox (AdaProx) automatically detects whether the problem is smooth or not, without requiring any prior tuning by the optimizer. As a result, AdaProx simultaneously achieves order-optimal convergence rates, i.e., it converges with a rate of $\mathcal{O}(1/T)$ iterations in smooth problems, and $\mathcal{O}(1/\sqrt{T})$ in non-smooth ones. Importantly, these guarantees do not require any of the standard boundedness or Lipschitz continuity conditions that are typically assumed in the literature; in particular, they apply even to problems with singularities (such as resource allocation problems and the like). This adaptation is achieved through the use of a geometric apparatus based on Finsler metrics and a suitably chosen mirror-prox template that allows us to derive sharp convergence rates for the methods at hand.

  Moving forward, we finally illustrate the full potential of our results. Namely, by employing the dual extrapolation (DualX) template run with a similar adaptive learning as is AdaProx, we are able to show optimal convergence rates for both deterministic and stochastic oracles and smooth and non-smooth settings.

1.4 NOTATIONAL CONVENTIONS

Throughout the sequel, $\mathcal{V} \cong \mathbb{R}^n$ will denote an n-dimensional space with norm $\|\cdot\|$ and $\mathcal{V}^*$ will denote its (algebraic) dual. We will also write $\langle w, x \rangle$ for the canonical pairing between $w \in \mathcal{V}^*$ and $x \in \mathcal{V}$, and $\|w\|_* \equiv \max\{\langle w, x \rangle : \|x\| \leq 1\}$ for the associated dual norm on $\mathcal{V}^*$. We also use the notation $\widetilde{\mathcal{O}}(\cdot)$ to dismiss logarithmic factors.

Part I

THE ROLE OF REGULARITY IN OPTIMIZATION

2

PRELIMINARIES

The main objective of this introductory chapter is to present the basic concepts of two general optimization scenarios: *a*) the time-varying setting of *online convex optimization* (OCO); and *b*) the operator-based setting of *variational inequalities* (VIs). In both frameworks, we seek to briefly review the main definitions, applications, and state-of-the-art solution methods.

To begin with, the online convex optimization setting – presented in detail in Section 2.1 – concerns decision-making processes that unfold in an otherwise unknown and time-varying environment. More precisely, the optimizer is assumed to be facing a sequence of convex losses f_t which evolves from round to round, possibly in an adversarial manner. This framework properly includes as special cases the class of convex minimization problems, deterministic and/or stochastic; these problems will be of individual interest throughout as well.

Moving forward, Section 2.2 provides a detailed description of an optimization framework that goes beyond ordinary minimization problems – the general setting of variational inequalities. This setup serves as a unifying framework for various "convex-structured" optimization problems so, in addition to standard minimization problems, it allows us to put under the same umbrella cases such as saddle-point, fixed-point and Nash equilibrium problems.

Having described these two settings of interest, in Section 2.3 we discuss two generic regularity conditions – boundedness and Lipschitz continuity of the defining operators of each problem class. Subsequently, in Section 2.4 we present the general framework of first-order methods which will be our main candidate solution methods. Moreover, we illustrate how the performance of these methods is influenced under each specific regularity condition, by providing "worst-case" optimal lower bounds. Finally, in Section 2.5 and Section 2.6 we present the state-of-the-art first-order methods that match the optimal lower bounds along with their adaptive counterparts.

2.1 ONLINE CONVEX OPTIMIZATION

2.1.1 *Problem setup and examples*

We begin by presenting the core protocol of *online convex optimization* (OCO), i.e., when the optimizer faces a sequence of time-varying loss functions f_t, $t = 1, 2, \ldots$, one at a time. Formally, this can be described by the following sequence of events:

1. At each round $t = 1, 2, \ldots$, the optimizer chooses an *action* X_t from a convex – but not necessarily closed or compact – subset $\mathcal{X}$ of an ambient normed space $\mathcal{V} \cong \mathbb{R}^n$.
2. The optimizer incurs a loss $f_t(X_t)$ based on some (a priori unknown) *loss function* $f_t \colon \mathcal{X} \to (-\infty, +\infty]$ which is assumed to be proper, lower semi-continuous (l.s.c.) and convex.
3. The optimizer updates their action and the process repeats.

Online Convex Optimization Protocol

This broad setting captures a wide range of convex problems, for instance, when drawing different sample points from a large training set [31, 105]. To that end, we distinguish below two iconic examples of OCO problems which are going to be of individual interest in the sequel.

Convex Minimization

Example 2.1 (Static convex minimization)**.** Consider a convex minimization problem of the general form:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \mathcal{X} \end{aligned} \tag{Opt}$$

where $f \colon \mathcal{X} \to \mathbb{R}$ is a convex function. The notion of "stationarity" refers here to the fact that (Opt) is obtained by the online protocol by assuming that the optimizer faces at each round the *same* convex loss function, i.e., $f_t = f$.

Stochastic Convex Minimization

Example 2.2 (Stochastic convex minimization)**.** A variant of (Opt) with important applications to machine learning, distributed control and data science is the so-called *stochastic* optimization problem:

$$\begin{aligned} &\text{minimize} && f(x) = \mathbb{E}\left[F(x;\omega)\right] \\ &\text{subject to} && x \in \mathcal{X}. \end{aligned} \tag{StochOpt}$$

where $F \colon \mathcal{X} \times \Omega \to \mathbb{R}$ is a stochastic objective defined over a (complete) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $F(\cdot;\omega)$ is assumed convex for all $\omega \in \Omega$. Clearly, (StochOpt) can be seen as a special case of an OCO protocol where the optimizer faces at each round the convex loss function $f_t = F(\cdot;\omega_t)$ with ω_t drawn i.i.d. from Ω at each round.

2.1.2 *Performance evaluation and merit functions*

The most widely used figure of merit in OCO problems is the optimizer's *regret*. Intuitively, this notion compares the average loss incurred by the agent to the minimum loss they could have incurred in hindsight by playing a fixed $x \in \mathcal{X}$. Formally, the *regret* of a policy $X_t \in \mathcal{X}$, $t = 1, 2, \ldots$, against a "benchmark action" $x \in \mathcal{X}$ is defined as

Regret

$$\mathrm{Reg}_x(T) = \sum_{t=1}^{T} [f_t(X_t) - f_t(x)] \tag{2.1}$$

and we define the optimizer's *static* (or *external*) regret (without any benchmark quantifiers) as

$$\mathrm{Reg}(T) = \sup_{x \in \mathcal{X}} \mathrm{Reg}_x(T) = \sup_{x \in \mathcal{X}} \sum_{t=1}^{T} [f_t(X_t) - f_t(x)]. \tag{2.2}$$

With all this in hand, a natural property that the optimizer would like to attain is for their regret to remain "small" over time; this amounts to the requirement:

$$\mathrm{Reg}_x(T) = o(T) \quad \text{for all } x \in \mathcal{X}. \tag{2.3}$$

No-Regret

This, in turn, yields that on average the cumulative loss compared to the best action in hindsight becomes asymptotically non-positive.

For concreteness, we discuss below the implications of attaining no regret in the special cases of static and stochastic minimization problems discussed above. To begin with, if the optimizer is facing (Opt) while deploying an iterative method generating the sequence of actions X_t, $t = 1, 2, \dots$, the regret given by (2.2) becomes

Regret Conversion

$$\mathrm{Reg}(T) = \mathrm{Reg}_{x^*}(T) = \sum_{t=1}^{T} f(X_t) - Tf(x^*) \tag{2.4}$$

with $x^* \in \arg\min_{x \in \mathcal{X}} f$ (assumed here to be nonempty). Now, since f is assumed to be convex, Jensen's inequality shows that the performance (in terms of function values) of the time-averaged sequence

$$\overline{X}_T = \frac{1}{T} \sum_{t=1}^{T} X_t \tag{2.5}$$

is bounded by the optimizer's regret as

Optimality Gap

$$f(\overline{X}_T) - f(x^*) \leq \frac{\mathrm{Reg}(T)}{T} \tag{2.6}$$

In a similar fashion, if the optimizer is facing (StochOpt), we get

Expected Optimality Gap

$$\mathbb{E}\left[f(\overline{X}_T) - f(x^*)\right] \leq \frac{\mathbb{E}\left[\mathrm{Reg}(T)\right]}{T}. \tag{2.7}$$

As a result, in view of (2.6) and (2.7), no-regret policies clearly guarantee an "optimality gap" $f(\cdot) - \min_{x \in \mathcal{X}} f(x)$ that vanishes asymptotically for the associated time-average sequence $\overline{X}_t$.

2.2 VARIATIONAL INEQUALITIES

2.2.1 *Problem setup and examples*

Despite the generality of OCO protocols, there are relevant instances that arise in practice and which necessitate a framework for "optimization beyond minimization". A large class of such problems can be captured by the *variational inequality* (VI) framework:

Variational Inequality Problem

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle A(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X} \tag{VI}$$

where $A \colon \mathcal{X} \to \mathcal{V}^*$ is a single-valued operator, which we call the problem's *defining vector field*. Moreover, for the time being we shall assume that the feasible region $\mathcal{X}$ is a convex and closed subset of $\mathbb{R}^n$. Following [40], we will refer to this problem as

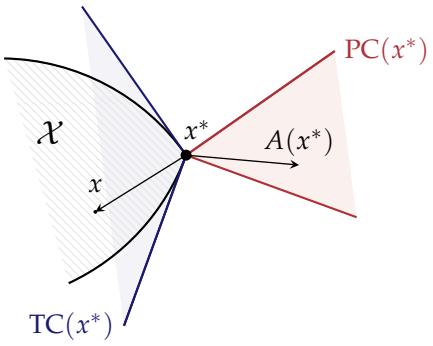


Figure 2.1: Schematic representation of a variational inequality problem: the defining vector field A at a solution x^* of (VI) belongs to the polar cone $\mathrm{PC}(x^*)$ to $\mathcal{X}$ at x^*.

$\mathrm{VI}(\mathcal{X}, A)$ and we will write $\mathcal{X}^* \equiv \mathrm{Sol}(\mathcal{X}, A)$ for its set of solutions.[1] Moreover, to avoid trivialities, we will also assume that the solution set $\mathcal{X}^*$ of (VI) is nonempty and we will reserve the notation x^* for solutions thereof.

Monotone Operators

In terms of blanket requirements, we will assume throughout that A is continuous and *monotone*, i.e.,

$$\langle A(x) - A(x'), x - x' \rangle \geq 0 \quad \text{for all } x, x' \in \mathcal{X}. \tag{Mon}$$

This condition translates the notion of convexity to the language of operators: indeed, if $A = \nabla f$ for some smooth function f, then A satisfies (Mon). For a panoramic overview of monotone operators we refer the reader to Bauschke and Combettes [18]

For illustration purposes, we present some archetypal examples of such problems below:

Minimization Problems

Example 2.3 (Function minimization)**.** If $A = \nabla f$ for some smooth convex function f on $\mathcal{X} = \mathbb{R}^n$, solutions of (VI) coincide with the global minimizers of f, i.e., the solutions of (Opt).

Min-Max Problems

Example 2.4 (Min-max optimization)**.** Suppose that $A = (\nabla_{x_1} f, -\nabla_{x_2} f)$ for some real-valued function $f(x_1, x_2)$ with $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, and $\mathcal{X}_1, \mathcal{X}_2$ convex. If f is convex-concave (i.e., convex in x_1 and concave in x_2), any solution $x^* = (x_1^*, x_2^*)$ of (VI) is a global saddle-point of f, i.e.,

$$f(x_1^*, x_2^*) \leq f(x_1, x_2^*) \quad \text{and} \quad f(x_1^*, x_2^*) \geq f(x_1^*, x_2) \tag{2.8}$$

for all $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$. Problems of this type have attracted considerable interest in the fields of machine learning and artificial intelligence because they constitute the basic optimization framework for GANs [46]. For a series of recent papers focusing on this interplay, see [37, 43, 69, 81, 116] and references therein.

Convex Games

Example 2.5 (Convex Games)**.** A *continuous game in normal form* is defined as follows: Consider a finite set of players $\mathcal{N} = \{1, \dots, N\}$, each with their own

1 In the literature, this formulation of the problem is sometimes referred to as a Stampacchia [40] or "strong" variational inequality [56, 90]. This is to distinguish with the *Minty* or "weak" variational inequality; these two formulations are equivalent when A is monotone, so we will not distinguish between them in the sequel.

action space $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ (convex but possibly not closed). During play, each player selects an action x_i from $\mathcal{X}_i$ with the aim of minimizing a loss determined by the ensemble $x \equiv (x_i; x_{-i}) = (x_1, \dots, x_N)$ of all players' actions. In more detail, writing $\mathcal{X} = \prod_i \mathcal{X}_i$ for the game's total action space, we assume that the loss incurred by the i-th player is $\ell_i(x_i; x_{-i})$, where $\ell_i \colon \mathcal{X} \to \mathbb{R}$ is the player's *loss function*.

In this context, a Nash equilibrium is any action profile $x^* \in \mathcal{X}$ that is *unilaterally stable*, i.e.,

$$\ell_i(x_i^*; x_{-i}^*) \leq \ell_i(x_i; x_{-i}^*) \quad \text{for all } x_i \in \mathcal{X}_i \text{ and all } i \in \mathcal{N}. \tag{NE}$$

In most cases of interest, the players' loss functions are *individually subdifferentiable* on a subset $\mathcal{X}'$ of $\mathcal{X}$ with $\operatorname{ri} \mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{X}$ [51, 102]. This means that there exists a (possibly discontinuous) vector field $A_i \colon \mathcal{X} \to \mathbb{R}^{n_i}$ such that

$$\ell_i(x_i'; x_{-i}) \geq \ell_i(x_i; x_{-i}) + \langle A_i(x), x_i' - x_i \rangle \tag{2.9}$$

for all $x \in \mathcal{X}'$, $x' \in \mathcal{X}$ and all $i \in \mathcal{N}$ [51]. In the simplest case, if ℓ_i is differentiable at x, then $A_i(x)$ can be interpreted as the gradient of ℓ_i with respect to x_i. In turn, this means that Nash equilibria of the game are solutions of $\mathrm{VI}(\mathcal{X}, A)$.

2.2.2 *Merit functions*

Due to the lack of a single objective function the quality of a candidate solution of (VI) becomes much trickier to assess compared to the minimization case. To that end, we start with the unconstrained case, i.e., when $\mathcal{X} = \mathbb{R}^n$. Then (VI) is reduced to the *zero-finding* problem: *Stationarity Problem*

$$\text{Find } \; x^* \in \mathbb{R}^n \; \text{ such that } \; A(x^*) = 0 \tag{Zer}$$

Therefore, a natural performance criterion of a given policy X_t for this case would be to examine how fast $\|A(X_t)\|_*$ converges to 0. [2]

However, if $\mathcal{X}$ is a strict subset of $\mathbb{R}^n$, i.e., when we are facing a genuine constrained problem, then the operator may not necessarily vanish at a solution of (VI). Therefore, we shall need a more general measure in order to be able to capture cases where the solution lies on the border of the domain $\mathcal{X}$.

A popular performance criterion in this context is that of the *restricted merit function*, first introduced in [11, 12]: *Gap Function*

$$\mathrm{Gap}_{\mathcal{C}}(\hat{x}) = \sup_{x \in \mathcal{C}} \langle A(x), \hat{x} - x \rangle, \tag{2.10}$$

where the "*test domain*" $\mathcal{C}$ is a nonempty convex subset of $\mathcal{X}$ [40, 56, 90]. The following proposition generalizes earlier characterizations by [11, 90] and justifies the use of $\mathrm{Gap}_{\mathcal{C}}(x)$ as a merit function for (VI); since every solution of (VI) is a zero of (2.10) and vice versa.

Proposition 2.1 (6). *Let $\mathcal{C}$ be a nonempty convex subset of $\mathcal{X}$. Then: a) $\mathrm{Gap}_{\mathcal{C}}(\hat{x}) \geq 0$ whenever $\hat{x} \in \mathcal{C}$; and b) if $\mathrm{Gap}_{\mathcal{C}}(\hat{x}) = 0$ and $\mathcal{C}$ contains a neighborhood of $\hat{x}$, then $\hat{x}$ is a solution of* (VI).

2 Recent developments on convergence results and rates can be found in [45, 117] and references therein.

Proof. Let $x^* \in \mathcal{X}$ be a solution of (VI) so $\langle A(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$. Then, by monotonicity, we get:

$$\begin{aligned} \langle A(x), x^* - x \rangle &\leq \langle A(x) - A(x^*), x^* - x \rangle + \langle A(x^*), x^* - x \rangle \\ &= -\langle A(x^*) - A(x), x^* - x \rangle - \langle A(x^*), x - x^* \rangle \leq 0, \end{aligned} \tag{2.11}$$

so $\text{Gap}_{\mathcal{C}}(x^*) \leq 0$. On the other hand, if $x^* \in \mathcal{C}$, we also get $\text{Gap}(x^*) \geq \langle A(x^*), x^* - x^* \rangle = 0$, so we conclude that $\text{Gap}_{\mathcal{C}}(x^*) = 0$.

For the converse statement, assume that $\text{Gap}_{\mathcal{C}}(\hat{x}) = 0$ for some $\hat{x} \in \mathcal{C}$ and suppose that $\mathcal{C}$ contains a neighborhood of $\hat{x}$ in $\mathcal{X}$. First, we claim that the following inequality holds:

$$\langle A(x), x - \hat{x} \rangle \geq 0 \quad \text{for all } x \in \mathcal{C}. \tag{2.12}$$

Indeed, assume to the contrary that there exists some $x_1 \in \mathcal{C}$ such that

$$\langle A(x_1), x_1 - \hat{x} \rangle < 0. \tag{2.13}$$

This would then give

$$0 = \text{Gap}_{\mathcal{C}}(\hat{x}) \geq \langle A(x_1), \hat{x} - x_1 \rangle > 0, \tag{2.14}$$

which is a contradiction. Now, we further claim that $\hat{x}$ is a solution of (VI),i.e.,:

$$\langle A(\hat{x}), x - \hat{x} \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{2.15}$$

If we suppose that there exists some $z_1 \in \mathcal{X}$ such that $\langle A(\hat{x}), z_1 - \hat{x} \rangle < 0$, then, by the continuity of A, there exists a neighborhood U' of $\hat{x}$ in $\mathcal{X}$ such that

$$\langle A(x), z_1 - x \rangle < 0 \quad \text{for all } x \in U'. \tag{2.16}$$

Hence, assuming without loss of generality that $U' \subset U \subset \mathcal{C}$ (the latter assumption due to the assumption that $\mathcal{C}$ contains a neighborhood of $\hat{x}$), and taking $\lambda > 0$ sufficiently small so that $x = \hat{x} + \lambda(z_1 - \hat{x}) \in U'$, we get that $\langle A(x), x - \hat{x} \rangle = \lambda \langle A(x), z_1 - \hat{x} \rangle < 0$, in contradiction to (2.12). We conclude that $\hat{x}$ is a solution of (VI), as claimed. □

2.3 LIPSCHITZ REGULARITY

Having described the problems of interest, besides the structural assumption of convexity (or monotonicity for the (VI) context) there are two additional regularity conditions which heavily determine the performance of the respective merit functions of each framework. In what follows, we shall present them in a nutshell. More precisely, given an operator $A : \mathcal{X} \to \mathbb{R}^n$ we have the following definitions:

Bounded Operators

1. A is *bounded*, i.e., there exists some positive constant $G > 0$ such that :

$$\|A(x)\|_* \leq G \text{ for all } x \in \mathcal{X} \tag{Bd}$$

Lipschitz Continuous Operators

2. A is *Lipschitz continous*, i.e., there exists some positive constant $\beta > 0$ such that :

$$\|A(x) - A(x')\|_* \leq \beta\|x - x'\| \quad \text{for all} \quad x, x' \in \mathcal{X} \tag{LC}$$

As said these conditions play a crucial role in determining the performance of the various algorithmic methods at play; this fact will become apparent in Section 2.4.

Now, when $A = \nabla f$ for some convex objective f, (Bd), (LC) give rise to a series of explicit properties for f. Starting with (Bd) one can straightforwardly derive that said property essentially boils to Lipschitz continuity of f,i.e., : *Lipschitz Objectives*

$$|f(x) - f(x')| \leq G\|x - x'\| \quad \text{for all} \quad x, x' \in \mathcal{X} \tag{2.17}$$

On the other hand under (LC), f satisfies the *descent inequality* *Descent Inequality*

$$f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{\beta}{2}\|x' - x\|^2 \quad \text{for all} \quad x, x' \in \mathcal{X}, \tag{2.18}$$

which lies at the core of the success of first order "descent" methods. In particular, we have the following proposition: *Smoothness Properties*

Proposition 2.2. *Assume that $\mathcal{X}$ is a convex and closed subset of $\mathbb{R}^n$ and $f : \mathcal{X} \to \mathbb{R}$ is a continuously differentiable convex function on* int $\mathcal{X}$*. Then, the following statements are equivalent:*

1. *∇f satisfies* (LC)
2. *f satisfies* (2.18)
3. *$\frac{\beta}{2}\|\cdot\|^2 - f$ is a convex function*
4. *$\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \leq \beta\|x - x'\|^2$ for all $x, x' \in \mathcal{X}$*

Finally, a quite interesting equivalence holds whenever $\mathcal{X} = \mathbb{R}^n$ which is known as the Baillon-Haddad theorem [15]. In particular, we have: *Baillon-Haddad Theorem*

Theorem 2.3. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function. Then, the following statements are equivalent:*

1. *∇f satisfies* (LC)
2. *∇f is $1/\beta$- cocoercive*[3]*:*

$$\frac{1}{\beta}\|\nabla f(x) - \nabla f(x')\|_*^2 \leq \langle \nabla f(x) - \nabla f(x'), x - x' \rangle \quad \text{for all} \quad x, x' \in \mathbb{R}^n \tag{2.19}$$

2.4 FIRST-ORDER METHODS

Now we turn our attention towards the respective iterative solution methods. In particular, our focal point would be the so-called *first-order methods*, i.e., methods that require at each iteration access on a first order/ gradient feedback. The surge of recent breakthroughs in machine learning and artificial intelligence has reaffirmed the prominence of these methods in solving large-scale optimization

3 For a panoramic overview of cocoercive operators we refer the reader to [18]

problems. One of the main reasons for this is that the computation of higher-order derivatives of functions with thousands – if not millions – of variables quickly becomes prohibitive; another is that gradient calculations are typically easier to distribute and parallelize, especially in large-scale problems. In view of this, first-order methods have met with prolific success in many diverse fields, from machine learning and signal processing to wireless communications, nuclear medicine, and many others [30, 103, 108]. In what follows, we present the main structure of these methods combined with the respective optimal lower bounds.

2.4.1 *Oracle mechanism and feedback*

From an algorithmic point of view, we aim to solve (Opt) and/or (VI) by using iterative methods that require access to a *stochastic first-order oracle* (SFO) [89]. This means that, at each stage of the process, the optimizer can query a black-box mechanism that returns an estimate of the objective's gradient (or subgradient) at the queried point. Formally, when called at $x \in \mathcal{X}$, an SFO is assumed to return a random (dual) vector $V(x;\omega) \in \mathcal{V}^*$ where ω belongs to some (complete) probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In practice, the oracle will be called repeatedly at a (possibly) random sequence of points $X_t \in \mathcal{X}$ generated by the algorithm under study. Thus, once X_t has been generated at stage t, the oracle draws an i.i.d. sample sample $\omega_t \in \Omega$ and returns the dual vector:

First Order Oracle

$$V_t \equiv V(X_t;\omega_t) = A(X_t) + U_t \tag{SFO}$$

with $U_t \equiv U(X_t;\omega_t) \in \mathcal{V}^*$ denoting the "measurement error" of the oracle. In terms of measurability, we will write $\mathcal{F}_t$ for the history (natural filtration) of X_t; in particular, X_t is $\mathcal{F}_t$-adapted, but ω_t, V_t and U_t are not.

First Order Oracle's Statistics

Finally, we will also make the following statistical assumptions. First, we shall assume that (SFO) is an unbiased estimator:

$$\mathbb{E}[U_t \,|\, \mathcal{F}_t] = 0 \tag{2.20}$$

Moreover, we shall assume that for some (known) $q \in (2, +\infty]$ we have:

$$\|U_t\|_{\mathcal{L}^q,*} = \mathbb{E}[\|U_t\|_*^q]^{1/q} \leq \sigma^2 \quad \text{for all } t = 1, 2, \dots \tag{2.21}$$

For concreteness, we will refer to the oracles with $\sigma = 0$ as "perfect" – since, in that case, $U_t = 0$ for all t almost surely. Otherwise, if $\|U_t\|_{\mathcal{L}^q,*} > 0$ the noise will be called *persistent* and the model will be called *stochastic*.

2.4.2 *Lower bounds*

With all this in hand, the first question that arises is what is the worst performance that the optimizer may expect and how is this influenced by the different feedback and regularity conditions at play. The answer to the above question is formally stated by the notion of *worst-case lower bounds* and differs depending on the respective setting. So, we shall investigate each setting individually.

- *Online Convex Optimization/ Stochastic Minimization*: We begin with the online convex optimization framework. More precisely, under (Bd) the regret optimal lower bound is

$$\mathrm{Reg}(T) = \Theta(1/\sqrt{T}) \tag{2.22}$$

Regret Lower Bound

Moreover, (LC) does not help the optimizer to improve upon this lower bound [1].

- *Static Convex Minimization*: Now, we turn our attention towards the particular case of (Opt). We distinguish the deterministic ($\sigma = 0$) and the (purely) stochastic ($\sigma > 0$) instances of (SFO). Starting with the deterministic one, the sub-optimality gap for first-order methods with perfect gradient input possesses a "worst-case" guarantee

Convex Minimization Lower Bounds

$$f(X_T) - f(x^*) = \Omega(1/\sqrt{T}) \tag{2.23}$$

under (Bd). This guarantee is improved significantly, i.e.,

$$f(X_T) - f(x^*) = \Omega(1/T^2) \tag{2.24}$$

whenever (LC) kicks in [89]. On the other hand, if the optimizer has only access to stochastic gradients (as is often the case in machine learning and distributed control), the corresponding lower bound for the expected sub-optimality gap is

$$\mathbb{E}[f(X_T) - f(x^*)] = \Omega(1/\sqrt{T}) \tag{2.25}$$

For details we refer the reader to [30, 86, 89].

- *Variational inequalities*: Finally, we describe the worst case guarantees for the generic framework of (VI). In doing so, if the optimizer has access to a perfect (SFO) oracle then the respective optimal lower bound for the restricted merit function (2.10) under (Bd) is:

Variational Inequality Lower Bound

$$\mathrm{Gap}_{\mathcal{C}}(X_T) = \Omega(1/\sqrt{T}), \tag{2.26}$$

while under (LC) a lower bound of $\Theta(1/T)$ is achievable [83, 84]; the latter illustrates also a significant gap between VI's and the static smooth minimization setting. Finally, for a purely stochastic (SFO) the respective lower bound relative to the restricted merit function (2.10) would be that of $\Theta(1/\sqrt{T})$ under (Bd).

2.5 FIRST-ORDER METHODS FOR ONLINE CONVEX OPTIMIZATION

2.5.1 *Gradient descent and its primal-dual variant*

For OCO, the most popular first order methods are the so called *greedy/lazy* (projected) gradient descent algorithms. In what follows, we describe these methods in detail.

To start with, the greedy version is defined formally as:

Gradient Descent

$$X_{t+1} = \mathrm{pr}_{\mathcal{X}}(X_t - \gamma_t V_t) \tag{GD}$$

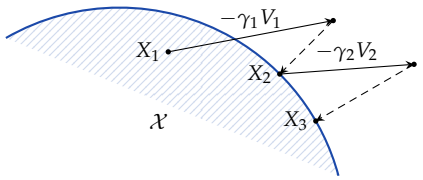


Figure 2.2: Schematic representation of (projected) gradient descent.

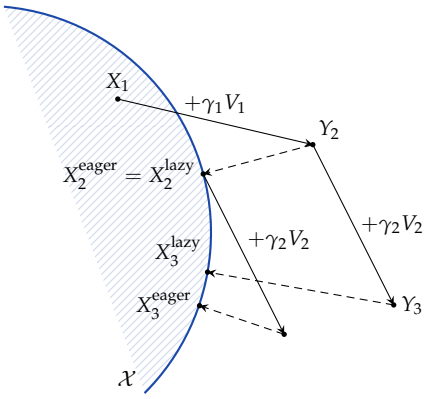


Figure 2.3: Lazy vs. ordinary gradient descent.

In the above $\mathrm{pr}_{\mathcal{X}}(x) = \arg\min_{x' \in \mathcal{X}} \|x' - x\|$ denotes the euclidean projection onto the convex and closed feasible domain $\mathcal{X}$, $\gamma_t > 0$ is the method's step-size and V_t is the (SFO) feedback at X_t. We refer to (GD) also as *greedy gradient descent* in order to distinguish it from the so-called *lazy* variant [105] which is defined by the following recursion:

Lazy Gradient Descent

$$\begin{aligned} Y_{t+1} &= Y_t - \gamma_t V_t \\ X_{t+1} &= \mathrm{pr}_{\mathcal{X}}(Y_{t+1}) \end{aligned} \tag{LGD}$$

A different perspective of the above method is given by the so-called *dual averaging* scheme, originally introduced by Nesterov in [91] and further developed in [115]. This is formally given by the following recursion:

Dual Averaging

$$\begin{aligned} Y_{t+1} &= Y_t - V_t \\ X_{t+1} &= \mathrm{pr}_{\mathcal{X}}(\eta_{t+1} Y_{t+1}) \end{aligned} \tag{DA}$$

More precisely, the critical difference between (LGD) and (DA) is that in the latter the learning rate η_t changes its role. In particular, in (DA) η_t acts as a post-multiplier over the ensemble aggregation of V_t instead of allocating a specific weight on each individual V_t. As we discuss in the sequel this key feature of (DA) would enable us to deal with unbounded feasible domains $\mathcal{X}$.

2.5.2 *Performance guarantees*

We now proceed to describe the regret minimization guarantees of the family of algorithms presented in Section 2.5.1. More precisely, if (GD)/(LGD) are run with

a "horizon"-dependent step-size policy $\gamma_t \equiv 1/\sqrt{T}$[4] we have the following the proposition [105, 122]:

Regret of Descent Variants (Horizon-dependent step-size)

Proposition 2.4. *Assume that X_t are the iterates of (GD) or (LGD) run with a step-size $\gamma_t = 1/\sqrt{T}$ and a "perfect" oracle feedback. Then, if f_t satisfies (2.17). for all $t = 1, 2, \ldots, T$ with $\sup_t \|\nabla f_t(x)\|_*^2 \leq G$, we have:*

$$\frac{1}{T}\left[\sum_{t=1}^{T} f_t(X_t) - \sum_{t=1}^{T} f_t(x)\right] = \mathcal{O}\left(\frac{\|X_1 - x\|^2 + G^2}{\sqrt{T}}\right) \text{ for all } x \in \mathcal{X} \tag{2.27}$$

Some comments concerning the particular step-size are in order. More precisely, this step-size policy is based on the idea on dividing the infinite play into epochs (or time-windows) of length T. Hence, the optimizer practically applies a constant, within the time window $[1, T]$ and then repeats the same idea for the next window $[T, 2T]$ and the procedure repeats to infinity. Moreover, if the feasible domain $\mathcal{X}$ is a compact set, one may apply a "dynamic" step-size $\gamma_t \propto 1/\sqrt{t}$ and derive an "any-time" regret bound in contrast to that of Proposition 2.4. This is described by the following proposition.

Regret of Descent Variants (Dynamic step-size)

Proposition 2.5. *Assume that $\mathcal{X}$ is a compact set and let X_t be the iterates of (GD) or (LGD) run with $\gamma_t \propto 1/\sqrt{t}$ and a "perfect" oracle feedback. Then, if f_t satisfy (2.17) with $\sup_t \|\nabla f_t(x)\|_*^2 \leq G^2$, we have:*

$$\frac{1}{T}\left[\sum_{t=1}^{T} f_t(X_t) - \sum_{t=1}^{T} f_t(x)\right] = \mathcal{O}\left(\frac{\operatorname{diam}\mathcal{X} + G^2}{\sqrt{T}}\right) \text{ for all } x \in \mathcal{X} \tag{2.28}$$

where $\operatorname{diam}\mathcal{X} = \sup_{x,x' \in \mathcal{X}} \|x - x'\|$.

Moving forward our next step is to illustrate the respective regret guarantees for (DA). Similarly with the above guarantees we have the following result for the (DA) [115]:

Regret of Dual Averaging

Proposition 2.6. *Assume that X_t are the iterates of (DA) run with a learning rate $\eta_t \propto 1/\sqrt{t}$ and a "perfect" oracle feedback. Then, if f_t satisfy (2.17) with $\sup_t \|\nabla f_t(x)\|_*^2 \leq G^2$, we have:*

$$\frac{1}{T}\left[\sum_{t=1}^{T} f_t(X_t) - \sum_{t=1}^{T} f_t(x)\right] = \mathcal{O}\left(\frac{\|x\|^2 + G^2}{\sqrt{T}}\right) \text{ for all } x \in \mathcal{X} \tag{2.29}$$

An important difference between Proposition 2.5 and Proposition 2.6 is that in the latter no compactness-or rather boundedness- assumption for the domain $\mathcal{X}$ is required.

2.5.3 *Sub-optimality for smooth minimization and accelerated methods*

In this section we shall investigate the particular case of (Opt) in a more detailed manner in accordance to the optimal worst case lower bounds (cf. Section 2.4.2).

4 The choice of such a step-size assumes a prior knowledge of the horizon of the process and is also referred to as "doubling" trick [105]

In doing so, the first candidate would be the (GD) methods presented in Section 2.5.1. More precisely, we first describe their performance under the different regularity conditions (Bd) and (LC). A preliminary result under (Bd) can be obtained via a straightforward adaptation of Propositions 2.5 and 2.6; more precisely this yields that under (Bd):

$$f(\overline{X}_t) - f(x^*) = \mathcal{O}(1/\sqrt{T}) \tag{2.30}$$

with $\overline{X}_T$ denoting the time average of the (GD)/(LGD) and (DA) iterates run with a step-size policy $\gamma_t \propto 1/\sqrt{t}$. Hence, the generic (GD) algorithms exhibit an optimal convergence rate within this class of objectives.

That said, the situation changes drastically under (LC). For that particular case, (GD) and (LGD) run with a constant step-size $\gamma_t \equiv \gamma \leq 1/\beta$ guarantees a performance rate of order $\mathcal{O}(1/T)$. This result confirms the sub-optimality of (GD) family of methods for smooth deterministic minimization problems, since its performance does not match the iconic $1/T^2$ lower bound. This $1/T^2$ rate was first achieved by Nesterov in his seminal paper [88]. This algorithm has since generated an immense literature with several hallmark contributions like the fast iterative shrinkage-thresholding algorithm (FISTA) method,[21], for composite minimization problems and many others. More precisely, following [13] we consider the improved interior gradient algorithm (IGA) algorithm:

Acceleration Schemes

$$\begin{aligned} Y_t &= (1-\lambda_t)X_t + \lambda_t Z_t \\ Z_{t+1} &= \mathrm{pr}_{\mathcal{X}}(Z_t - \frac{\lambda_t}{\beta} V_t) \\ X_{t+1} &= (1-\lambda_t)X_t + \lambda_t Z_{t+1} \end{aligned} \tag{IGA}$$

with the weight sequence λ_t being defined recursively as follows:

$$\frac{1-\lambda_{t+1}}{\lambda_{t+1}^2} = \frac{1}{\lambda_t^2} \tag{2.31}$$

The crucial difference of (IGA) is the particular averaging part that serves as an acceleration mechanism of the (GD) template. This is described formally by the following proposition.

Acceleration Rate

Proposition 2.7. *Assume that X_t are the iterates of* (IGA) *run with a step-size given by* (2.31)*. Then, if f satisfies* (LC) *we have:*

$$f(X_t) - f(x^*) \leq \frac{2\beta\|X_1 - x^*\|^2}{T^2} \tag{2.32}$$

Variants of this method can be also found in [49].

2.6 OPTIMAL METHODS: THE VARIATIONAL INEQUALITY CASE

2.6.1 *Extra-Gradient method and its primal-dual variant*

Now we turn our attention towards defining optimal iterative methods for (VI). Perhaps the most widely used solution method for VIs is the EG algorithm of

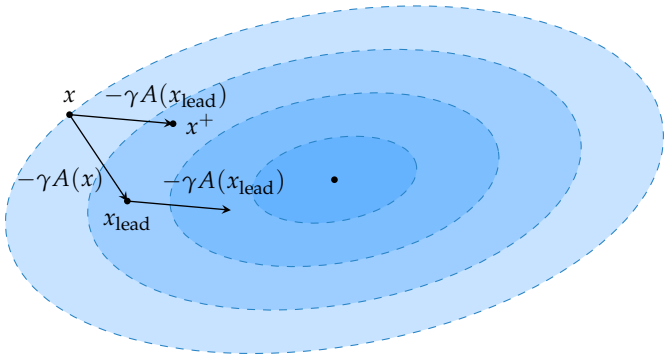


Figure 2.4: Schematic representation of the extra-gradient algorithm.

Korpelevich (1976) and its variants [74, 99, 100]. This algorithm has a rich history in optimization, and it has recently attracted considerable interest in the fields of machine learning and AI, see e.g., [33, 37, 43, 52, 53, 81, 82] and references therein. In its simplest form, for problems with closed and convex domains, the algorithm proceeds recursively as

$$\begin{aligned} X_{t+1/2} &= \mathrm{pr}_x(X_t - \gamma_t V_t) \\ X_{t+1} &= \mathrm{pr}_{\mathcal{X}}(X_t - \gamma_t V_{t+1/2}) \end{aligned} \tag{EG}$$

In a nutshell (EG) suggests first to generate a leading state $X_{t+1/2}$ by taking a "gradient" step as usual. Then, instead of continuing from $X_{t+1/2}$, (EG) samples $V_{t+1/2}$ and goes back to the original state X_t in order to generate a new state X_{t+1} via a "gradient" step along the direction of $V_{t+1/2}$. *Extra-Gradient*

Let us now present its primal-dual counterpart, firstly introduced by Nesterov in [90]. In particular, the (euclidean based) dual extrapolation (DualX) method is given via the following recursive formula: *Dual Extrapolation*

$$\begin{aligned} X_{t+1/2} &= \mathrm{pr}_x(X_t - \gamma_t V_t) \\ Y_{t+1} &= Y_t - V_{t+1/2} \\ X_{t+1} &= \mathrm{pr}_{\mathcal{X}}(\gamma_{t+1} Y_{t+1}) \end{aligned} \tag{DualX}$$

In turn, the (DualX) template hinges on a combination of the (GD) and (DA) methods. In particular, it suggests the following updating rule: first generate a leading state $X_{t+1/2}$ by taking a "gradient" step as in (EG) and again samples $V_{t+1/2}$. Then, the method aggregates these feedbacks and finally the method's update is obtained by applying a dual averaging step.

2.6.2 *Performance guarantees*

Building on the templates of (EG) and (DualX) in this section we present the performance guarantees (in terms of the restricted merit function (2.10)) under the light of the different regularity conditions (Bd) and/or (LC). Starting with the (EG) template, we have the proposition for the case where A is not necessarily Lipschitz continuous:

Extra-Gradient Guarantees (Stochastic)

Proposition 2.8 (Juditsky et al. [56]). *Assume that $X_t, t = 1, 1/2, \dots$ are the iterates of (EG) with an oracle satisfying (SFO) with $\sigma^2 > 0$ and A satisfies (Bd). Moreover, let $\overline{X}_T = \left[\sum_{t=1}^T \gamma_t\right]^{-1} \sum_{t=1}^T \gamma_t X_{t+1/2}$, let $\mathcal{C}$ be a compact neighbourhood of a solution of the (VI) and set $D^2 = \sup_{x\in\mathcal{C}} \|x - X_1\|^2$. Then, if (EG) run with a decreasing (deterministic) step-size γ_t satisfies the following estimate:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{D^2 + \left[G^2 + \sigma^2\right] \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t} \tag{2.33}$$

In particular, if (EG) is run with $\gamma_t \propto 1/\sqrt{t}$, then

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{2.34}$$

We distinguish the "perfect" (SFO) case, i.e., $\sigma^2 = 0$, where the particular influence of the respective regularity conditions becomes more apparent. In particular, this is illustrated by the following result.

Extra-Gradient Guarantees (Deterministic)

Proposition 2.9 (Nemirovski [85]). *Assume that X_t, $X_{t+1/2}$ are the iterates of (EG) with a "perfect" (SFO) and A satisfies (Bd). Let us denote $\overline{X}_T = \left[\sum_{t=1}^T \gamma_t\right]^{-1} \sum_{t=1}^T \gamma_t X_{t+1/2}$, $\mathcal{C}$ is a compact neighbourhood of a solution of the (VI) and $D = \sup_{x\in\mathcal{C}} \|x - X_1\|^2$. Then, the following hold:*

1. *Under (Bd) then,*

$$\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) \leq \frac{D + G^2 \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t} \tag{2.35}$$

   In particular, if $\gamma_t \propto 1/\sqrt{t}$, then $\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) = \mathcal{O}(1/\sqrt{T})$

2. *Under (LC) and $0 < \inf_t \gamma_t \leq \gamma_t \leq 1/\beta$ then,*

$$\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) \leq \frac{D}{2T \inf_t \gamma_t} \tag{2.36}$$

On the other the hand, for the (DualX) template we may obtain similar convergence rate guarantees by consider the time average, $1/T \sum_{t=1}^T X_{t+1/2}$, as the method's output. Formally, we have the following proposition.[5]

Dual-Extrapolation Guarantees (Stochastic)

Proposition 2.10. *Assume that X_t, $X_{t+1/2}$ are the iterates of (DualX) with an oracle satisfying (SFO) with $\sigma^2 > 0$. Moreover, $\overline{X}_T = 1/T \sum_{t=1}^T X_{t+1/2}$ and $\mathcal{C}$ is a compact neighbourhood of a solution of the (VI) and $D^2 = \sup_{x\in\mathcal{C}} \|x - X_1\|^2$. Then, if (DualX) is run with a (deterministic) decreasing step-size γ_t the following holds:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{D + \left[G^2 + \sigma^2\right] \sum_{t=1}^T \gamma_t}{T} \tag{2.37}$$

In particular, if (DualX) is run with $\gamma_t \propto 1/\sqrt{t}$ then

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}). \tag{2.38}$$

5 Proposition 2.10 hinges on the methodology of. the Dual-Extapolation method of [91]. However, we are not aware of a specific paper which provides a proof for it. We revisit this from a more general point of view in Chapter 7.

Now in the same spirit as for the (EG) for the deterministic case we obtain the respective range of rates as in Proposition 2.11:

Dual- Extrapolation Guarantees (Deterministic)

Proposition 2.11 (Nesterov [91]). *Assume that X_t, $X_{t+1/2}$ are the iterates of* (DualX) *with a "perfect"* (SFO). *Moreover, let us denote $\overline{X}_T = 1/T \sum_{t=1}^{T} X_{t+1/2}$, $\mathcal{C}$ is a compact neighbourhood of a solution of the* (VI) *and $D^2 = \sup_{x \in \mathcal{C}} \|x - X_1\|^2$. Then, the following hold:*

1. *Under* (Bd) *then,*

$$\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) \leq \frac{D + G^2 \sum_{t=1}^{T} \gamma_t}{T} \tag{2.39}$$

   In particular, if and $\gamma_t \propto 1/\sqrt{t}$ then $\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) = \mathcal{O}(1/\sqrt{T})$.

2. *Under* (LC) *and $\gamma_t \leq 1/\beta$ then,*

$$\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) \leq \frac{D}{T} \tag{2.40}$$

As one may observe the set of results presented above rely their success on prior knowledge of the Lipschitz modulus of the associated operator. In what follows, we describe the state-of-the-art methods which transcend this restriction.

2.7 ADAPTIVE METHODS

Having described the performance guarantees under different regularity conditions and step-sizes, we move forward by introducing a range of "adaptive" methods that automatically detect the level of regularity in the problem and the quality of the oracle. In particular, adaptivity of a method refers (at least) to two different scenarios:

- The method automatically adjusts its performance to parameters within a fixed operator class (Lipschitz/Hölder smoothness and the like).
- The method automatically detects the respective Lipschitz modulus at hand and exhibits (optimal) rate interpolation guarantees between different classes- for example between non-smooth and smooth objectives etc.[6]

Of course, one may straightforwardly recognize the fact that the second type properly includes the other two. In Section 2.7.1 and Section 2.7.2, we squarely focus on the latter. Moreover, in what will follow we address each framework individually.

2.7.1 *The minimization case*

We assume first that the optimizer is facing an unconstrained (Opt) version and/or (StochOpt). For this particular framework one may show that the (GD) template run with the adaptive step-size of the form:

AdaGrad Step-Size

$$\gamma_t = \frac{1}{\sqrt{\sum_{j=1}^{t} \|V_j\|_*^2}} \tag{2.41}$$

6 The methods that satisfy this property are also denoted in the literature as *universal*.

achieve the following convergence rate:

- *For the deterministic case,* (GD) run with the step-size policy (2.41) interpolates between $\mathcal{O}(1/\sqrt{T})$ for non-smooth and stochastic regimes and $\mathcal{O}(1/T)$ whenever smoothness kicks in.
- For the stochastic case, guarantees an $\mathcal{O}(1/\sqrt{T})$ under (Bd) and $\mathcal{O}(\frac{\beta}{T} + \frac{\sigma}{\sqrt{T}})$ under (LC).

Remark 2.1. This iterative scheme is often refered as adaptive inverse-norm-squared (AdaNorm) and is a simplified variant of the general AdaGrad firstly introduced in [39, 76].

AdaGrad Guarantees

More precisely, the following result describes formally the above properties for (GD). For detailed proof we defer the reader to [67].

Proposition 2.12 (Levy et al. [67]). *Assume X_t are the iterates of* (GD) *run with adaptive step-size policy* (2.41) *and an oracle feedback of the form* (SFO) *and $\overline{X}_T = 1/T \sum_{t=1}^{T} X_t$. Moreover we assume that $\sup_{t\in\mathbb{N}} \|X_t - x^*\| \leq D$. Then the following hold:*

1. *Under* (Bd) *we have:*

$$\mathbb{E}\left[f(\overline{X}_T) - f(x^*)\right] \leq \frac{D^2(G+\sigma)}{2\sqrt{T}} \tag{2.42}$$

2. *Under* (LC) *we have:*

$$\mathbb{E}\left[f(\overline{X}_T) - f(x^*)\right] \leq \frac{\beta D^2}{T} + \frac{\sigma D}{\sqrt{T}} \tag{2.43}$$

The above result indicates that even if (GD) is run with an adaptive step-size this does not seem to match the worst-case $1/T^2$ lower bound remaining sub-optimal for smooth objectives. In order to overcome this and achieve an adaptive optimal rate interpolation from $\mathcal{O}(1/\sqrt{T})$ to $\mathcal{O}(1/T^2)$, more elaborate schemes are required.

The first result of this kind is done by Nesterov in [92]; in particular it shown that under "perfect" oracle feedback an optimal rate interpolation for objectives with gradient whose variance satisfy: For some $\beta > 0$ and $q \in [0,1]$[7]:

$$\|\nabla f(x) - \nabla f(x')\|_* \leq \beta \|x - x'\|^q \quad \text{for all} \quad x, x' \in \mathbb{R}^n \tag{2.44}$$

That said, since this result requires a perfect oracle feedback in an essential manner er we shall not dive into more detail. A different approach which captures adaptivity even for noisy settings is to mimic the idea of (2.41); a step-size that is updated "on the fly". That idea is incarnated by AcceleGrad [67] and Unixgrad [60] methods for the unconstrained and the constrained case. In particular, for $\mathcal{X} = \mathbb{R}^n$ the Accelegrad method suggests:

AcceleGrad Scheme

$$\begin{aligned} X_{t+1} &= \lambda_t Z_t + (1-\lambda_t) Y_t \\ Z_{t+1} &= \mathrm{pr}_{\mathcal{K}}(Z_t - \alpha_t \gamma_t V_t) \\ Y_{t+1} &= X_{t+1} - \gamma_t V_t \end{aligned} \tag{AcceleGrad}$$

7 Of course (2.44) includes the classes of (Bd) and (LC) as extreme cases for $q = 0$ and $q = 1$ respectively.

Some notational comments are here in order to describe each part of the method. In particular, $\mathcal{K}$ is a convex and compact subset of $\mathbb{R}^n$ and denotes a "domain of interest",i.e., an initial speculation of a subset where the global minimizer lives. Moreover, D denotes the diameter of $\mathcal{K}$. In terms of the weighting sequences we set $\alpha_t = t$, $\lambda_t = 1/\alpha_t$. Having all this in hand, the method's step-size step-size is defined:

$$\gamma_t = 2D \left[\sqrt{\theta^2 + \sum_{j=1}^{t} \alpha_j^2 \|V_j\|_*^2} \right]^{-1} \tag{2.45}$$

The following result describes the precise convergence rates of (AcceleGrad); for details we refer the reader to [67]. *AcceleGrad Guarantees*

Proposition 2.13 (Levy et al. [67]). *Assume that Y_t are the iterates of* (AcceleGrad) *and let $\overline{Y}_T = \left[\sum_{t=1}^T \alpha_t\right]^{-1} \sum_{t=1}^T \alpha_t Y_t$ with oracle feedback satisfying* (SFO). *Then, the following hold:*

1. If $\sigma = 0$, *then*
   - *Under* (Bd) *we have:*

$$f(\overline{Y}_T) - f(x^*) \leq \frac{GD\sqrt{\log T}}{\sqrt{T}} \tag{2.46}$$

   - *Under* (LC) *we have:*

$$f(\overline{Y}_T) - f(x^*) \leq \frac{DG^2 + \beta D^2 \log(\beta D/G)}{T^2} \tag{2.47}$$

2. If $\sigma > 0$, *then*

$$\mathbb{E}\left[f(\overline{Y}_T) - f(x^*)\right] \leq \frac{GD\sqrt{\log T}}{\sqrt{T}} \tag{2.48}$$

Now, we turn our attention towards the constrained setting and the (UniXGrad) method, which hinges on the (EG) template. More precisely, this is given by the following: *UniXGrad Scheme*

$$\begin{aligned} X_{t+1/2} &= \mathrm{pr}_{\mathcal{X}}(X_t - \alpha_t \gamma_t V_t) \\ X_{t+1} &= \mathrm{pr}_{\mathcal{X}}(X_t - \alpha_t \gamma_t V_{t+1/2}) \end{aligned} \tag{UniXGrad}$$

The crucial difference with generic (EG) is that V_t and $V_{t+1/2}$ are the oracle queries for the gradient evaluated at the averaged points:

$$\overline{X}_t = \frac{\alpha_t X_t + \sum_{j=1}^{t-1} \alpha_j X_{j+1/2}}{\sum_{j=1}^{t} \alpha_j} \quad \text{and} \quad \overline{X}_{t+1/2} = \frac{\sum_{j=1}^{t} \alpha_j X_{j+1/2}}{\sum_{j=1}^{t} \alpha_j} \tag{2.49}$$

with $\alpha_t = t$. Having induced this acceleration mechanism in the (EG) routine, one may obtain the first result concerning the universal properties of (UniXGrad) for the deterministic framework. *UniXGrad Guarantees*

Proposition 2.14 (Kavis et al. [60]). *Assume that X_t $t = 1, 1/2, \dots$ are the iterates of* (UniXGrad) *under an oracle of the form* (SFO). *Then, we have the following:*

1. *If f satisfies* (Bd)*, then,*

$$\mathbb{E}\left[f(\overline{X}_{T+1/2}) - f(x^*)\right] \leq \frac{6D}{T^2} + \frac{14\sigma D}{\sqrt{T}} \tag{2.50}$$

2. *If f satisfies* (LC)*, then,*

$$\mathbb{E}\left[f(\overline{X}_{T+1/2}) - f(x^*)\right] \leq \frac{224\sqrt{14}D^2L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}} \tag{2.51}$$

An important remark concerning Proposition 2.14 is that the compactness assumption for the feasible region $\mathcal{X}$ is crucial for establishing the desired agnostic rate interpolation.

2.7.2 *The variational inequality case*

Now we move forward towards adaptive methods for (VI). To that end a reasonable candidate would be that of the (EG). Indeed, in [14] a novel adaptive step-size is proposed for constrained (VI) problems in the following manner: If $\mathcal{X}$ is convex and compact with diam $\mathcal{X} = D$, then, Bach-Levy in [14] propose:

Universal Extra-Gradient

$$\gamma_t = \frac{2D}{\sqrt{\theta^2 + \sum_{j=1}^{t-1} Z_j^2}} \tag{2.52}$$

with $\theta > 0$ being an arbitrarily chosen positive constant and Z_j^2:

$$Z_j^2 = \frac{\|X_{j+1/2} - X_j\|^2 + \|X_{j+1/2} - X_{j+1}\|^2}{\gamma_j^2} \tag{2.53}$$

As it becomes apparent the (2.53) is the crucial ingredient of the adaptive step-size (2.52). In terms of convergence rate guarantees for the stochastic case [14] provides us the following result:

Universal Extra-Gradient Guarantees

Proposition 2.15 (Bach and Levy [14])**.** *Assume that X_t, $X_{t+1/2}$ are the iterates of* (EG) *un with the adaptive step-size policy* (2.52) *and an oracle feedback of the form* (SFO)*. Moreover, assume that $\overline{X}_T = 1/T\sum_{t=1}^{T} X_{t+1/2}$ and $\mathcal{C}$ is a convex and compact neighbourhood of a solution x^* of the* (VI)*. Then, the following hold:*

1. *If A satisfies* (Bd)*, then*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{\alpha D(G+\sigma)\sqrt{\log T}}{\sqrt{T}} \tag{2.54}$$

2. *If A satisfies* (Bd) *and* (LC)*, then*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{\alpha GD + \alpha^2\beta D^2 + \beta D^2 \log \beta D/\theta_0}{T} + \frac{\alpha}{\sqrt{T}} \tag{2.55}$$

The analysis of Proposition 2.15 is that in order to achieve rate adaptivity, even for the case of perfect oracle feedback hinges on the following limitations:

- Compactness of the feasible domain $\mathcal{X}$.
- The associated operator A should satisfy simultaneously both (Bd) and (LC).

As a prelude of our contributions our general beyond Lipschitz analysis will allow us to drop both these restrictions.

3

BEYOND LIPSCHITZ REGULARITY

 # This section incorporates material from the papers [6–8]

THROUGHOUT this chapter we focus on extending the Lipschitz regularity conditions presented in Section 2.3. Even though boundedness and Lipschitz continuity conditions (Bd) and (LC) appear to be fairly mild, they may fail to hold in a wide array of practical applications. These unboundedness issues also arise for the case of bounded domains. Indeed, consider as a toy example the 1- dimensional minimization objective:

$$f(x) = -\log x \quad \text{for} \quad x > 0. \tag{3.1}$$

In that case, since $\nabla f(x) = -\frac{1}{x}$, one may straightforwardly verify that ∇f remains unbounded for all positive intervals that include the origin; so it fails to satisfy both (Bd) and (LC). The main objective would be to design efficient definitions which are able to account for possible "blow ups" of the associated operators. In doing so, one should apply more "geometry aware" toolkits than the standard geometry-blind Euclidean setup of the previous chapter. Therefore in Section 3.2 we present two frameworks of that kind. We first describe the notion of a regularization function (or regularizer for short) along with the associated Bregman divergence. The Bregman divergence will serve in the sequel as a generalized distance function surrogate; despite the fact that it is not a distance function per se (it does not satisfy neither symmetry nor the triangle inequality). Moreover, drawing arguments from differential geometry we provide an alternative approach based on the notion of a Finsler metric. This framework allows us to induce families of local[1] norms over the ambient space which are able to capture the geometry of the feasible region in a more efficient way. Armed with these mathematical tools, we introduce in Section 3.3 and Section 3.4 the main classes of objectives that transcend the traditional Lipschitz regularity conditions. To motivate all the above, we first present some prominent examples of widely studied problems with "gradient singularities" in Section 3.1.

3.1 MOTIVATING EXAMPLES

3.1.1 *Poisson Inverse Problems*

Poisson Inverse Problems

Many problems in machine learning and the imaging sciences focus on the reconstruction of an unknown object from a set of imperfect observations (e.g., noisy 2D

1 Local for this context refers to the fact that the said family of norms dependents on the point upon which it is evaluated.

cross-sections of a 3D object). This is especially true in the fields of emission tomography and optical/infrared astronomy, where images are obtained by counting particles (usually photons) reaching a detector. In this case, factors such as fluorescence emissions, radioactive decay and thermal noise can severely affect particle counts, typically by introducing Poisson-distributed errors in the measurement process [22].

Mathematically, inverse problems of this kind boil down to solving linear systems of the form

$$y = Hx + z \tag{3.2}$$

where:

- $x \in \mathbb{R}^n_+$ is the object under study (a signal, image, ...).
- $y \in \mathbb{R}^m_+$ is the observed data (usually $m \ll n$).
- The *kernel matrix* $H \in \mathbb{R}^{m\times n}_+$ is a representation of the data-gathering protocol and is typically ill-conditioned (e.g., a Toeplitz matrix in the case of image deconvolution problems).
- $z \in \mathbb{R}^m$ is the noise affecting the measurements.

When data points are obtained by means of a counting process, measurements can be modeled as Poisson random variables of the form $y_j \sim \text{Pois}(Hx)_j$. [2] Then, up to an additive constant, the log-likelihood of $x \in \mathbb{R}^n$ given an observation $y \in \mathbb{R}^m$ will be

$$\ell(x; y) = -\sum_{j=1}^{m} \left[y_j \log \frac{y_j}{(Hx)_j} + (Hx)_j - y_j \right]. \tag{3.3}$$

Hence, obtaining a maximum likelihood estimate for x leads to the archetypal *Poisson inverse problem*:

$$\begin{aligned} &\text{minimize} && f(x) \equiv D_{KL}(y, Hx), \\ &\text{subject to} && x \in \mathbb{R}^n_+, \end{aligned} \tag{PIP}$$

where $D_{KL}(p, q) = \sum_{j=1}^{m} [p_j \log(p_j/q_j) + q_j - p_j]$ denotes the generalized KL divergence on $\mathbb{R}^m_+$.

In many cases of practical interest, measurements arrive in distinct batches over time – e.g., as sequential optical sections in microscopy and tomography. Moreover, due to the large numbers of pixels/voxels involved (a typical range of values for m is between 10^6 and 10^7), gradients of f are very costly to compute; as such, optimization methods that rely on accurate gradient data are difficult to apply in this setting. Accordingly, a natural workaround to this obstacle is to exploit the online nature of the measurement process, model (PIP) as an *online* optimization problem, and then to use an online-to-batch conversion to get a candidate solution [105].

On the downside, this online optimization analysis crucially requires the loss functions faced by the optimizer to be Lipschitz continuous, and this assumption

2 In the above, we are ignoring background emission noise which, in many applications, can be eliminated by pre-processing the detected image [22].

does not hold for (PIP): Indeed, if $f_j(x) = -y_j \log(y_j/(Hx)_j)$ denotes the singular part of the KL divergence for the j-th sample, we readily get

$$\frac{\partial f_j}{\partial x_j} = \frac{y_j H_{ji}}{(Hx)_j}. \tag{3.4}$$

This shows that the gradient of f_j exhibits an $\mathcal{O}(1/x)$ singularity at the boundary of $\mathbb{R}^n_+$, so f cannot be Lipschitz under *any* global norm on $\mathbb{R}^n$. The same of course holds for (LC).

3.1.2 *Resource sharing problems*

Resource Sharing Formulation

Consider a set of *resources* $r \in \mathcal{R} = \{1, \dots, R\}$ serving a stream of *demands* that arrive at a rate of ρ per unit of time (for instance, a GPU cluster or a computing grid processing a stream of jobs). If the load on the r-th resource is x_r, the expected service time in the standard Kleinrock model [62] is given by the M/M/1 loss function

$$\ell_r(x_r) = \frac{1}{c_r - x_r}, \tag{3.5}$$

where c_r denotes the capacity of the resource. In this setting, the set of feasible resource allocations is $\mathcal{X} \equiv \{(x_1, \dots, x_r) : 0 \le x_r < c_r, x_1 + \dots + x_R = \rho\}$,[3] and we say that a resource allocation profile $x^* \in x^*$ is at *Nash/Wardrop equilibrium* [94, 113] if

$$\ell_r(x_r^*) \le \ell_r(x_r) \quad \text{for all } x \in \mathcal{X} \text{ and all } r \in \mathcal{R} \text{ such that } x_r^* > 0 \tag{3.6}$$

i.e., when no job would be better served by transferring it to a different priority queue. In this case, if we let $A(x) = (\ell_1(x_1), \dots, \ell_R(x_R))$, a standard calculation shows that x^* is an equilibrium allocation if and only if it solves the associated variational inequality problem for A.

3.1.3 *Fisher market model*

Fisher Model Formulation

Following [94], a Fisher market consists of a set $\mathcal{N} = \{1, \dots, N\}$ of N *buyers* – or *players* – that seek to share a set $\mathcal{A} = \{1, \dots, n\}$ of n perfectly divisible goods (ad space, CPU/GPU runtime, bandwidth, etc.). The allocation mechanism for these goods follows a proportionally fair price-setting rule that is sometimes referred to as a *Kelly auction* [61]: each player $i = 1, \dots, N$ bids x_{ia} per unit of the a-th good, up the player's individual budget; for the sake of simplicity, we assume that this budget is equal to 1 for all players, so $\sum_{a=1}^n x_{ia} \le 1$ for all $i = 1, \dots, N$. The price of the p-th good is then set to be the sum of the players' bids, i.e., $p_a = \sum_{i \in \mathcal{N}} x_{ia}$; then, each player gets a prorated fraction of each good, namely $w_{ia} = x_{ia}/p_a$.

Now, if the marginal utility of the i-th player per unit of the a-th good is θ_{ip}, the agent's total utility will be

$$u_i(x_i; x_{-i}) = \sum_{a \in \mathcal{A}} \theta_{ia} w_{ia} = \sum_{a \in \mathcal{A}} \frac{\theta_{ia} x_{ia}}{\sum_{j \in \mathcal{N}} x_{ja}}, \tag{3.7}$$

3 For posterity, note here that $\mathcal{X}$ is convex but it is not necessarily closed.

where $x_i = (x_{ia})_{a\in\mathcal{A}}$ denotes the bid profile of the i-th player, and we use the shorthand $(x_i; x_{-i}) = (x_1, \dots, x_i, \dots, x_N)$. A *Fisher equilibrium* is then reached when the players' prices bids follow a profile $x^* = (x_1^*, \dots, x_N^*)$ such that

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \tag{Eq}$$

for all $i \in \mathcal{N}$ and all $x_i = (x_{ia})_{a\in\mathcal{A}}$ such that $x_{ia} \geq 0$ and $\sum_{a\in\mathcal{A}} x_{ia} = 1$.[4]

As was observed by Shmyrev [107], the equilibrium problem (Eq) can be rewritten equivalently as

$$\begin{aligned} &\text{minimize} \quad F(x;\theta) \equiv \sum_{a\in\mathcal{A}} p_a \log p_a - \sum_{i\in\mathcal{N}} \sum_{a\in\mathcal{A}} x_{ia} \log \theta_{ia} \\ &\text{subject to} \quad p_a = \sum_{i\in\mathcal{N}} x_{ia}, \ \sum_{a\in\mathcal{A}} x_{ia} = 1, \text{ and } x_{ia} \geq 0 \text{ for all } a \in \mathcal{A}, i \in \mathcal{N}, \end{aligned} \tag{Opt}$$

with the standard continuity convention $0 \log 0 = 0$. In the above, the agents' marginal utilities are implicitly assumed fixed throughout the duration of the game. On the other hand, if these utilities fluctuate stochastically over time, the corresponding reformulation instead involves the *mean* objective

$$f(x) = \mathbb{E}[F(x;\omega)]. \tag{3.8}$$

Because of the logarithmic terms involved, F (and, a fortiori, f) cannot be Lipschitz continuous or smooth in the standard sense.

3.2 TOOLS FOR TRANSCENDING THE EUCLIDEAN FRAMEWORK

In this section we present the necessary mathematical machinery that will allow us to generalize the notions of (Bd) and (LC). In doing so, we shall use two key notions. The first is that of the so-called Bregman divergence, whereas the second consists of the geometrical tool of a local norm, i.e., a norm that depends on the point upon which it is calculated.

3.2.1 *Bregman functions and divergences*

The notion of a Bregman divergence was first introduced by Bregman [29]. The building block for this pseudo-distance function is that of a suitable "reference" *Bregman function*. This is defined as follows:

Bregman Functions

Definition 3.1. A convex l.s.c. function $h\colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is a *Bregman function* on $\mathcal{X}$, if

1. The subdifferential of h admits a continuous selection, i.e., there exists a continuous mapping

$$\nabla : \operatorname{dom} \partial h \to \nabla h(x) \in \partial h(x) \tag{3.9}$$

for all $x \in \operatorname{dom} \partial h$.

4 It is trivial to see that, in this market problem, all users would saturate their budget constraints at equilibrium, i.e., $\sum_{a\in\mathcal{A}} x_{ia} = 1$ for all $i \in \mathcal{N}$.

2. h is strongly convex, i.e., there exists some $K > 0$ such that

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \frac{K}{2}\|x' - x\|^2 \tag{3.10}$$

for all $x \in \operatorname{dom} \partial h$, $x' \in \operatorname{dom} \partial h$.

The induced *Bregman divergence* of h is then defined for all $x \in \operatorname{dom} \partial h$, $x' \in \operatorname{dom} h$ as

Bregman Divergence

$$D(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x \rangle. \tag{3.11}$$

Remark. Our definition follows [56, 87, 91], but there are variant definitions where h is not necessarily assumed strongly convex, cf. [7, 32, 34] and references therein.

Some standard examples of Bregman functions are as follows:

Example 3.1. Euclidean regularizer: Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^n$ endowed with the Euclidean norm $\|\cdot\|_2$. Then, the *Euclidean regularizer* on $\mathcal{X}$ is defined as $h(x) = \|x\|_2^2/2$ and the induced Bregman divergence is the standard square distance $D(x', x) = \|x' - x\|_2^2$ for all $x, x' \in \mathcal{X}$

Example 3.2. Entropic regularizer: Let $\mathcal{X} = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$ be the unit simplex of $\mathbb{R}^n$ endowed with the L^1-norm $\|\cdot\|_1$. Then, the *entropic regularizer* on $\mathcal{X}$ is $h(x) = \sum_i x_i \log x_i$ and the induced divergence is the relative entropy $D(x', x) = \sum_i x'_i \log(x'_i/x_i)$ for all $x' \in \mathcal{X}$ $x \in \operatorname{ri} \mathcal{X}$. In particular, h is 1- strongly convex with respect to $\|\cdot\|_1$.

Example 3.3. Log-barrier: Let $\mathcal{X} = \mathbb{R}^n_{++}$ denote the (open) positive orthant of $\mathbb{R}^n$. Then, the *log-barrier* regularizer on $\mathcal{X}$ is defined as $h(x) = -\sum_{i=1}^n \log x_i$ for all $x \in \mathbb{R}^n_{++}$. The corresponding divergence is known as the *Itakura-Saito divergence* and is given by $D(x, x') = \sum_{i=1}^n (x_i/x'_i - \log(x_i/x'_i) - 1)$ [34].

We conclude this presentation by providing some elementary properties of a Bregman [56].

Lemma 3.1. *Let h be a Bregman function on $\mathcal{X}$ with associated divergence D. Then:*

1. *$D(x', x)$ is convex with respect to x' (but not necessarily with respect to x).*
2. *$D(x, x') \geq \frac{K}{2}\|x - x'\|^2$ for all $x \in \operatorname{dom} h, x' \in \operatorname{dom} \partial h$.*

Remark. In a nutshell, the first part of Lemma 3.1 is directly derived by the convexity of h, whereas the second is obtained by collecting the terms that constitute the Bregman divergence in (3.16). In the sequel we shall revisit Lemma 3.1 under the light of the notion of local norms.

3.2.2 *Finsler geometry and local norms*

Following [17, 35] a *Finsler metric* [17, 35] is described as follows:

Definition 3.2. A *Finsler metric* on a convex subset $\mathcal{X}$ of $\mathcal{V}$ is a continuous function $\Phi \colon \mathcal{X} \times \mathcal{V} \to \mathbb{R}_+$ which satisfies the following properties for all $x \in \mathcal{X}$ and all $z, z' \in \mathcal{V}$:

Finsler Metrics

1. *Subadditivity:* $\Phi(x; z + z') \leq \Phi(x; z) + \Phi(x; z')$.
2. *Absolute homogeneity:* $\Phi(x; \lambda z) = |\lambda|\Phi(x; z)$ for all $\lambda \in \mathbb{R}$.
3. *Positive-definiteness:* $\Phi(x; z) \geq 0$ with equality if and only if $z = 0$.

Finslerian Local Norms

Given a Finsler metric on $\mathcal{X}$, the induced *primal/dual local norms* on $\mathcal{X}$ are respectively defined as

$$\|z\|_x = \Phi(x; z) \quad \text{and} \quad \|w\|_{x,*} = \max\{\langle w, z\rangle : \Phi(x; z) = 1\} \tag{3.12}$$

for all $x \in \mathcal{X}$ and all $z, w \in \mathcal{V}$. We will also say that a Finsler metric on $\mathcal{X}$ is *regular* when $\|w\|_{x',*}/\|w\|_{x,*} = 1 + \mathcal{O}(\|x' - x\|_x)$ for all $x, x' \in \mathcal{X}$, $w \in \mathcal{V}^*$. Finally, for simplicity, we will also assume in the sequel that $\|\cdot\|_x \geq \nu\|\cdot\|$ for some $\nu > 0$ and all $x \in \mathcal{X}$ (this last assumption is for convenience only, as the norm could be redefined to $\|\cdot\|_x \leftarrow \|\cdot\|_x + \nu\|\cdot\|$ without affecting our theoretical analysis).

When $\mathcal{X}$ is equipped with a regular Finsler metric as above, we will say that it is a *Finsler space*.

Example 3.4. Let $\Phi(x; z) = \|z\|$ where $\|\cdot\|$ denotes the reference norm of $\mathcal{X} = \mathcal{V}$. Then the properties of Definition 3.2 are satisfied trivially.

Example 3.5. For a more interesting example of a Finsler structure, consider the set $\mathcal{X} = (0, 1]^n$ and the metric $\|z\|_x = \max_i |z_i|/x_i$, $z \in \mathbb{R}^n$, $x \in \mathcal{X}$. In this case $\|w\|_{x,*} = \sum_{i=1}^n x_i|w_i|$ for all $w \in \mathbb{R}^n$, and the only property of Definition 3.2 that remains to be proved is that of regularity. To that end, we have

$$\|w\|_{x',*} - \|w\|_{x,*} \leq \sum_{i=1}^n |w_i| \cdot |x'_i - x_i| = \sum_{i=1}^n x_i|w_i| \cdot |x'_i - x_i|/x_i \leq \|w\|_{x,*} \cdot \|x' - x\|_x. \tag{3.13}$$

Hence, by dividing by $\|w\|_{x,*}$, we readily get $\|w\|_{x',*}/\|w\|_{x,*} \leq 1 + \|x - x'\|_x$ i.e., $\|\cdot\|_x$ is regular in the sense of Definition 3.2.

Example 3.6 (Riemannian metrics)**.** In its simplest form, a *Riemannian metric* on $\mathcal{C} \subseteq \mathbb{R}^n$ is a field of positive-definite matrices $g(x) \succ 0$, $x \in \mathcal{C}$; for a panoramic view of the subject we refer the reader to [9, 66]. This defines a local norm as $\|z\|_x = \sqrt{z^\top g(x) z}$, and a dual local norm as $\|w\|_{x,*} = \sqrt{w^\top g(x)^{-1} w}$. In this way, Riemannian metrics can be seen as special cases of Finsler metrics; the converse however is not true (35; see also Example 3.7 below).

Examples of Finsler Spaces

Example 3.7 (Shahshahani p-norm)**.** Consider the Finsler metric on $\mathcal{C} = \mathbb{R}^n_{++}$ given by

$$\Phi(x; z) = \left(\sum_{i=1}^n |z_i|^p / x_i\right)^{1/p} \tag{3.14}$$

By a straightforward application of Hölder's inequality, the associated dual norm is given by

$$\|w\|_{x,*} = \left(\sum_{i=1}^n x_i^{c-1} |w|_i^c\right)^{1/q} \tag{3.15}$$

with the convention $p^{-1} + q^{-1} = 1$. This metric is known as the Shahshahani p-norm [104] and it plays an important role in game theory, optimal transport, evolutionary biology, and many other fields – see e.g., [2, 3, 58, 108], and references therein. The Shahshahani p-norm comes from a Riemannian metric if $p = 2$ but not otherwise (since it does not satisfy the parallelogram law for $p \neq 2$).

We are now in the position to revisit Definition 3.1 under the light of the local norms. More precisely, we may assume that the respective Bregman function is compatible with the said norm, resulting to the new notion of a *Bregman-Finsler regularizer*. In particular, we propose the following definition firstly introduced in Antonakopoulos et al. (2019):

Bregman-Finsler Functions

Definition 3.3 (Antonakopoulos et al. [6, 7, 8])**.** Let $\|\cdot\|_x$ be a local norm. We say that $h\colon \mathcal{V} \to \mathbb{R}$ is a *Bregman-Finsler function* on $\mathcal{X}$ if: h is a Bregman function in the sense of Definition 3.1 and h is strongly convex relative to the underlying local norm, i.e.,

$$h(p) \geq h(x) + \langle \nabla h(x), p - x \rangle + \tfrac{1}{2}\alpha \|p - x\|_x^2 \tag{3.16}$$

for some $\alpha > 0$ and all $p \in \mathcal{X}$, $x \in \operatorname{dom} \partial h$.

As a consequence of the above, we have:

Lemma 3.2. *A Bregman function h is α-strongly convex relative to $\|\cdot\|_x$ if and only if*

$$D(p, x) \geq \tfrac{1}{2}\alpha \|p - x\|_x^2 \quad \text{for all } p \in \mathcal{X} \text{ and all } x \in \operatorname{dom} \partial h. \tag{3.17}$$

Proof. The result follows by rearranging (3.16) along with the definition of the Bregman divergence described in (3.11). □

The main difference between Definition 3.3 and Definition 3.1 or the standard assumptions in the literature [19, 27, 28, 56, 77–79, 87, 90, 91] is the strong convexity requirement relative to the local norm $\|\cdot\|_x$ (whose choice, in turn, is aimed to capture the singularity landscape of the operator). We illustrate this with two examples of Bregman-Finsler functions below:

Example 3.8. Suppose that $\mathcal{X} = \mathbb{R}^n$ is endowed with the Euclidean norm. Then, setting $h(x) = (1/2)\|x\|_2^2$, we get the standard expression $D(p, x) = (1/2)\|p - x\|_2^2$ for the associated Bregman divergence. Obviously, h is 1-strongly convex relative to $\|\cdot\|_2$.

Example 3.9. Let $\mathcal{X} = [0, 1)^n$ (so $\mathcal{X}$ is neither open nor closed), and consider the local norm $\|z\|_x^2 = \sum_{i=1}^n |z|_i^2/(1 - x_i)^2$ for $x \in \mathcal{X}$, $z \in \mathbb{R}^n$ (cf. Example 3.7 above). If we set

$$h(x) = \textstyle\sum_{i=1}^n 1/(1 - x_i) \tag{3.18}$$

a straightforward calculation gives

$$D(p, x) = \sum_{i=1}^n \frac{(p_i - x_i)^2}{(1 - p_i)(1 - x_i)^2} \geq \sum_{i=1}^n \frac{(p_i - x_i)^2}{(1 - x_i)^2} = \|p - x\|_x^2, \tag{3.19}$$

i.e., h is strongly convex relative to $\|\cdot\|_x$. Importantly, since $\|\cdot\|_x \geq \|\cdot\|_2$, this Bregman function is also strongly convex relative to the standard Euclidean norm. However, even though the Euclidean regularizer of Example 3.8 is strongly convex relative to *any* global norm on $\mathcal{X}$, it cannot be strongly convex relative to the local norm $\|\cdot\|_x$ because of the singularity of the latter when $x_i \to 1^-$.

3.3 SURROGATES FOR OPERATOR BOUNDEDNESS

Having described this background material, we now proceed to discuss the particular generalizations of Section 2.3 in order to account for problems with singular objective functions. We divide our presentation into two parts. The first concerns these notions that are based on the Bregman divergence, whereas the second part considers the notions based on Finsler induced norms.
The first extension of (LC) is due to [111].

Weakly Continuous Objectives

Definition 3.4 (Teboulle [111])**.** An operator A is said to be W$[h]$-continuous relative to h on $\mathcal{X}$ if there exists some $G > 0$ such that, for all $t > 0$, we have

$$t\langle A(x), x - x'\rangle - D(x', x) \leq \frac{t^2}{2}G^2 \quad \text{for all } x' \in \operatorname{dom} h,\, x \in \operatorname{dom} \partial h. \tag{W}$$

As a prelude we mention that (W) notion intends to single out sufficient conditions for the convergence of "proximal-like" methods like mirror descent. The standard Euclidean (Bd) condition satisfies Definition 3.4. Indeed, if one chooses $h(x) = 1/2\|x\|^2$ and its respective Bregman divergence $D(x, x') = 1/2\|x - x'\|^2$ then by applying Fenchel-Young inequality we get:

$$\begin{aligned} t\langle \nabla f(x), x - x'\rangle - \frac{1}{2}\|x' - x\|^2 &\leq \frac{t^2\|\nabla f(x)\|_*^2}{2} + \frac{1}{2}\|x' - x\|^2 - \frac{1}{2}\|x' - x\|^2 \\ &\leq \frac{t^2G^2}{2} \end{aligned}$$

which yields that f is weakly continuous. Moreover, another related notion is that of (RC), as introduced by [71] and extended further in a recent paper by [119]:

Relative Continuous Objectives

Definition 3.5 (Zhou et al. [119])**.** An operator $A : \mathcal{X} \to \mathbb{R}^n$ is said to be *relatively continuous* if there exists some $G > 0$ such that

$$\langle A(x), x - x'\rangle \leq G\sqrt{2D(x', x)} \quad \text{for all } x \in \operatorname{dom} h,\, x' \in \operatorname{dom} \partial h. \tag{RC}$$

The two notions above are linked in the following manner: Consider an objective f which satisfies Definition 3.4. Then, we have:

$$t\langle \nabla f(x), x - x'\rangle - D(x', x) \leq \frac{t^2}{2}G^2 \quad \text{for all } x' \in \operatorname{dom} h,\, x \in \operatorname{dom} \partial h. \tag{W}$$

By rearranging the above quadratic polynomial in t, we note that its discriminant is $\Delta = [\langle \nabla f(x), x - x'\rangle]^2 - 2G^2D(x', x)$, so it is immediate to check that (RC) holds.

Let us now turn our attention towards the Finsler driven generalized notion of (Bd). In order to give some intuition, let us recall the toy-example presented in Section 3.1. In particular, the $1-$ dimensional logistic regression $f(x) = -\log x$ for $x > 0$ one may straightforwardly detect that the optimizer is dealing with a gradient "singularity" of order $\mathcal{O}(1/x)$. Therefore, if one chooses a local norm of the form:

$$\|x'\|_x = |x'|/x \;\text{ for all }\; x' \in \mathbb{R}, x > 0$$

along with the corresponding dual (local) norm:

$$\|w\|_{x,*} = x|w| \quad \text{for all} \quad w \in \mathbb{R}, x > 0 \tag{3.20}$$

which allows us to obtain:

$$\|\nabla f(x)\|_{x,*} = x(1/x) = 1 \quad \text{for all} \quad x > 0 \tag{3.21}$$

Inspired by the above toy-example a robust theoretical is provided via a local norm induced by Definition 3.2. Formally, we propose the following definition for a generic operator A; firstly introduced in Antonakopoulos et al. (2021).

Metric Boundedness

Definition 3.6 (Antonakopoulos et al. [8])**.** Let $\|\cdot\|_x$, $x \in \mathcal{X}$ be a local norm. We say that A is *metrically bounded* relative to $\|\cdot\|_x$, if there exists some $G > 0$ such that:

$$\|A(x)\|_{x,*} \leq G \quad \text{for all} \quad x \in \mathcal{X} \tag{MB}$$

Remark 3.1. Of course, the standard Euclidean (Bd) is directly recovered by considering $\|\cdot\|_x = \|\cdot\|$. Moreover, in Antonakopoulos et al. (2020) a Riemann-Lipschitz continuity condition is introduced and extends (LC) for the OCO setting as follows. Let $\|\cdot\|_x$ be a family of local norms on $\mathcal{X}$, induced by an appropriate Riemannian metric, and let $\|w\|_{x,*} = \max_{\|x'\|_x \leq 1} \langle w, x' \rangle$ denote the corresponding dual norm. Then, f is *Riemann–Lipschitz continuous* relative to $\|\cdot\|_x$ if there exists some $G > 0$ such that:

Riemann Lipschitz Continuity

$$\|\nabla f(x)\|_{x,*} \leq G \quad \text{for all } x \in \mathcal{X}. \tag{RLC}$$

That said, for the sake of generality we prefer the more general formulation of Definition 3.6.

We conclude this section by presenting the connection between (W) and (MB). More precisely, given a Bregman-Finsler fuction (cf. Definition 3.3) Fenchel-Young inequality ensures:

$$t\langle \nabla f(x), x - x' \rangle - D(x', x) \leq \frac{t^2 \|\nabla f(x)\|_{x,*}^2}{2K} + \frac{K}{2}\|x' - x\|_x^2 - D(x', x) \tag{3.22}$$

which yields that f is weakly K-continuous.

3.4 SURROGATES FOR OPERATOR LIPSCHITZ CONTINUITY

Now we turn our attention towards the generalization of (LC). A popular notion, closely linked with the particular minimization framework, is that of *Lipschitz-like* (or relative smoothness) introduced by [19] (see also [23] [72]). Formally following [19] we have:

Lipschitz-like Objectives

Definition 3.7 (Bauschke et al. [19])**.** A convex l.s.c. function $f \colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is said to be *Lipschitz-like* if there exists some $\beta > 0$ such that

$$\beta h - f \quad \text{is convex on int dom } h. \tag{RS}$$

Here, we recall that the domain of the respective regularizer h is contained in the domain of f. The main motivation behind this elegant definition Definition 3.7 is to generalize the standard descent inequality satisfied by smooth objectives:

$$f(x) \leq f(x') + \langle \nabla f(x'), x - x' \rangle + \beta \|x - x'\|^2 \tag{3.23}$$

by substituting it with the more geometry-sensitive Bregman divergence, i.e.,

$$f(x) \leq f(x') + \langle \nabla f(x'), x - x' \rangle + \beta D(x, x') \tag{3.24}$$

Lipschitz-like Properties

For the sake of completeness we provide an overview of some properties of Lipschitz-like functions; proofs of the following appear in [19] (see also [23, 72]), so we omit it.

Proposition 3.3 (Bauschke et al. [19])**.** *The following statements are equivalent:*

1. *f satisfies* (RS) *in* $\operatorname{int} \mathcal{X}$.
2. *f satisfies the inequality $f(x) \leq f(x') + \langle \nabla f(x'), x - x' \rangle + \beta D(x, x')$, for all $x, x' \in \operatorname{int} \mathcal{X}$*
3. *f satisfies the inequality $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \leq \beta \left[D(x, x') + D(x', x) \right]$.*

That said, the success of (RS) condition is limited for minimization settings; the reason is that descent-type inequalities of the form (3.23) and/or (3.24), are not available for general (VI) problems. Therefore, in order to be able to include optimization beyond minimization settings we should follow a different approach. In particular, by applying similar reasoning with Definition 3.6, we have the following definition again for a generic operator A. Following Antonakopoulos et al. (2021), we propose a novel regularity condition based on the local norm framework.

Metric Smoothness and Variants

Definition 3.8 (Antonakopoulos et al. [8])**.** Given a local norm $\|\cdot\|_x$, $x \in \mathcal{X}$, we say that A is *metrically smooth* (relative to $\|\cdot\|_x$) if

$$\|A(x) - A(x')\|_{x,*} \leq \beta \|x - x'\|_{x'} \quad \text{for all } x, x' \in \operatorname{dom} A. \tag{MS}$$

Following Antonakopoulos et al. (2019), one may also get a similar notion to (MS); namely that of *Bregman continuity* of an operator. Formally, this is given by the following.

Definition 3.9. [Antonakopoulos et al. [6]] Let h be a Bregman-Finsler regularizer relative to some local norm $\|\cdot\|_x$ on $\mathcal{X}$. The operator $A \colon \mathcal{X} \to \mathcal{V}^*$ is said to be *β-Bregman continuous* if

$$\|A(x') - A(x)\|_{x,*} \leq \beta \sqrt{2D(x, x')} \quad \text{for all } x, x' \in \operatorname{dom} A. \tag{BC}$$

Remark 3.2. If A satisfies (MS) and we are given with some h Finsler-Bregman regularizer adapted to the associated norm then one may straightforwardly obtain that (MS) implies (BC).

Finally, we conclude by describing the connection between (RS) and (MS) conditions; of course under the assumption that $A = \nabla f$ for some f convex function. Indeed, if h is a Bregman-Finsler function (cf. Definition 3.3), we have:

$$\begin{aligned}\langle \nabla f(x) - \nabla f(x'), x - x' \rangle &\leq \|\nabla f(x) - \nabla f(x')\|_{x,*} \|x - x'\|_x \\ &\leq \beta \|x - x'\|_{x'} \|x - x'\|_x \\ &\leq \frac{\beta}{2} \left[\|x - x'\|_{x'}^2 + \|x - x'\|_x^2 \right]. \end{aligned} \tag{3.25}$$

Thus by the compatibility of h and $\|\cdot\|_x$, we readily obtain

$$\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \leq \frac{\beta}{K} \left[D(x, x') + D(x', x) \right]. \tag{3.26}$$

Therefore, the claim that (MS) implies (RS) follows from Proposition 3.3.

4

BREGMAN FIRST ORDER METHODS

We now turn to the presentation of the necessary algorithmic machinery that will allow us transcend the limitations of Euclidean based algorithmic schemes. The main ingredient of these methods boils down to a generalized notion of "projections" which are based on a appropriately chosen reference function in the sense of Definition 3.1 and/or Definition 3.3. We describe this toolkit in Section 4.1, where we also present more precisely the so-called *prox* and *mirror* mappings.

As an additional feature we present a novel primal-dual variant of the Bregman divergences; the so-called *Fenchel coupling*. This will serve a "primal-dual" measure of distance and will come in handy for the particular analysis of primal-dual methods in the sequel.

Finally, in Section 4.2 and Section 4.2 we illustrate the explicit defining recursive formulas of the particular algorithmic schemes. In particular, these methods are obtained by revisiting the (GD) and (EG) methods (see in Section 2.5.1 and Section 2.6.1) under the light of these new projection operators, i.e., the prox and mirror mappings. These generic iterative methods will enable us in the sequel to capture in an efficient manner the finer geometrical aspects that arise from the non-Lipschitz framework.

4.1 PROX-AND MIRROR MAPPINGS

In order to describe the algorithmic methods we first present their key ingredients, that of Bregman prox- and mirror mappings. In addition we provide their crucial properties and template inequalities. Versions of these are known in the literature [see e.g., 20, 34, 91, 105, and references therein] and mostly rely on *global* – norms. However, in our case we revisit these results and provide here complete statements and proofs armed with the notion *local* – norm.

In particular, we shall assume that h is a Bregman-Finsler regularization function in the sense of Definition 3.3. To begin, we introduce two key notions that will be useful in the sequel. The first is the convex conjugate of a Bregman function h, i.e.,

$$h^*(y) = \max_{x \in \operatorname{dom} h} \{\langle y, x\rangle - h(x)\} \tag{4.1}$$

and the associated primal-dual *mirror map* $Q\colon \mathcal{V}^* \to \operatorname{dom} \partial h$: *Mirror Map*

$$Q(y) = \arg\max_{x \in \operatorname{dom} h} \{\langle y, x\rangle - h(x)\} \tag{4.2}$$

That the above is well-defined is a consequence of the fact that h is proper, l.s.c., convex and coercive; [1] in addition, the fact that Q takes values in $\operatorname{dom}\partial h$ follows from the fact that any solution of (4.2) must necessarily have nonempty subdifferential (see below Lemma 4.1). We also recall here the definition of the Bregman *proximal mapping*:

Proximal Mapping

$$P_x(w) = \operatorname*{arg\,min}_{x' \in \operatorname{dom} h} \{\langle w, x - x' \rangle + D(x', x)\} \tag{4.3}$$

valid for all $x \in \operatorname{dom}\partial h$ and all $w \in \mathcal{V}^*$.

Mirror and Prox Mapping Links

We then have the following basic lemma connecting the above notions:

Lemma 4.1. *Let h be a K-strongly convex Bregman–Finsler regularizer. Then, for all $x \in \operatorname{dom}\partial h$ and all $w, y \in \mathcal{V}^*$ we have:*

1. $x = Q(y) \iff y \in \partial h(x)$.
2. $x^+ = P_x(w) \iff \nabla h(x) + w \in \partial h(x) \iff x^+ = Q(\nabla h(x) + w)$.
3. *Finally, if $x = Q(y)$ and $p \in \mathcal{X}$, we get:*

$$\langle \nabla h(x), x - p \rangle \leq \langle y, x - p \rangle. \tag{4.4}$$

Proof. For the first equivalence, note that x solves (4.1) if and only if $0 \in y - \partial h(x)$ and hence if and only if $y \in \partial h(x)$. Working in the same spirit for the second equivalence, we get that x^+ solves (4.3) if and only if $\nabla h(x) + w \in \partial h(x^+)$ and therefore if and only if $x^+ = Q(\nabla h(x) + w)$.

For our last claim, by a simple continuity argument, it is sufficient to show that the inequality holds for the relative interior $\operatorname{ri}\mathcal{X}$ of $\mathcal{X}$ (which, in particular, is contained in $\operatorname{dom}\partial h$). In order to show this, pick a base point $p \in \operatorname{ri}\mathcal{X}$, and let

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, t(p - x) \rangle] \quad \text{for all } t \in [0, 1]. \tag{4.5}$$

Since, h is strongly convex and $y \in \partial h(x)$ due to the first equivalence, it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Since, $\psi(t) = \langle \nabla h(x + t(p - x)) - y, p - x \rangle$ is a continuous selection of subgradients of ϕ and both ϕ and ψ are continuous over $[0, 1]$, it follows that ϕ is continuously differentiable with $\phi' = \psi$ on $[0, 1]$. Hence, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0, 1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, p - x \rangle \geq 0$ and thus we obtain the result. □

To proceed, the basic ingredient for establishing connections between Bregman proximal steps is a generalization of the rule of cosines which is known in the literature as the "three-point identity" [34]. This will be our main tool for deriving the main estimates for our results. Being more precise, we have the following lemma:

Bregman Three-Point Identity

Lemma 4.2 (Chen and Teboulle [34])**.** *Let h be a Bregman–Finsler regularizer. Then, for all $p \in \operatorname{dom} h$ and all $x, x' \in \operatorname{dom}\partial h$, we have:*

$$D(p, x') = D(p, x) + D(x, x') + \langle \nabla h(x') - \nabla h(x), x - p \rangle. \tag{4.6}$$

1 The latter holds because h is strongly convex relative to $\|\cdot\|_x$, and $\|\cdot\|_x$ has been tacitly assumed bounded from below by a multiple $\mu\|\cdot\|$ of $\|\cdot\|$.

Proof. By definition:

$$\begin{aligned} D(p,x') &= h(p) - h(x') - \langle \nabla h(x'), p - x' \rangle \\ D(p,x) &= h(p) - h(x) - \langle \nabla h(x), p - x \rangle \\ D(x,x') &= h(x) - h(x') - \langle \nabla h(x'), x - x' \rangle. \end{aligned} \tag{4.7}$$

The lemma then follows by adding the two last lines and subtracting the first. □

Remark 4.1. As one may directly observe from the proof of the above, Lemma 4.2 holds for a general convex function h. However, for the need of our analysis we constrain our selves to the specific regularizer class of interest; namely that of Bregman-Finsler ones.

Thanks to the three-point identity, we obtain the following estimate for the Bregman divergence before and after a mirror descent step:

One Step Mirror Template Inequality

Proposition 4.3. *Let h be a Bregman–Finsler function with strong convexity modulus $K > 0$. Fix some $p \in \operatorname{dom} h$ and let $x^+ = P_x(w)$ for some $x \in \operatorname{dom} \partial h$ and $w \in \mathcal{V}^*$. We then have:*

$$D(p,x^+) \leq D(p,x) - D(x^+,x) + \langle w, x^+ - p \rangle \tag{4.8}$$

and

$$D(p,x^+) \leq D(p,x) + D(x,x^+) - \langle w, x - p \rangle. \tag{4.9}$$

Proof. By the three-point identity established in Lemma 4.2, we have:

$$D(p,x) = D(p,x^+) + D(x^+,x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p \rangle \tag{4.10}$$

Rearranging terms then yields:

$$D(p,x^+) = D(p,x) - D(x^+,x) + \langle \nabla h(x^+) - \nabla h(x), x^+ - p \rangle \tag{4.11}$$

By (4.4) and the fact that $x^+ = P_x(w)$ so $\nabla h(x) + w \in \partial h(x^+)$, the first inequality follows; the second one is obtained similarly. □

Thanks to the above estimations, we obtain the following inequalities relating the Bregman divergence between *two* prox-steps:

Two Steps Mirror Template Inequality

Proposition 4.4. *Let h be a Bregman function on $\mathcal{X}$ and fix some $p \in \mathcal{X}$, $x \in \mathcal{X}^\circ$. Letting $x_1^+ = P_x(w_1)$ and $x_2^+ = P_x(w_2)$, we have:*

$$D(p,x_2^+) \leq D(p,x) + \langle w_2, x_1^+ - p \rangle + [\langle w_2, x_2^+ - x_1^+ \rangle - D(x_2^+,x)] \tag{4.12}$$

and

$$\begin{aligned} D(p,x_2^+) \leq D(p,x) + \langle w_2, x_1^+ - p \rangle + \langle w_2 - w_1, x_2^+ - x_1^+ \rangle \\ - D(x_2^+,x_1^+) - D(x_1^+,x). \end{aligned} \tag{4.13}$$

Proof. For the first inequality, by applying (4.8) for $x_2^+ = P_x(w_2)$, we get:

$$D(p,x_2^+) \leq D(p,x) - D(x_2^+,x) + \langle w_2, x_2^+ - p \rangle$$

$$= D(p,x) + \langle w_2, x_1^+ - p\rangle + [\langle w_2, x_2^+ - x_1^+\rangle - D(x_2^+, x)] \tag{4.14}$$

For the second inequality, we need to bound $\langle w_2, x_2^+ - x_1^+\rangle - D_h(x_2^+, x)$. In particular, applying again (4.8) for $p = x_2^+$, we get:

$$D(x_2^+, x_1^+) \leq D(x_2^+, x) + \langle w_1, x_1^+ - x_2^+\rangle - D(x_1^+, x) \tag{4.15}$$

and hence:

$$D(x_2^+, x) \geq D(x_2^+, x_1^+) + D(x_1^+, x) - \langle w_1, x_1^+ - x_2^+\rangle. \tag{4.16}$$

So, combining the above inequalities we get:

$$\begin{aligned}\langle w_2, x_2^+ - x_1^+\rangle - D(x_2^+, x) \leq \langle w_2, x_2^+ - x_1^+\rangle - D(x_2^+, x_1^+) \\ - D(x_1^+, x) - \langle w_1, x_2^+ - x_1^+\rangle\end{aligned} \tag{4.17}$$

and thus we get the second inequality as well. □

On the other hand, much of our analysis of primal-dual methods revolves around a "primal-dual" divergence between a target point $p \in \mathcal{X}$ and a dual vector $y \in \mathcal{Y}$. This gives rise to the primal-dual counterpart of the Bregman divergence, the so-called *Fenchel coupling*. Following [79], this is defined as follows for all $p \in \mathcal{X}$, $y \in \mathcal{Y}$:

Fenchel Coupling

$$F(p,y) = h(p) + h^*(y) - \langle y, p\rangle. \tag{4.18}$$

The following lemma illustrates basic properties of the Fenchel coupling and generalizes similar properties derived in [79]:

Fenchel Coupling Properties

Lemma 4.5. *Let h be a Bregman-Finsler regularizer on $\mathcal{X}$ with convexity modulus α. Then, for all $p \in \mathcal{X}$ and all $y \in \mathcal{Y}$, we have:*

1. $F(p,y) \geq D(p, Q(y))$.
2. *Moreover,*

$$F(p,y) = D(p,Q(y)) \;\; \textit{if} \;\; Q(y) \in \mathcal{X}^\circ \;\; \textit{(but not necessarily otherwise).} \tag{4.19}$$

3. *If $x = Q(y)$, then $F(p,y) \geq \frac{\alpha}{2}\|x - p\|_x^2$*

Proof. For the first inequality we have,

$$\begin{aligned}F(p,y) &= h(p) + h^*(y) - \langle y, y\rangle \\ &= h(p) - h(Q(y)) + \langle y, Q(y)\rangle + \langle y, -p\rangle \\ &= h(p) - h(Q(y)) - \langle y, p - Q(y)\rangle\end{aligned}$$

Since $y \in \partial h(Q(x))$, by Lemma 4.1 we get

$$\langle \nabla h(Q(y)), Q(y) - p\rangle \leq \langle y, Q(y) - p\rangle$$

With all the above we then have

$$F(p,y) = h(p) - h(Q(y)) - \langle y, p - Q(y)\rangle$$

$$\begin{aligned} &\geq h(p) - h(Q(y)) - \langle \nabla h(Q(y)), p - Q(y) \rangle \\ &= D(p, Q(y)) \end{aligned}$$

For the equality, let $x = Q(y)$. Then, by definition we have:

$$\begin{aligned} F(p,y) &= h(p) - \langle y, Q(y) \rangle - h(Q(y)) - \langle y, p \rangle \\ &= h(p) - h(x) - \langle y, p - x \rangle. \end{aligned}$$

Since $y \in \partial h(x)$, we have $h'(x; p - x) = \langle y, p - x \rangle$ whenever $x \in \mathcal{X}^\circ$, thus proving our first claim. For our second claim, working in the previous spirit we get that:

$$F(p,y) = h(p) - h(x) - \langle y, p - x \rangle \tag{4.20}$$

Thus, we obtain the result by recalling the strong convexity assumption for h with respect to the local norm $\|\cdot\|_x$. □

We continue with some basic relations connecting the Fenchel coupling relative to a target point before and after a gradient step. The basic ingredient for this is a primal-dual analogue of Lemma 4.2

Fenchel Three-Point Identity

Lemma 4.6. *Let h be a Bregman-Finsler regularizer on $\mathcal{X}$. Fix some $p \in \mathcal{X}$ and let $y, y^+ \in \mathcal{Y}$. Then, letting $x = Q(y)$, we have*

$$F(p,y^+) = F(p,y) + F(x,y^+) + \langle y^+ - y, x - p \rangle. \tag{4.21}$$

Proof. By definition, we get:

$$\begin{aligned} F(p,y^+) &= h(p) + h^*(y^+) - \langle y^+, p \rangle \\ F(p,y) &= h(p) + h^*(y) - \langle y, p \rangle. \end{aligned} \tag{4.22}$$

Then, by subtracting the above we get:

$$\begin{aligned} F(p,y^+) - F(p,y) &= h(p) + h^*(y^+) - \langle y^+, p \rangle - h(p) - h^*(y) + \langle y, p \rangle \\ &= h^*(y^+) - h^*(y) - \langle y^+ - y, p \rangle \\ &= h^*(y^+) - \langle y, Q(y) \rangle + h(Q(y)) - \langle y^+ - y, p \rangle \\ &= h^*(y^+) - \langle y, x \rangle + h(x) - \langle y^+ - y, p \rangle \\ &= h^*(y^+) + \langle y^+ - y, x \rangle - \langle y^+, x \rangle + h(x) - \langle y^+ - y, p \rangle \\ &= F(x,y^+) + \langle y^+ - y, x - p \rangle \end{aligned} \tag{4.23}$$

and our proof is complete. □

4.2 BREGMAN FIRST ORDER METHODS

Armed with the mirror and prox-mappings presented in Section 4.1, we are now in the position to revisit the Sections 2.5 and 2.6. More precisely, we start with the Bregman version of (GD); widely known as mirror descent (MD) algorithm.
The template upon which the template of MD hinges is the following recursion:

Mirror Descent

$$X_{t+1} = P_{X_t}(-\gamma_t V_t) \tag{MD}$$

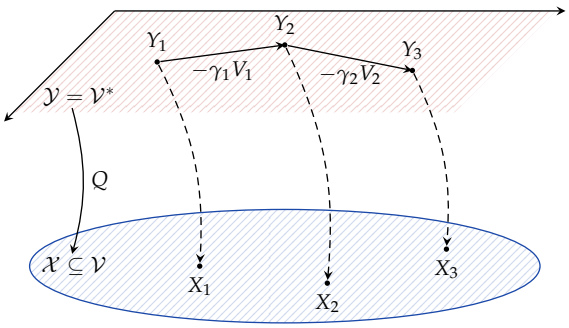


Figure 4.1: Schematic representation of lazy mirror descent.

In the above, the standard notation of Section 2.5 is preserved. In particular, $X_t \in \operatorname{dom} \partial h$ denotes the current state of the algorithm, $V_t \in \mathcal{V}^*$ denotes a generic search direction, $\gamma_t > 0$ is a step-size parameter, and X_{t+1} is the new state generated after taking a Bregman proximal step from x along $-\gamma_t V_t$. (MD) closely resembles the projected gradient update (GD) and, indeed, (GD) is recovered if we take $h(x) = (1/2)\|x\|_2^2$ (cf. Example 3.8). In addition, the abstact template:

$$x^+ = P_x(-\gamma w) \tag{4.24}$$

is *well-posed* in our setting. Therefore, its allows us to iterate (MD) in perpetuity. Formally, we have the following result:

Proposition 4.7. *The abstract recursion,*

$$x^+ = P_x(-\gamma w) \tag{4.25}$$

satisfies $x^+ \in \operatorname{dom} \partial h$ *for all* $x \in \operatorname{dom} \partial h$ *and all* $V \in \mathcal{V}^*$.

Proposition 4.7 is a direct corollary of Lemma 3.1, so we omit its proof. Now we turn our attention towards generalizing the primal-dual methods described in (LGD) and (DA); more precisely, we shall revisit these methods under the lens of the mirror mapping (4.2). In particular, the lazy version of (MD) is given by the following:

Lazy Mirror Descent

$$\begin{aligned} Y_{t+1} &= Y_t - \gamma_t V_t \\ X_{t+1} &= Q(Y_{t+1}) \end{aligned} \tag{LMD}$$

Mirror Dual Averaging whereas the (DAvg) is given by:

$$\begin{aligned} Y_{t+1} &= Y_t - V_t \\ X_{t+1} &= Q(\gamma_{t+1} Y_{t+1}) \end{aligned} \tag{DAvg}$$

Now, building on these templates we are in a position to generalize the method of Section 2.6.1 along with its primal-dual counterpart. In particular, the (EG) template is extended by applying the prox-mapping (4.3). More precisely, Mirror-Prox is derived by applying the following recursion:

Mirror-Prox

$$\begin{aligned} X_{t+1/2} &= P_{X_t}(-\gamma_t V_t) \\ X_{t+1} &= P_{X_t}(-\gamma_t V_{t+1/2}) \end{aligned} \tag{MP}$$

As it becomes apparent from (MP), the method consists of two (MD) steps. Therefore,if we take $h(x) = (1/2)\|x\|_2^2$ then (MP) boils down to the (EG) template. On the other hand, the respective primal-dual version of (MP); namely that of the Bregman generalization of (DualX). In particular, this given by the following recursion:

$$\begin{aligned} X_{t+1/2} &= P_{X_t}(-\gamma_t V_t) \\ Y_{t+1} &= Y_t - V_{t+1/2} \\ X_{t+1} &= Q(\gamma_{t+1} Y_{t+1}) \end{aligned} \tag{DualX}$$

Mirror Dual Extrapolation

Once more, if the optimizer chooses the euclidean regularizer $h(x) = 1/2\|x\|_2^2$ then (DualX) boils down directly to (DualX).

Having described the crucial algorithmic methods in what follows we shall exploit their adaptivity to the particular geometrical features of problems which exhibit "gradient" singularities.

Part II

PROPOSED METHODS AND THEIR GUARANTEES

5

REGRET MINIMIZATION BEYOND LIPSCHITZ CONTINUITY

 # This section incorporates material from the paper [7]

Our first set of results concerns the generic framework of OCO problems. In particular, in Section 5.1 we present an optimal regret minimization result for convex losses which satisfy (RLC) instead of the traditional (euclidean based) Lipschitz continuity condition. Moving forward, in Section 5.2 we apply our regret minimization result to the particular case of stochastic non-smooth convex minimization and provide the respective convergence rates for such problems. In addition, as an extra feature we also provide an almost sure convergence result towards the problem's set of f minimizers.

5.1 REGRET MINIMIZATION

Online Convex Optimization Setting

Throughout this section, we make the following blanket assumptions:

1. The t-th stage loss function $f_t \colon \mathcal{X} \to \mathbb{R}$ is convex and satisfies (MB) with constant G_t.
2. The optimizer's aggregate loss $\sum_{t=1}^{T} f_t$ attains its minimum value at some $x^* \in \mathcal{X}$.

The purpose of the last assumption is to avoid cases where the infimum of a loss function is not attained within the problem's feasible region (such as e^{-x} over $\mathbb{R}_+$).

Online Lazy Mirror Descent

Moreover, from an algorithmic point of view we consider the (LMD) template

$$\begin{aligned} Y_{t+1} &= Y_t - \gamma_t V_t \\ X_{t+1} &= Q(Y_{t+1}) \end{aligned} \tag{5.1}$$

Blanket Assumptions

which satisfies the following blanket assumption:

1. The underlying regularizer h is a Bregman-Finsler function, i.e., it satisfies Definition 3.3.
2. The algorithm is initialized at the "prox-center" $x_c = \arg\min h$ of $\mathcal{X}$ and is run with (constant) step-size $\alpha / T^{1/2}$ for some $\alpha > 0$ chosen by the optimizer.
3. Finally, for (SFO). we make the following assumptions:

   a) Unbiasedness: $\mathbb{E}[\hat{g}_t \mid \mathcal{F}_t] = \nabla f_t(X_t)$. (5.2a)

$$b)\ \textit{Finite mean square:} \quad \mathbb{E}[\|\hat{g}_t\|_*^2 \mid \mathcal{F}_t] \leq M_t^2. \tag{5.2b}$$

NoLips Regret Guarantees

Having established the algorithmic framework, we are in the position to introduce our first result:

Theorem 5.1 (Antonakopoulos et al. [7]). *Let* $\mathrm{Reg}(T) \equiv \mathrm{Reg}_{x^*}(T)$ *with* $x^* \in \arg\min_{x\in\mathcal{X}} \sum_{t=1}^T f(x)$ *and* $\overline{M}_T^2 = T^{-1}\sum_{t=1}^T M_t^2$. *Then,* (LMD) *algorithm with noisy feedback of the form* (SFO) *enjoys the mean regret bound:*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \left[\frac{D(x^*, x_c)}{\alpha} + \frac{\alpha \overline{M}_T^2}{2\alpha}\right]\sqrt{T} \tag{5.3}$$

In particular, if $\sup_{t\mathbb{N}} M_t < +\infty$ *then the method guarantees* $\mathcal{O}(\sqrt{T})$ *regret.*

The main idea behind the proof of Theorem 5.1 is to relate the Finslerian structure of $\mathcal{X}$ to the Bregman regularization framework underlying (LMD). A first such link is provided by the Bregman divergence; however, because of the primal-dual interplay between $X_t \in \mathcal{X}$ and $Y_t \in \mathcal{V}^*$, the Bregman divergence is not sufficiently adapted. At this point is were the *Fenchel coupling*, defined in (4.18), comes in handy. Armed with toolkit, we are able to establish the main "energy" inequality for this section. Formally, we have the following result.

Regret Template Inequality

Proposition 5.2 (Antonakopoulos et al. [7]). *Let* h *be a Bregman-Finsler regularizer on* $\mathcal{X}$ *with convexity modulus* α, *fix some* $p \in \mathcal{X}$, *let* $x = Q(y)$ *for some* $y \in \mathcal{Y}$. *Then, for all* $w \in \mathcal{Y}$, *we have:*

$$F(p, y+w) \leq F(p,y) + \langle w, x - p\rangle + \frac{1}{2\alpha}\|w\|_{x,*}^2 \tag{5.4}$$

Proof. By the three-point identity (4.21), we get

$$F(p,y) = F(p, y+w) + F(Q(y+w), y) + \langle y - (y+w), Q(y+w) - p\rangle \tag{5.5}$$

and hence, after rearranging:

$$\begin{aligned} F(p, y+w) &= F(p,y) - F(Q(y+w), y) + \langle w, Q(y+w) - p\rangle \\ &= F(p,y) - F(Q(y+w), y) + \langle w, x - p\rangle + \langle w, Q(y+w) - x\rangle \end{aligned} \tag{5.6}$$

By Young's inequality [102], we also have

$$\langle w, Q(y+w) - x\rangle \leq \frac{\alpha}{2}\|Q(y+w) - x\|_x^2 + \frac{1}{2\alpha}\|w\|_{x,*}^2 \tag{5.7}$$

Our claim then follows by the fact that $F(Q(y+w), y) \geq \frac{\alpha}{2}\|Q(y+w) - x\|_x^2$ (cf. Lemma 4.5). □

Proof of Theorem 5.1. Now, applying Proposition 5.2 to (LMD), we get:

$$F(x^*, Y_{t+1}) \leq F(x^*, Y_t) - \gamma\langle \hat{g}_t, X_t - x^*\rangle + \frac{\gamma^2}{2\alpha}\|\hat{g}_t\|_{X_t,*}^2$$

$$= F(x^*, Y_t) + \gamma\langle \nabla f_t(X_t), x^* - X_t\rangle - \gamma\langle U_{t+1}, X_t - x^*\rangle + \frac{\gamma^2}{2\alpha}\|\hat{g}_t\|^2_{X_t,*}. \tag{5.8}$$

Hence, after rearranging and telescoping, we obtain

$$\mathrm{Reg}(T) \leq \sum_{t=1}^{T}\langle \nabla f_t(X_t), X_t - x^*\rangle \leq \frac{D(x^*, x_c)}{\gamma} + \sum_{t=1}^{T}\xi_{t+1} + \frac{\gamma}{2\alpha}\sum_{t=1}^{T}\|\hat{g}_t\|^2_{X_t,*} \tag{5.9}$$

where, in the last line, we used the definition of the Finsler dual norm $\|\cdot\|_* \equiv \|\cdot\|_{x^*,*}$, and we set $\xi_{t+1} = \langle U_{t+1}, x^* - X_t\rangle$. By taking expectations on both sides, we have:

$$\mathbb{E}\left[\mathrm{Reg}(T)\right] \leq \frac{D(x^*, x_c)}{\gamma} + \sum_{t=1}^{T}\mathbb{E}\left[\xi_{t+1}\right] + \sum_{t=1}^{T}\mathbb{E}\left[\|\hat{g}_t\|^2_{X_t,*}\right] \tag{5.10}$$

We examine each (RHS) term individually. In particular, we have:

- For the term $\sum_{t=1}^{T}\mathbb{E}\left[\xi_{t+1}\right]$ we have:

$$\begin{aligned}\mathbb{E}\left[\xi_{t+1}\right] &= \mathbb{E}\left[\langle U_{t+1}, x^* - X_t\rangle\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\langle U_{t+1}, x^* - X_t\rangle | \mathcal{F}_t\right]\right] \\ &= \mathbb{E}\left[\langle \mathbb{E}\left[U_{t+1}|\mathcal{F}_t\right], x^* - X_t\rangle\right] \\ &= 0\end{aligned}$$

  with the last equality being obtained by the unbiasedness of (SFO)

- For the term $\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{g}_t\|^2_{X_t,*}\right]$ we have by the finite mean square assumption of (SFO):

$$\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{g}_t\|^2_{X_t,*}\right] = \mathcal{O}(\sqrt{T}) \tag{5.11}$$

Finally the result follows by combining the above. □

Remark 5.1. We emphasize here that the $\mathcal{O}(\sqrt{T})$ regret bound above is achieved even if $\mathcal{X}$ is unbounded or if the range of h $H \equiv \sup_{x\in\mathcal{X}} D(x, x_c) = \sup h - \inf h$ of $\mathcal{X}$ is infinite.To see this, simply note that $D(x, x_c) = h(x) - h(x_c) - \langle \nabla h(x_c), x - x_c\rangle < \infty$ for all $x \in \mathcal{X} = \mathrm{dom}\, h$ (recall also that, since $x_c = \arg\min h$, we have $0 \in \partial h(x_c)$ so $x_c \in \mathrm{dom}\,\partial h$). Of course, if $H < \infty$ and G (or M) is known to the optimizer, (5.3) can be optimized further by tuning α.

Our analysis hinges on controlling the second-order error term in (5.2) by means of the (MB) continuity assumption. It is precisely this primal-dual inequality which allows us to go beyond the standard Lipschitz framework: compared to (primal-primal) inequalities of a similar form for global norms [16, 64, 87, 91, 120], the distinguishing feature of (5.2) is the advent of the Finsler induced norm $\|w\|_{x,*}$. Thanks to the intricate connection between the Finsler norm and h, the second-order term in (5.2) can be controlled even when the received gradient is unbounded relative to *any* global norm, i.e., even if the objective is singular.

The main obstacle to achieve this is that the underlying local norm, the Fenchel coupling F and the Bregman divergence D (all state-dependent notions of distance)

need not be compatible with one another. That this is indeed the case is obtained by Lemma 4.5; what plays a crucial role in deriving (5.2) is to introduce the local strong convexity with respect to the the Finsler norm to the *second* argument of the Bregman divergence instead of the first (or any other point in-between). Any other relation between the local norms and h along these lines is not amenable to analyzing (LMD) in this framework.

5.2 APPLICATION TO STOCHASTIC NON-SMOOTH MINIMIZATION

As announced, the second part of our analysis focuses on the application of the above regret analysis on stochastic non-smooth optimization problems of the form:

Non-Smooth Stochastic Minimization

$$\begin{aligned} &\text{minimize} && f(x) = \mathbb{E}[F(x;\omega)] \\ &\text{subject to} && x \in \mathcal{X} \end{aligned} \tag{Opt}$$

with the expectation taken over some model sample space Ω. Our first result here is as follows:

NonLips Non-Smooth Guarantees

Theorem 5.3. *Assume that f is convex and satisfies (MB) in mean square, i.e.,*

$$\sup_x \mathbb{E}[\|\nabla F(x;\omega)\|_{x,*}^2] \leq M^2 \tag{5.12}$$

for some $M > 0$. If (LMD) is run for T iterations with a constant step-size of the form $\alpha/\sqrt{T}$ and stochastic gradients $\hat{g}_t = \nabla F(X_t;\omega_t)$ generated by an i.i.d. sequence $\omega_t \in \Omega$, we have

$$\mathbb{E}[f(\bar{X}_T)] \leq \min f + \left[\frac{D_c}{\alpha} + \frac{\alpha M^2}{2\alpha}\right]\frac{1}{\sqrt{T}} \tag{5.13}$$

where $\bar{X}_T = (1/T)\sum_{t=1}^T X_t$ is the "ergodic average" of X_t and

$$D_c = \inf_{x^* \in \arg\min f} D(x^*, x_c) < \infty \tag{5.14}$$

denotes the Bregman distance of the prox-center x_c of $\mathcal{X}$ to $\arg\min f$.

The key novelty in Theorem 5.3 is that the optimal $\mathcal{O}(T^{-1/2})$ convergence rate of (LMD) is maintained *even if the stochastic gradients of F become singular at residual points $x \in \operatorname{cl}(\mathcal{X}) \setminus \mathcal{X}$*. The proof of Theorem 5.3 likewise relies on an online-to-batch conversion of the regret guarantees of (LMD) for the sequence of stochastic gradients $\nabla F(\cdot;\omega_t)$ of f.

To go beyond the ergodic guarantees of Theorem 5.3, we also analyze below the convergence of the "last iterate" of online mirror descent (OMD), i.e., the *actual sequence of generated points X_t*. This is of particular interest for non-convex problems where ergodic convergence results are of limited value (because Jensen's inequality no longer applies). To obtain global convergence results in this setting, we focus on a class of functions which satisfy a weak secant inequality of the form

Weak Secant Inequality

$$\inf\{\langle \nabla f(x), x - x^*\rangle : x^* \in \arg\min f, x \in \mathcal{K}\} > 0 \tag{SI}$$

for every closed subset $\mathcal{K}$ of $\mathcal{X}$ that is separated by neighborhoods from $\arg\min f$. Variants of this condition have been widely studied in the literature and include non-

convex functions with complicated ridge structures [26, 40, 55, 59, 70, 93, 118, 120]. In this very general setting, we have:

Last Iterate Convergence

Theorem 5.4. *Assume f satisfies (SI) and satisfies (MB) in L^2. Suppose further that $\arg\min f$ is bounded and (LMD) is run with a sequence of stochastic gradients $\hat{g}_t = \nabla F(X_t; \omega_t)$, a Bregman–Finsler regularizer h, and a variable step-size γ_t such that $\sum_{t=1}^{\infty} \gamma_t = \infty$, $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$. Then, with probability 1, X_t converges to some (possibly random) $x^* \in \arg\min f$.*

We begin by recalling two important results from probability theory. The first is a version of the law of large numbers for martingale difference sequences that are bounded in L^2 [47]:

Law of Large Numbers for Martingales

Theorem 5.5 (Hall and Heyde [47])**.** *Let $Y_t = \sum_{i=1}^{t} \zeta_i$ be a martingale and β_t a non-decreasing positive sequence such that $\lim_{t\to\infty} \beta_t = \infty$. Then,*

$$\lim_{t\to\infty} Y_t / \beta_t = 0 \ \text{ almost surely} \tag{5.15}$$

on the set $\sum_{t=1}^{\infty} \beta_t^{-2} \mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] < \infty$.

The second is a convergence result for quasi-supermartingales due to Robbins and Sigmund [101]:

Stochastic Quasi-Fejer Sequences

Lemma 5.6 (Robbins and Sigmund [101])**.** *Let $(\mathcal{F}_t)_{t\in\mathbb{N}}$ be a non-decreasing sequence of $\sigma-$ algebras. Let $(\alpha_t)_{t\in\mathbb{N}}$, $(\theta_t)_{t\in\mathbb{N}}$ non-negative $\mathcal{F}_t-$ measurable random variables, $(\eta_t)_{t\in\mathbb{N}}$ is an $\mathcal{F}_t-$ measurable non-negative summable random variable and the following inequality holds:*

$$\mathbb{E}[\alpha_{t+1} \mid \mathcal{F}_t] \leq \alpha_t - \theta_t + \eta_t \ \text{ almost surely} \tag{5.16}$$

Then, $(\alpha_t)_{t\in\mathbb{N}}$ converges almost surely towards a $[0,\infty)$-valued random variable.

An application of this lemma leads us to the following result which is of independent interest:

Proposition 5.7 (Antonakopoulos et al. [7])**.** *Let X_t be the sequence of iterates generated by (LMD) run with a step-size sequence γ_t such that $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$ and a stochastic oracle as in the statement of Theorems 5.3 and 5.4. Then, for all $x^* \in \arg\min f$, $F(x^*, Y_t)$ converges with probability 1.*

Proof. Let $x^* \in \arg\min f$. Recalling our main estimation:

$$F(x^*, Y_{t+1}) \leq F(x^*, Y_t) - \gamma_t \langle \hat{g}_t, X_t - x^* \rangle_x + \frac{\gamma_t^2}{2\alpha} \|\hat{g}_t\|^2_{X_t,*} \tag{5.17}$$

and taking conditional expectations on both sides, we get due to $\mathcal{F}_t-$ measurability arguments:

$$\mathbb{E}[F(x^*, Y_{t+1}) | \mathcal{F}_t] \leq F(x^*, Y_t) - \gamma_t \langle \hat{g}_t, X_t - x^* \rangle_x + \frac{\gamma_t^2}{2\alpha} \mathbb{E}[\|\hat{g}_t\|^2_{X_t,*} | \mathcal{F}_t]. \tag{5.18}$$

Since, $(2\alpha)^{-1} \sum_{t=1}^{\infty} \gamma_t^2 \mathbb{E}[\|\hat{g}_t\|^2_{X_t,*} | \mathcal{F}_t] \leq M(2\alpha)^{-1} \sum_{t=1}^{\infty} \gamma_t^2 < \infty$ by applying the above we get the result. □

Almost Sure Boundedness

Having this at hand, we can establish the following proposition:

Proposition 5.8. *Let X_t be the sequence of iterates generated by* (LMD) *with assumptions as in Theorem 5.4. Then, for all $x^* \in \arg\min f$, the sequence $\|X_t - x^*\|_{X_t}$ is bounded with probability* 1.

Proof. Recalling our main estimation and taking condition expectations on both sides, we get:

$$\mathbb{E}[F(x^*, Y_{t+1}) \mid \mathcal{F}_t] \leq F(x^*, Y_t) - \gamma_t \langle \hat{g}_t, X_t - x^* \rangle_x + \frac{\gamma_t^2}{2\alpha} \mathbb{E}[\|\hat{g}_t\|_{X_t,*}^2 | \mathcal{F}_t] \tag{5.19}$$

Hence, by the above corollary, we have that the sequence $F(x^*, Y_t)$ converges with probability 1 for all $x^* \in \arg\min f$. Thus, it is also bounded with probability 1 for all x^*. We then get

$$\|X_t - x^*\|_{X_t}^2 \leq \frac{2}{\alpha} F(x^*, Y_t) \tag{5.20}$$

which concludes our proof. □

Existence of Convergent Sub-sequence

We continue by showing that X_t possesses a subsequence that converges to $\arg\min f$:

Proposition 5.9 (Antonakopoulos et al. [7])**.** *Let X_t be the sequence of iterates generated by* (LMD) *with assumptions as in Theorem 5.4. Then, with probability* 1*, there exists a* (*possibly random*) *subsequence of X_t which converges to* $\arg\min f$.

Proof. Assume to the contrary that, with positive probability, the sequence X_t generated by (LMD) admits no limit points in $\arg\min f$. Conditioning on this event, there exists a (nonempty) closed set $\mathcal{C} \subset \mathcal{X}$ which is separated by neighborhoods from $\arg\min f$ and is such that $X_t \in \mathcal{C}$ for all suffiently large t. Then, by relabeling X_t if necessary, we can assume without loss of generality that $X_t \in \mathcal{C}$ for all $t \in \mathbb{N}$. Thus, by Proposition 5.2, we get:

$$\begin{aligned} F(x^*, Y_{t+1}) &\leq F(x^*, Y_t) - \gamma_t \langle \hat{g}_t, X_t - x^* \rangle + \frac{\gamma_t^2}{2\alpha} \|\hat{g}_t\|_{X_t,*}^2 \\ &= F(x^*, Y_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle - \gamma_t \langle U_{t+1}, X_t - x^* \rangle + \frac{\gamma_t^2}{2\alpha} \|\hat{g}_t\|_{X_t,*}^2 \\ &\leq F(x^*, Y_t) - \gamma_t \delta(\mathcal{C}) + \gamma_t \xi_{t+1} + \frac{\gamma_t^2}{2\alpha} \|\hat{g}_t\|_{X_t,*}^2 \end{aligned} \tag{5.21}$$

where in the last line we set $\delta(\mathcal{C}) = \inf\{\langle \nabla f(x), x - x^* \rangle : x^* \in \arg\min f, x \in \mathcal{C}\} > 0$ (by (SI)), $U_{t+1} = \hat{g}_t - \nabla f(X_t)$, $\xi_{t+1} = -\langle U_{t+1}, X_t - x^* \rangle$ and $\beta_t = \sum_{i=1}^t \gamma_i$. Thus, by telescoping and factorizing we get:

$$F(x^*, Y_{t+1}) \leq F(x^*, Y_1) - \beta_t \left[\delta(\mathcal{C}) - \frac{\sum_{s=1}^t \gamma_s \xi_{s+1}}{\beta_t} - \frac{\sum_{s=1}^t \gamma_s^2 \|\hat{g}_s\|_{X_s,*}^2}{2\alpha\beta_t} \right] \tag{5.22}$$

By the unbiasedness assumption for U_t, we have $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = \langle \mathbb{E}[U_{t+1} \mid \mathcal{F}_t], X_t - x^* \rangle = 0$. Moreover, for all $x^* \in \arg\min f$, we have

$$\sum_{t=1}^{\infty} \gamma_t^2 \, \mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] \leq \sum_{t=1}^{\infty} \gamma_t^2 \|X_t - x^*\|_{X_t}^2 \, \mathbb{E}[U_{t+1} \mid \mathcal{F}_t] \leq \sum_{t=1}^{\infty} \gamma_t^2 F(x^*, Y_t) \, \mathbb{E}[U_{t+1} \mid \mathcal{F}_t] < \infty \tag{5.23}$$

where the last (strict) inequality is obtained due to the finite mean square property, the boundness of $F(x^*, Y_t)$ and the fact that $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$. Thus, we can apply the law of large numbers for L^2- martingales stated above and conclude that $\beta_t^{-1} \sum_{s=1}^{t} \gamma_s \xi_{s+1}$ converges to 0 almost surely. On the other hand, for the term $S_{t+1} = \sum_{s=1}^{t} \gamma_s^2 \|\hat{g}_s\|_{X_t,*}^2$, since $\hat{g}_{s+1}$ is $\mathcal{F}_s$-measurable for all $s = 1, 2 \dots, t-1$ we have:

$$\mathbb{E}[S_{t+1} \mid \mathcal{F}_t] = \mathbb{E}\left[\sum_{i=1}^{t-1} \gamma_t^2 \|\hat{g}_i\|_{x_i,*}^2 + \gamma_t^2 \|\hat{g}_t\|_{X_t,*}^2 \,\middle|\, \mathcal{F}_t \right] = S_t + \gamma_t^2 \, \mathbb{E}\left[\|\hat{g}_t\|_{X_t,*}^2 \,\middle|\, \mathcal{F}_t \right] \geq S_t \tag{5.24}$$

so S_t is a submartingale with respect to $\mathcal{F}_t$. Furthermore, by the law of total expectation, we also get:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}[\mathbb{E}[S_{t+1} \mid \mathcal{F}_t]] \leq \sigma^2 \sum_{i=1}^{t} \gamma_i^2 \leq \sigma^2 \sum_{t=1}^{\infty} \gamma_t^2 < \infty, \tag{5.25}$$

implying that S_t is bounded in L^1. Thus, due to Doob's submartingale convergence theorem [47], we coclude that S_t converges to some (almost surely finite) random variable S_∞ so $\lim_{t\to\infty} \frac{S_{t+1}}{\beta_t} = 0$ with probability 1.

Now, by letting $t \to \infty$ in (5.22), we get $F(x^*, Y_t) \to -\infty$, a contradiction. Going back to our original assumption, this shows that there exists a subsequence of X_t which converges to $\arg\min f$ with probability 1, as claimed. □

With all this at hand, we proceed to the proof of our convergence result:

Proof of Theorem 5.4. By the boundedness (and hence compactness) of $\arg\min f$, Proposition 5.9 implies that, with probability 1, there exists some $x^* \in \arg\min f$ such that $X_{t_k} \to x^*$ for some (possibly random) subsequence X_{t_k} of X_t. By the Riemann–Legendre property of h, it follows that $F(x^*, Y_{t_k}) = D(x^*, X_{t_k}) \to 0$ as $k \to \infty$, implying in turn that $\lim_{t\to\infty} D(x^*, X_t) = 0$ (by Proposition 5.7). Since $D(x^*, X_t) \geq \alpha \|X_t - x^*\|_{X_t}^2 \geq \mu \|X_t - x^*\|^2$, we conclude that $X_t \to x^*$, and our proof is complete. □

The above result is a first step towards establishing convergence rate guarantees for NoLips (non-smooth) convex minimization problems. In what will follow, we dive into more detail regarding the interplay between the different NoLips regularity conditions.

5.3 NUMERICAL EVALUATION IN POISSON INVERSE PROBLEMS

For the purposes of validation, we proceed with an application of our algorithmic results to a broad class of Poisson inverse problems that arise in tomography

Figure 5.1: Reconstruction of the Lena test image from a sample contaminated with Poisson noise. Left to right: (*a*) the contaminated sample; (*b*) RMD reconstruction; and (*c*) Poisson likelihood loss at each iteration. The RMD process provides a sharper definition of image features relative to the CMP algorithm (which is the second-best).

problems; the objective of interest here is the Poisson likelihood loss (generalized Kullback–Leibler divergence):

$$f(x) = \sum_{j=1}^{N} \left[x_j \log \frac{x_j}{(Hx)_j} + (Hx)_j - x_j \right] \tag{5.26}$$

where $x \in \mathbb{R}^N_+$ is a vector of Poisson data observations (e.g., pixel intensities) and $H \in \mathbb{R}^{N \times n}$ is an ill-conditioned matrix representing the data-gathering protocol. From an algorithmic point of view we will restrict ourselves to the Riemannian framework. Since the generalized KL objective of (5.26) exhibits an $\mathcal{O}(1/x)$ singularity at the boundary of the orthant, we consider the Poincaré metric $g(x) = \operatorname{diag}(1/x_1, \dots, 1/x_n)$ under which the KL divergence is Riemann–Lipschitz continuous. (Going back to PIP, a suitable Riemannian regularizer for this metric is $h(x) = \sum_{i=1}^{N} 1/x_i^2$, which is 1-strongly convex relative to g. We then run the induced mirror descent algorithm with an online-to-batch conversion mechanism as described in Section 5.2. For reference purposes, we call the resulting process *Riemannian mirror descent* (RMD).

Subsequently, we ran RMD on a Poisson denoising problem for a 384×384 test image contaminated with Poisson noise (so $n \approx 10^5$ in this case). For benchmarking, we also ran a fast variant of the widely used Lucy–Richardson (LR) algorithm [22], and the recent composite mirror prox (CMP) method of [50]; all methods were run with stochastic gradients and the same minibatch size. Because of the "dark area" gradient singularities when $[Hx]_j \to 0$, Euclidean stochastic gradient methods oscillate without converging, so they are not reported. As we see in Fig. 5.1, the RMD process provides the sharpest reconstruction of the original. In particular, after an initial warm-up phase, the last iterate of Riemannian mirror descent consistently outperforms the LR algorithm by 7 orders of magnitude, and CMP by 3. We also note that the Poisson likelihood loss decreases faster under the last iterate of RMD relative to the different algorithmic variants that we tested, exactly because of the hysteresis effect that is inherent to ergodic averaging.

Overall, we note that the introduction of an additional degree of freedom (the choice of Bregman function and that of the local Riemannian norm), makes RMD a particularly flexible and powerful paradigm for loss models with singularities.

We find these results particularly encouraging for further investigations on the interplay between Riemannian geometry and Bregman-proximal methods.

6

NOLIPS MINIMIZATION PROBLEMS

 # This section incorporates material from the paper [4]

In this chapter, we proceed to treat in depth the specific case of convex minimization problems. The key feature which differentiates the present analysis with the results of Chapter 5 is that here we aim to establish order optimal convergence rate guarantees for both deterministic and/or stochastic oracle feedback and (RC) and (RS) respectively.

In doing so, a first issue that should be tackled is to determine what *optimal interpolation* means for the NoLips framework, i.e., to determine the respective worst case lower bounds under (RC) and (RS). To begin with, Theorem 5.3 provides a convergence rate of order $\mathcal{O}(1/\sqrt{T})$ under (RC) which matches the worst-case lower bound presented in Section 2.4.2.

Therefore, the most intriguing part is whether an $\mathcal{O}(1/T^2)$ bound can also be achieved in the (RS) case. This question remained open until Dragomir et al. (2019) established the discrepancy between standard smoothness and (RC). More precisely, they showed that the worst case lower bound under (RS) is $\Omega(1/T)$ and hence the optimal rate for relatively smooth problems does not match the $\mathcal{O}(1/T^2)$ rate for standard Lipschitz smooth problems.[1]

With all this in hand, in Section 6.1 we begin by describing the respective "universal" step-size policy that we will study to interpolate between (RC) and (RS). Moving forward, in Section 6.2 we present our results with respect to a deterministic oracle. More precisely, we show that the time-averages of the (MD) iterates run with our adaptive step-size policy achieve simultaneously order optimal guarantees for both (RS) and (RC) objectives. As an additional feature we establish that the actual iterates of the method - before any averaging occurs- converge towards the solution set.

We conclude this chapter by providing the respective convergence rate guarantees for stochastic case. More precisely, in Section 6.3 we illustrate the respective stochastic rates under (RC) and (RS) via explicit upper bounds.

6.1 A UNIVERSAL STEP-SIZE

Throughout this chapter, the blanket algorithmic template which we will be focusing on is that of (MD), i.e.,

Mirror Descent Template

1 In [49] proposed a tentative path towards faster convergence in certain beyond Lipschitz problems. However, in doing so they require some strict regularity conditions.

		Constr. / Uncon.	Stoch. (L)	(RC)	(RS)	Stoch. (R)
AdaGrad	[39]	✓/✓	✓	×	×	×
AcceleGrad	[67]	×/ ✓	✓	×	×	×
UniXGrad	[60]	✓/ ×	✓	×	×	×
UPGD	[92]	✓/ ✓	×	×	×	×
GMP	[110]	✓/ ✓	×	×	$1/T$	×
AdaProx	[8]	✓/ ✓	×	partial	partial	×
AdaMir	[4]	✓/ ✓	✓	$1/\sqrt{T}$	$1/T$	$1/\sqrt{T}$

Table 6.1: Overview of related adaptive methods for convex optimization. For the purposes of this table, (L) refers to "Lipschitz" and (R) to "relative" continuity or smoothness respectively. In the case of AdaProx, "partial" means that the non-Lipschitz conditions under which it guarantees convergence form a subset of (RC) / (RS). Logarithmic factors are ignored throughout; we also note that the $\mathcal{O}(1/T)$ rate in the column (RS) is, in general, unimprovable [38].

$$X^+ = P_X(-\gamma V) \tag{6.1}$$

where P is a (Bregman) proximal operator associated to a Bregman function h as per Definition 3.1 and $V \in \mathcal{V}^*$. The next important element for our analysis is to define the method's step-size. In the unconstrained case, as we described in Section 2.7 , a popular adaptive choice is the so-called "inverse-sum-of-squares" policy:

$$\gamma_t = 1 \Big/ \sqrt{\textstyle\sum_{s=1}^{t} \|\nabla f(X_s)\|_*^2}, \tag{6.2}$$

where X_t is the series of iterates produced by the algorithm. However, in relatively continuous/smooth problems, this definition encounters two crucial issues. First, because the gradient of f is unbounded (even over a bounded domain), the denominator of (6.2) may grow at an uncontrollable rate, leading to a step-size policy that vanishes too fast to be of any practical use. The second is that, if the problem is constrained, the extra terms entering the denominator of γ_t do not vanish as the algorithm approaches a solution, so (6.2) may still be unable to exploit the smoothness of the objective.

We begin by addressing the second issue. In the Euclidean case, the key observation is that the difference $\|x^+ - x\|$ must always vanish near a solution (even near the boundary), so we can use it as a proxy for $\nabla f(x)$ in constrained problems. This idea is formalized by the notion of the *gradient mapping* [89] that can be defined here as

Gradient Mapping

$$\delta = \|x^+ - x\|/\gamma. \tag{6.3}$$

On the other hand, in a Bregman setting, the prox-mapping tends to deflate gradient steps, so the norm difference between two successive iterates x^+ and x of (MD) could be very small relative to the oracle signal that was used to generate the update. As a result, the Euclidean residual (6.3) could lead to a disproportionately large step-size that would be harmful for convergence. For this reason, we consider

a gradient mapping that takes into account the Bregman geometry of the method and we set

$$\delta = \sqrt{D(x, x^+) + D(x^+, x)}/\gamma. \tag{6.4}$$

Bregman Gradient Mapping

Obviously, when $h(x) = (1/2)\|x\|_2^2$, we readily recover the definition of the Euclidean gradient mapping (6.3). In general however, by the strong convexity of h, the value of this "Bregman residual" exceeds the corresponding Euclidean definition, so the induced step-size exhibits smoother variations that are more adapted to the framework in hand.

Having all this hand, we are in a position to put everything together and define our adaptive (MD) method. In this regard, combining the abstract template (MD) with the Bregman residual and "inverse-sum-of-squares" approach discussed above, we will consider the recursive policy

AdaMir Algorithm

$$X_{t+1} = P_{X_t}(-\gamma_t V_t) \tag{6.5}$$

with V_t, $t = 1, 2, \ldots$, coming from a oracle model of the form (SFO), and with γ_t defined as

$$\gamma_t = \frac{1}{\sqrt{\sum_{s=0}^{t-1} \delta_s^2}} \qquad \text{with } \delta_s^2 = \frac{D(X_s, X_{s+1}) + D(X_{s+1}, X_s)}{\gamma_s^2}. \tag{Adapt}$$

In the sequel, we will use the term to refer interchangeably to the update $X_t \leftarrow X_{t+1}$ and the specific step-size policy used within. The convergence properties of (MD) run with (Adapt); abbreviated as *AdaMir* are discussed in detail in the next two sections in both deterministic and stochastic problems.

6.2 DETERMINISTIC ANALYSIS

We are now in a position to state our main convergence results for our method. We begin with the deterministic analysis ($\sigma = 0$), treating both the method's "time-average" as well as the induced trajectory of query points; the analysis for the stochastic case ($\sigma > 0$) is presented in the next section.

6.2.1 *Ergodic convergence and rate interpolation*

We begin by showing the convergence rate guarantees of the method's "time-averaged" state, i.e., $\overline{X}_T = (1/T)\sum_{t=1}^T X_t$. More precisely, we show that our method simultaneously achieves an $\mathcal{O}(1/\sqrt{T})$ value convergence rate under (RC) and $\mathcal{O}(1/T)$ under (RS). Moreover, if both regularity conditions are satisfied we are able to obtain a more detailed "any-time" rate. Formally, we have the following result.

Theorem 6.1. *Let X_t, $t = 1, 2, \ldots$, denote the sequence of iterates generated by AdaMir, and let $D_1 = D(x^*, X_1)$. Then, AdaMir simultaneously enjoys the following guarantees:*

1. *If f satisfies (RC), we have:*

$$f(\overline{X}_T) - \min f \leq \frac{\sqrt{2}G\left[D_1 + 8G^2/\delta_0^2 + 2\log(1 + 2G^2T/\delta_0^2)\right]}{\sqrt{T}} + \frac{3\sqrt{2}G + 4G^2/\delta_0^2}{T}. \tag{6.6}$$

2. *If f satisfies (RS), we have $f(\overline{X_T}) - \min f = \mathcal{O}(D_1/T)$.*

3. *If f satisfies (RS) and (RC), we have:*

$$f(\overline{X}_T) - \min f \leq \left[f(X_1) - \min f + \left(2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{4\beta^2}{\delta_0^2}\right)\beta\right]^2 \frac{D_1}{T}. \tag{6.7}$$

Universality Guarantees (Deterministic)

As we already mentioned Theorem 6.1 shows that, up to logarithmic factors, AdaMir achieves the optimal lower bounds for objectives which satisfy either belong to the (RC) oracle complexity class or satisfy (RS). The key element of the proof is to the following regret bound:

Regret Analysis of AdaMir

Proposition 6.2. *With notation as in Theorem 6.1, AdaMir enjoys the regret bound*

$$\sum_{t=1}^{T}[f(X_t) - f(x^*)] \leq \frac{D_1}{\gamma_T} + \frac{\sum_{t=1}^{T}\gamma_t^2\delta_t^2}{\gamma_T} + \sum_{t=1}^{T}\gamma_t\delta_t^2. \tag{6.8}$$

Proof. By the convexity of f and the definition of the Bregman proximal step in Proposition 4.3, we have:

$$f(X_t) - f(x^*) \leq \langle V_t, X_t - x^*\rangle \leq \frac{1}{\gamma_t}\langle \nabla h(X_t) - \nabla h(X_{t+1}), X_t - x^*\rangle. \tag{6.9}$$

Hence, by applying again the three-point identity (Lemma 4.2), we obtain:

$$\begin{aligned} f(X_t) - f(x^*) &\leq \frac{D(x^*, X_t) - D(x^*,)}{\gamma_t} + \frac{D(X_t,)}{\gamma_t} \\ &\leq \frac{D(x^*, X_t) - D(x^*,)}{\gamma_t} + \frac{D(X_t,) + D(, X_t)}{\gamma_t} \\ &= \frac{D(x^*, X_t) - D(x^*,)}{\gamma_t} + \gamma_t\delta_t^2 \end{aligned} \tag{6.10}$$

where the last equality follows readily from the definition (6.4) of δ_t. Therefore, by summing through $t = 1, 2, \ldots, T$, we obtain:

$$\sum_{t=1}^{T}[f(X_t) - f(x^*)] \leq \frac{D(x^*, X_1)}{\gamma_1} + \sum_{t=2}^{T}\left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right]D(x^*, X_t) + \sum_{t=1}^{T}\gamma_t\delta_t^2. \tag{6.11}$$

Now, we are left to bound from above the second term on the right-hand side (RHS) of (6.11). By the second part of Proposition 4.3, we have:

$$\begin{aligned} D(x^*, X_{s+1}) &\leq D(x^*, X_s) - \gamma_t\langle V_t, X_t - x^*\rangle + D(X_s, X_{s+1}) \\ &\leq D(x^*, X_s) + D(X_s, X_{s+1}) \\ &\leq D(x^*, X_s) + D(X_{s+1}, X_s) + D(X_s, X_{s+1}) \end{aligned} \tag{6.12}$$

Thus, by telescoping through $s = 1, 2, \ldots, t$, we obtain:

$$\begin{aligned} D(x^*, X_t) &\leq D(x^*, X_1) + \sum_{s=1}^{t} [D(X_s, X_{s+1}) + D(X_{s+1}, X_s)] \\ &\leq D(x^*, X_1) + \sum_{s=1}^{T} [D(X_s, X_{s+1}) + D(X_{s+1}, X_s)] \\ &= D(x^*, X_1) + \sum_{s=1}^{T} \gamma_s^2 \delta_s^2 \end{aligned} \tag{6.13}$$

where the last equality follows from the definition (6.4) of δ_t. So, summarizing

$$\begin{aligned} \sum_{t=2}^{T} \left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right] D(x^*, X_t) &\leq \sum_{t=2}^{T} \left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right] \left(D(x^*, X_1) + \sum_{s=1}^{T} \gamma_s^2 \delta_s^2\right) \\ &\leq \frac{D(x^*, X_1)}{\gamma_T} - \frac{D(x^*, X_1)}{\gamma_1} + \sum_{s=1}^{T} \gamma_s^2 \delta_s^2 \cdot \sum_{t=1}^{T} \left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right] \\ &\leq \frac{D(x^*, X_1)}{\gamma_T} - \frac{D(x^*, X_1)}{\gamma_1} + \frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\gamma_T}. \end{aligned} \tag{6.14}$$

Hence, by combining the above with (6.11), our claim follows. □

The proof of Proposition 6.2 hinges on the specific definition of the adaptive step-size, and the exact functional form of the regret bound (6.8) plays a crucial role in the sequel. Specifically, under the regularity conditions (RC) and (RS), we respectively obtain the following key lemmas:

Boundedness of the Residuals Under (RC)

Lemma 6.3. *Under* (RC)*, the sequence of the Bregman residuals δ_t of is bounded as $\delta_t^2 \leq 2G^2$ for all $t \geq 1$.*

Proof. By the definition of the Bregman proximal step in (MD) and Proposition 4.3, we have:

$$\begin{aligned} D(X_t, X_{t+1}) + D(X_{t+1}, X_t) &= \langle \nabla h(X_t) - \nabla h(X_{t+1}), X_t - X_{t+1} \rangle \\ &\leq \gamma_t \langle V_t, X_t - X_{t+1} \rangle. \end{aligned} \tag{6.15}$$

Hence, by invoking (RC) we get:

$$\begin{aligned} D(X_t, X_{t+1}) + D(X_{t+1}, X_t) &\leq \gamma_t G \sqrt{2D(X_{t+1}, X_t)} \\ &\leq \gamma_t G \sqrt{2\left[D(X_{t+1}, X_t) + D(X_t, X_{t+1})\right]} \end{aligned} \tag{6.16}$$

We thus get:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq 2\gamma_t^2 G^2. \tag{6.17}$$

Hence, by the definition (6.4) of δ_t^2, we conclude that

$$\delta_t^2 = \frac{D(X_t, X_{t+1}) + D(X_{t+1}, X_t)}{\gamma_t^2} \leq 2G^2. \tag{6.18}$$

□

Summability of the Residuals Under (RS)

Lemma 6.4. *Under (RS), the sequence of the Bregman residuals δ_t is square-summable, i.e., $\sum_t \delta_t^2 < \infty$. Consequently, the method's step-size converges to a positive limit $\gamma_\infty > 0$.*

Proof. Since the adaptive step-size policy γ_t is decreasing and bounded from below $(\gamma_t) \geq 0$ we get that its limit exists,i.e.,

$$\lim_{t \to +\infty} \gamma_t = \gamma_\infty \quad \text{for some} \quad \gamma_\infty \geq 0 \tag{6.19}$$

Assume that $\gamma_\infty = 0$. By Proposition 3.3, we obtain:

$$\begin{aligned} f(X_{t+1}) &\leq f(X_t) + \langle \nabla f(X_t), X_{t+1} - X_t \rangle + \beta D(X_{t+1}, X_t) \\ &\leq f(X_t) - \frac{1}{\gamma_t} D(X_t, X_{t+1}) \\ &\quad - \frac{1}{\gamma_t} D(X_{t+1}, X_t) + \beta \left[D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \right] \end{aligned} \tag{6.20}$$

whereas by recalling the definition of the residuals (Adapt) the above can be rewritten as follows:

$$f(X_{t+1}) \leq f(X_t) - \gamma_t \delta_t^2 + \beta \gamma_t^2 \delta_t^2 = f(X_t) - \frac{1}{2} \gamma_t \delta_t^2 - \frac{1}{2} \gamma_t \delta_t^2 + \beta \gamma_t^2 \delta_t^2 \tag{6.21}$$

Moreover, by rearranging and factorizing the common term $\gamma_t \delta_t^2$ we get:

$$\frac{1}{2} \gamma_t \delta_t^2 \leq f(X_t) - f(X_{t+1}) + \gamma_t \delta_t^2 \left[\beta \gamma_t - \frac{1}{2} \right] \tag{6.22}$$

Now, by the fact that $\left[\beta \gamma_t - \frac{1}{2} \right] \leq 0$ for $\gamma_t \leq 1/2\beta$ and the fact that γ_t converges to 0 by assumption, we get that there exists some $t_0 \in \mathbb{N}$ such that:

$$\left[\beta \gamma_t - \frac{1}{2} \right] \leq 0 \quad \text{for all } t > t_0 \tag{6.23}$$

Hence, by telescoping for $t = 1, 2, \ldots, T$ for sufficiently large T, we have

$$\begin{aligned} \frac{1}{2} \sum_{t=1}^{T} \gamma_t \delta_t^2 &\leq f(X_1) - f(X_{T+1}) + \sum_{t=1}^{t_0} \left[\beta \gamma_t - \frac{1}{2} \right] \gamma_t \delta_t^2 \\ &\leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \left[\beta \gamma_t - \frac{1}{2} \right] \gamma_t \delta_t^2 \end{aligned} \tag{6.24}$$

Now, by applying the (LHS) of Lemma A.4 we get:

$$\frac{1}{2} \left[\frac{1}{\gamma_T} - \delta_0 \right] \leq \frac{1}{2} \sqrt{\delta_0^2 + \sum_{t=1}^{T-1} \gamma_t \delta_t^2} \leq \sum_{t=1}^{T} \gamma_t \delta_t^2 \leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \left[\beta \gamma_t - \frac{1}{2} \right] \gamma_t \delta_t^2 \tag{6.25}$$

Since $\gamma_t \to 0$ we get that $1/\gamma_t \to +\infty$ and hence the above yields that $+\infty \leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \left[\frac{\beta}{K} \gamma_t - \frac{1}{2} \right] \gamma_t \delta_t^2$, a contradiction. Therefore we get that:

$$\lim_{t \to +\infty} \gamma_t = \gamma_\infty > 0 \tag{6.26}$$

Moreover, by recalling the definition of the adaptive step-size policy γ_t:

$$\gamma_t = \frac{1}{\sqrt{\delta_0^2 + \sum_{s=1}^{t-1} \delta_s^2}} \tag{6.27}$$

and after rearranging we obtain:

$$\sum_{s=1}^{t-1} \delta_s^2 = \frac{1}{\gamma_t^2} - \delta_0^2 \tag{6.28}$$

and therefore by taking limit on both sides we obtain:

$$\sum_{t=1}^{+\infty} \delta_t^2 = \lim_{t \to +\infty} \sum_{s=1}^{t-1} \delta_s^2 = \lim_{t \to +\infty} \frac{1}{\gamma_t^2} - \delta_0^2 = \frac{1}{\gamma_\infty^2} - \delta_0^2 < +\infty \tag{6.29}$$

and hence the result follows. □

As we explain below, the boundedness estimate of Lemma 6.3 is necessary to show that the iterates of the method do not explode; however, without further assumptions, it is not possible to sharpen this bound. The principal technical difficulty – and an important novelty of our analysis – is the stabilization of the step-size to a strictly positive limit in Lemma 6.4. This property plays a crucial role because the method is not slowed down near a solution. To the best of our knowledge, there is no comparable result for the step-size of parameter-agnostic methods in the literature.[2]

Armed with these two lemmas, we will establish below the following series of estimates:

1. Under (RC), the terms in the RHS of (6.8) can be bounded respectively as $\mathcal{O}(G\sqrt{T})$, $\mathcal{O}(\log(G^2T)\sqrt{T})$, and $\mathcal{O}(G\sqrt{T})$. As a result, we obtain an $\tilde{\mathcal{O}}(1/\sqrt{T})$ rate of convergence.
2. Under (RS), all terms in the RHS of (6.8) can be bounded as $\mathcal{O}(1)$, so we obtain an $\mathcal{O}(1/T)$ convergence rate for $\bar{X}_T$.

We formalize all this below:

Proof of Theorem 6.1. Repeating the statement of Proposition 6.2, the iterate sequence X_t generated by ADM enjoys the bound:

$$\sum_{t=1}^{T} [f(X_t) - f(x^*)] \leq \frac{D(x^*, X_1)}{\gamma_T} + \frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\gamma_T} + \sum_{t=1}^{T} \gamma_t \delta_t^2 \tag{6.8}$$

We now proceed to bound each term on the RHS of (6.8) from above. We consider three separate cases, first only under (RC),then under (RS) and finally when both (RC) and (RS) holds.

2 In more detail, [67], [68] and [60] establish the summability of a suitable residual sequence to sharpen the $\mathcal{O}(1/\sqrt{T})$ rate in their respective contexts, but this does not translate to a step-size stabilization result. Under (RC)/(RS), controlling the method's step-size is of vital importance because the gradients that enter the algorithm may be unbounded even over a bounded domain; this crucial difficulty does not arise in any of the previous works on adaptive methods for ordinary Lipschitz problems.

Analysis Under (RC)

1. Under (RC): We begin with problems satisfying (RC).

   - For the first term, Lemma 6.3 gives:

$$\frac{D(x^*, X_1)}{\gamma_T} = D(x^*, X_1)\sqrt{\sum_{t=0}^{T-1} \delta_t^2} \leq D(x^*, X_1)\sqrt{2G^2T}. \tag{6.30}$$

   - For the second term, we have:

$$\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 \leq \sum_{t=1}^{T} \frac{\delta_t^2}{\sum_{s=0}^{t-1} \delta_s^2} = \sum_{t=1}^{T} \frac{\delta_t^2}{\delta_0^2 + \sum_{s=1}^{t-1} \delta_s^2}. \tag{6.31}$$

     Hence, by Lemmas 6.3 and A.5, we get:

$$\begin{aligned}\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 &\leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\left(1 + \sum_{t=1}^{T-1} \frac{\delta_t^2}{\delta_0^2}\right) \\ &= 2 + \frac{8G^2}{\delta_0^2} + 2\log\left(\sum_{t=0}^{T-1} \frac{\delta_t^2}{\delta_0^2}\right) \\ &\leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{2G^2T}{\delta_0^2}.\end{aligned} \tag{6.32}$$

   - Finally, for the third term, we get:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 = \sum_{t=1}^{T} \frac{\delta_t^2}{\sqrt{\sum_{s=0}^{t-1} \delta_t^2}} = \sum_{t=1}^{T} \frac{\delta_t^2}{\sqrt{\delta_0^2 + \sum_{s=1}^{t-1} \delta_t^2}}. \tag{6.33}$$

     Hence, Lemmas 6.3 and A.4 again yield:

$$\begin{aligned}\sum_{t=1}^{T} \gamma_t \delta_t^2 &\leq \frac{4G^2}{\delta_0} + 3\sqrt{2}G + 3\sqrt{\delta_0^2 + \sum_{t=1}^{T-1} \delta_t^2} \\ &\leq \frac{4G^2}{\delta_0} + 3\sqrt{2}G + 3\sqrt{\sum_{t=0}^{T-1} \delta_t^2} \\ &\leq \frac{4G^2}{\delta_0} + 3\sqrt{2}G + 3\sqrt{2G^2T}.\end{aligned} \tag{6.34}$$

   The claim of Theorem 6.1 then follows by combining the above within the regret bound (6.8).

Analysis Under (RS)

2. Under (RS): We now turn to problems satisfying (RS). Recalling Lemma 6.4, we shall revisit the terms of (6.8). In particular, we have:

   - For the first term, we have:

$$\frac{D(x^*, X_1)}{\gamma_T} = D(x^*, X_1)\sqrt{\sum_{t=0}^{T-1} \delta_t^2} \leq \frac{D(x^*, X_1)}{\gamma_\infty} \tag{6.35}$$

- For the second term, we have:

$$\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 \leq \frac{1}{\delta_0^2} \sum_{t=1}^{T} \delta_t^2 \leq \frac{1}{\delta_0^2 \gamma_\infty^2} - 1 \tag{6.36}$$

- Finally, for the third term, we get:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq \frac{1}{\delta_0} \sum_{t=1}^{T} \delta_t^2 \leq \frac{1}{\delta_0 \gamma_\infty^2} - \delta_0 \tag{6.37}$$

Combining all the above, the result follows.

3. Under (RS) and (RC): Finally, we consider objectives where (RC) and (RS) hold simultaneously. Now, by working in the same spirit as in the proof of Lemma 6.4 we get:

Analysis Under (RC) & (RS)

$$\frac{1}{2} \gamma_t \delta_t^2 \leq f(X_t) - f(X_{t+1}) + \gamma_t \delta_t^2 \left[\beta \gamma_t - \frac{1}{2} \right] \tag{6.38}$$

which after telescoping $t = 1, \ldots, T$ it becomes:

$$\frac{1}{2} \sum_{t=1}^{T} \gamma_t \delta_t^2 \leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{T} \gamma_t \delta_t^2 \left[\beta \gamma_t - \frac{1}{2} \right] \tag{6.39}$$

Now, after denoting:

$$t_0 = \max\{t \in \mathbb{N} : 1 \leq t \leq T \text{ such that } \gamma_t \geq \frac{1}{2\beta}\} \tag{6.40}$$

and decomposing the sum we get:

$$\begin{aligned} \frac{1}{2} \sum_{t=1}^{T} \gamma_t \delta_t^2 &\leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \gamma_t \delta_t^2 \left[\beta \gamma_t - \frac{1}{2} \right] + \sum_{t=t_0+1}^{T} \gamma_t \delta_t^2 \left[\beta \gamma_t - \frac{1}{2} \right] \\ &\leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \gamma_t \delta_t^2 \left[\beta \gamma_t - \frac{1}{2} \right] \\ &\leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \beta \sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 \end{aligned} \tag{6.41}$$

On the other hand, by applying Lemma A.5, we have:

$$\begin{aligned} \sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 &\leq 2 + \frac{8G^2}{\delta_0^2} + 2 \log \left(1 + \sum_{t=1}^{t_0 - 1} \frac{\delta_t^2}{\delta_0^2} \right) \\ &= 2 + \frac{8G^2}{\delta_0^2} + 2 \log \left(\frac{1}{\delta_0^2} \left[\delta_0^2 + \sum_{t=1}^{t_0 - 1} \delta_t^2 \right] \right) \\ &= 2 + \frac{8G^2}{\delta_0^2} + 2 \log \frac{1}{\delta_0^2 \gamma_{t_0}^2} \end{aligned} \tag{6.42}$$

and by the definition of t_0 we get:

$$\sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 \leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{4\beta^2}{\delta_0^2}. \tag{6.43}$$

which yields:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq f(X_1) - \min_{x\in\mathcal{X}} f(x) + \beta\left[2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{4\beta^2}{\delta_0^2}\right] \tag{6.44}$$

The result then follows by plugging in the above bounds in (6.8).

□

Having established the convergence rate for the time-average iterates as output of our method, we proceed with examining the asymptotic behaviour of the iterates of the method per se.

6.2.2 *Other modes of convergence*

In complement to the analysis above, we provide below a spinoff result for the method's "last iterate", i.e., the actual trajectory of queried points. In particular, these results become more appealing for the more general non-convex landscapes. Formalizing the blanket assumption in order to get the said last-iterate convergence results we shall assume throughout this section that the underlying objective f satisfies the so-called 'secant condition" [26, 121]:

Weak Secant Inequality

$$\inf\{\langle \nabla f(x), x - x^*\rangle : x^* \in \arg\min f, x \in \mathcal{K}\} > 0 \tag{SI}$$

for every closed subset $\mathcal{K}$ of $\mathcal{X}$ that is separated by neighborhoods from $\arg\min f$. The formal statement is as follows.

Last Iterate Convergence

Theorem 6.5. *Suppose that f satisfies* (RC) *or* (RS) *along with* (SI) *condition. Then X_t converges to* $\arg\min f$.

The main idea of the proof consists of two steps. The first key step is to show that, under (RC) ∪ (RS), the iterates have convergent subsequences, i.e., $\liminf f(X_t) = \min f$. In particular, we have the following result.

Extracting a Convergenr Sub-sequence

Proposition 6.6. *Assume that f satisfies* (RC) *or* (RS) *along with the* (SI) *and X_t are the iterates generated by* AdaMir. *Then there exists a subsequence X_{k_t} which converges to the solution set $\mathcal{X}^*$.*

Proof. Assume to the contrary that the sequence X_t generated by AMD admits no limit points in $\mathcal{X}^* = \arg\min f$. Then there exists a (non-empty) closed set $\mathcal{K} \subseteq \mathcal{X}$ which is separated by neighborhoods from $\arg\min f$ and is such that $X_t \in \mathcal{C}$ for all sufficiently large t. Then, by relabelling X_t if necessary, we can assume without loss of generality that $X_t \in \mathcal{K}$ for all $t \in \mathbb{N}$. Thus, we have:

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) - \gamma_t\langle \nabla f(X_t), X_t - x^*\rangle + D(X_t, X_{t+1})$$

$$\begin{aligned} &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + [D(X_t, X_{t+1}) + D(X_{t+1}, X_t)] \\ &= D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \delta_t^2 \end{aligned} \tag{6.45}$$

with the last equality being obtained by the definition of (6.4). Now, applying (SI) we get:

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) - \gamma_t \delta(\mathcal{K}) + \gamma_t^2 \delta_t^2 \tag{6.46}$$

with $\delta(\mathcal{K}) = \inf\{\langle \nabla f(x), x - x^* \rangle : x^* \in \arg\min f, x \in \mathcal{K}\} > 0$. Hence, by telescoping $t = 1, \ldots, T$, factorizing and setting $\beta_t = \sum_{t=1}^{T} \gamma_t$ we have:

$$D(x^*, X_{T+1}) \leq D(x^*, X_1) - \beta_t \left[\delta(\mathcal{K}) - \frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\beta_t} \right] \tag{6.47}$$

(6.47) will be the crucial lemma that will walk throughout our analysis. In particular, we will treat the different regularity conditions of (RC) and (RS) seperately.

1. The (RC) case: Assume that f satisfies (RC). By examining the asymptotic behaviour of each term individually, we obtain:

   Sub-sequence Under (RC)

   - For the term $\beta_T = \sum_{t=1}^{T} \gamma_t$, we have:

   $$\beta_T = \sum_{t=1}^{T} \frac{1}{\sqrt{\delta_0^2 + \sum_{j=1}^{t-1} \delta_t^2}} \geq \sum_{t=1}^{T} \frac{1}{\sqrt{\delta_0^2 + 2G^2 t}} \tag{6.48}$$

   which yields that $\beta_T \to +\infty$ and more precisely $\beta_T = \Omega(\sqrt{T})$.

   - For the term $\frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\beta_T}$, for the numerator we have:

   $$\begin{aligned} \sum_{t=1}^{T} \gamma_t^2 \delta_t^2 &= \sum_{t=1}^{T} \frac{\delta_t^2}{\delta_0^2 + \sum_{j=1}^{t-1} \delta_j^2 / \delta_0^2} \\ &\leq 2 + 8G^2/\delta_0^2 + 2\log(1 + \sum_{t=1}^{T-1} \delta_t^2/\delta_0^2) \\ &\leq 2 + 8G^2/\delta_0^2 + 2\log(1 + 2G^2 T/\delta_0^2) \end{aligned} \tag{6.49}$$

   which yields that $\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 = \mathcal{O}(\log T)$, and combined with the fact that $\beta_t = \Omega(\sqrt{T})$ we readily get:

   $$\frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\beta_T} \to 0 \tag{6.50}$$

   So, combining all the above and letting $T \to +\infty$ in (6.47), we get that $D(x^*, X_{T+1}) \to -\infty$, a contradiction. Therefore, the result under (RC) follows.

2. The (RS) case: On the other hand, assume that f satisfies (RS). Recalling Lemma 6.4 and the fact that γ_t is decreasing we have:

   Sub-sequence Under (RS)

   $$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq \sum_{t=1}^{+\infty} \delta_t^2 < +\infty \tag{6.51}$$

which by working as in Lemma 6.4 also yields:

$$\lim_{t\to+\infty} \gamma_t = \gamma_\infty > 0 \tag{6.52}$$

Additionally, since γ_t is decreasing and bounded we also have that $\gamma_\infty = \inf_t \gamma_t$. Now, we shall re-examine the terms of (6.47). More precisely, we have:

- For β_T we have:

$$\beta_T = \sum_{t=1}^{T} \gamma_t \geq \gamma_\infty \sum_{t=1}^{T} 1 = \gamma_\infty T \tag{6.53}$$

  which in turn yields that $\beta_T \to +\infty$ and more precisely $\beta_T = \Omega(T)$.

- For the term $\frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\beta_T}$, for the numerator we have by the fact that $\gamma_t \leq 1/\delta_0$ and Lemma 6.4:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq \frac{1}{\delta_0} \sum_{t=1}^{T} \delta_t^2 < +\infty \tag{6.54}$$

  which yields that $\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 = \mathcal{O}(1)$, which combined with (6.53) gives that:

$$\frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\beta_T} \to 0 \tag{6.55}$$

so, again combing the above and letting $T \to +\infty$ in (6.47), we get that $D(x^*, X_{T+1}) \to -\infty$, a contradiction. Therefore, the result follows also under (RS).

□

Now, given the existence of a convergent subsequence, the rest of our proof strategy branches out depending on whether f satisfies (RC) or (RS). Under (RS), the analysis relies on arguments that involve a quasi-Fejér argument as in [26, 36]; this is described by the following lemma.

Quasi-Fejer Seuences

Lemma 6.7. *Let $\chi \in (0,1]$, $(\alpha_t)_{t\in\mathbb{N}}$, $(\beta_t)_{t\in\mathbb{N}}$ non-negative sequences and $(\varepsilon_t)_{t\in\mathbb{N}} \in l^1(\mathbb{N})$ such that $t = 1, 2, \ldots$:*

$$\alpha_{t+1} \leq \chi\alpha_t - \beta_t + \varepsilon_t \tag{6.56}$$

Then, α_t converges.

Proof. First, one shows that $\alpha_{t\in\mathbb{N}}$ is a bounded sequence. Indeed, one can derive directly that:

$$\alpha_{t+1} \leq \chi^{t+1}\alpha_0 + \sum_{k=0}^{t} \chi^{t-k}\varepsilon_k \tag{6.57}$$

Hence, $(\alpha_t)_{t\in\mathbb{N}}$ lies in $[0, \alpha_0 + \varepsilon]$, with $\varepsilon = \sum_{t=0}^{+\infty} \varepsilon_t$. Now, one is able to extract a convergent subsequence $(\alpha_{k_t})_{t\in\mathbb{N}}$, let say $\lim_{t\to+\infty} \alpha_{k_t} = \alpha \in [0, \alpha_0 + \varepsilon]$ and fix

$\delta > 0$. Then, one can find some t_0 such that $\alpha_{k_{t_0}} - \alpha < \frac{\delta}{2}$ and $\sum_{m>t_{k_{t_0}}} \varepsilon_m < \frac{\delta}{2}$. That said, we have:

$$0 \leq \alpha_t \leq \alpha_{k_{t_0}} + \sum_{m>t_{k_{t_0}}} \varepsilon_m < \frac{\delta}{2} + \alpha + \frac{\delta}{2} = \alpha + \delta \tag{6.58}$$

Hence, $\limsup_t \alpha_t \leq \liminf_t \alpha_t + \delta$. Since, δ is chosen arbitrarily the result follows. □

However, under (RC), the quasi-Fejér property fails, so we prove the convergence of X_t via a novel induction argument that shows that the method's iterates remain trapped within a Bregman neighborhood of x^* if they enter it with a sufficiently small step-size. Therefore, we provide the relevant details of Theorem 6.5.

Proof of Theorem 6.5. We will divide our proof in two parts by distinguishing the two different regularity cases.

1. The (RC) case: Given that γ_t is decreasing and bounded from below we have that its limit exists, denoted by $\gamma_\infty \geq 0$. We shall consider two cases: *Last Iterate Under (RC)*

   a) $\gamma_\infty > 0$: Following the same reasoning with Lemma 6.4 we get that:

   $$\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 \leq \sum_{t=1}^{+\infty} \delta_t^2 < +\infty \tag{6.59}$$

   Hence, by recalling the inequality:

   $$D(x^*, X_{t+1}) \leq D(x^*, X_t) + \gamma_t^2 \delta_t^2 \quad \text{for all} \quad x^* \in \mathcal{X}^* \tag{6.60}$$

   whereas after taking infima on both sides with respect to $\mathcal{X}^*$, we get:

   $$\inf_{x^* \in \mathcal{X}^*} D(x^*, X_{t+1}) \leq \inf_{x^* \in \mathcal{X}^*} D(x^*, X_t) + \gamma_t^2 \delta_t^2 \tag{6.61}$$

   and since the sequence $\gamma_t^2 \delta_t^2$ is summable we can directly apply Lemma 6.7 which yields that the sequence $\inf_{x^* \in \mathcal{X}^*} D(x^*, X_t)$ is convergent. Now, since by Proposition 6.6, AMD possesses a convergent subsequence towards the solution set $\mathcal{X}^*$ the result follows.

   b) $\gamma_\infty = 0$: Pick some $\varepsilon > 0$ and consider the Bregman zone:

   $$D_\varepsilon = \{x \in \mathcal{X} : D(\mathcal{X}^*, x) < \varepsilon\}. \tag{6.62}$$

   Then, it suffices to show that $X_t \in D_\varepsilon$ for all sufficiently large t. In doing so, consider the inequality:

   $$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \delta_t^2 \\ &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \frac{2G^2}{K} \end{aligned} \tag{6.63}$$

   with the second inequality being obtained by Lemma 6.3. To proceed, assume inductively that $X_t \in D_\varepsilon$. By the regularity assumptions of the

regularizer h, it follows that there exists a $\delta-$ neighbourhood contained in the closure of $D_{\varepsilon/2}$. So, by the (SI) condition we have:

$$\langle f(x), x - x^* \rangle \geq c > 0 \text{ for some } c \equiv c(\varepsilon) > 0 \text{ and for all } x \in D_\varepsilon \setminus D_{\varepsilon/2} \text{ and } x^* \in \mathcal{X}^* \tag{6.64}$$

We consider two cases:

- $X_t \in D_\varepsilon \setminus D_{\varepsilon/2}$: In. this case, we have:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \frac{2G^2}{K} \\ &\leq D(x^*, X_t) - \gamma_t c + \gamma_t^2 \frac{2G^2}{K} \end{aligned} \tag{6.65}$$

  Thus, provided that $\gamma_t \leq \frac{cK}{2G^2}$ we get that $D(x^*, X_{t+1}) \leq D(x^*, X_t)$. Hence, by taking infima on both sides relative to $x^* \in \mathcal{X}^*$, we get that $D(\mathcal{X}^*, X_{t+1}) \leq D(\mathcal{X}^*, X_t) < \varepsilon$.

- $X_t \in D_{\varepsilon/2}$: In this case, we have:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \frac{2G^2}{K} \\ &\leq D(x^*, X_t) + \gamma_t^2 \frac{2G^2}{K} \end{aligned} \tag{6.66}$$

  with the second inequality being obtained by the optimality of x^*. Now, provided that $\gamma_t^2 \leq \frac{\varepsilon K}{4G^2}$ or equivalently $\gamma_t \leq \frac{\sqrt{\varepsilon K}}{2G}$ we have:

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) + \frac{\varepsilon}{2} \tag{6.67}$$

  whereas again by taking infima on both sides we get that $D(\mathcal{X}^*, X_{t+1}) \leq D(\mathcal{X}^*, X_t) + \frac{\varepsilon}{2} < \varepsilon$.

Hence, summarizing we have that $X_{t+1} \in D_\varepsilon$ whenever $X_t \in D_\varepsilon$ and $\gamma_t \leq \min\{\frac{cK}{2G^2}, \frac{\sqrt{\varepsilon K}}{2G}\}$. Hence, the result follows by. Proposition 6.6 and the fact that $\gamma_t \to 0$.

Last Iterate Under (RS)

2. The (RS) case Recall that we have the following inequality,

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) + \gamma_t^2 \delta_t^2 \text{ for all } x^* \in \mathcal{X}^* \tag{6.68}$$

whereas taking infima on both sides relative to $\mathcal{X}^*$ we readily get:

$$\inf_{x^* \in \mathcal{X}^*} D(x^*, X_{t+1}) \leq \inf_{x^* \in \mathcal{X}^*} D(x^*, X_t) + \gamma_t^2 \delta_t^2 \tag{6.69}$$

Now, by recalling that by Lemma 6.4, we have $\gamma_t^2 \delta_t^2$ is summable. we can apply directly Lemma 7.10. Thus, we have the sequence $\inf_{x^* \in \mathcal{X}^*} D(x^*, X_t)$ is convergent. Moreover, Proposition 6.6 guarantees that there a subsequence of $\inf_{x^* \in \mathcal{X}^*} \|X - x^*\|^2$ that converges to 0. We obtain that there exists also a subsequence of $\inf_{x^* \in \mathcal{X}^*} D(x^*, X_t)$ that converges to 0 and since $\inf_{x^* \in \mathcal{X}^*} D(x^*, X_t)$ is convergent, we readily get that:

$$\inf_{x^* \in \mathcal{X}^*} \|x^* - X_t\|^2 \leq \inf_{x^* \in \mathcal{X}^*} D(x^*, X_t) \to 0 \tag{6.70}$$

and the proof is complete.

□

Even more generally, Lemma 6.4 also allows us to derive results for general non-convex problems. Indeed, the proof of Proposition 3.3 shows that $\min_{1\leq t\leq T}\delta_t^2 = \mathcal{O}(1/T)$ *without* requiring any properties on f other than (RS). As a result, we conclude that the "best iterate" of the method – i.e., the iterate with the least residual – decays as $\mathcal{O}(1/\sqrt{T})$. This fact partially generalizes a similar result obtained in [68, 112] for AdaGrad applied to non-convex problems; however, an in-depth discussion of this property would take us too far afield, so we do not attempt it.

6.3 THE STOCHASTIC CASE

In this last section, we focus on the stochastic case ($\sigma > 0$). Our main results here are as follows.

Theorem 6.8. *Let X_t, $t = 1, 2, \ldots$, denote the sequence of iterates generated by* AdaMir, *and let $D_1 = D(x^*, X_1)$ and $G_\sigma = G + \sigma/\sqrt{K}$. Then, under (RC), we have*

$$\mathbb{E}\left[f(\bar{X}_T) - f(x^*)\right] \leq (D_1 + H)\sqrt{\frac{\delta_0^2 + 2G_\sigma^2}{T}} \tag{6.71}$$

where $H = 8G_\sigma^2/\delta_0^2 + 2\log(1 + 2G_\sigma^2 T/\delta_0^2)$.

Guarantees of AdaMir (Stochastic)

Moreover, if (RS) kicks in, we have the sharper guarantee:

Theorem 6.9. *With notation as above, if f satisfies (RS), it enjoys the bound*

$$\mathbb{E}[f(\bar{X}_T) - f(x^*)] \leq (2 + D_1 + H)\left[\frac{A}{T} + \frac{B\sigma}{\sqrt{T}}\right] \tag{6.72}$$

where:

$$a)\quad A = \delta_0 + 2[f(X_1) - \min f] + \beta\Big(2 + 8G_\sigma^2/\delta_0^2 + 2\log(4\beta^2/\delta_0^2)\Big). \tag{6.73a}$$

$$b)\quad B = \sqrt{(4 + 2H)/K}. \tag{6.73b}$$

The proof of Theorems 6.8 and 6.9 hinges on the following key steps:

Step 1: We first show that, under (RC), the method's residuals are bounded as $\delta_t^2 \leq 2G_\sigma^2$ (a.s.).

Step 2: With this at hand, the workhorse for our analysis is the following boxing bound for the mean "weighted" regret $\sum_{t=1}^T \mathbb{E}[\gamma_t\langle\nabla f(X_t), X_t - x^*\rangle]$:

$$\mathbb{E}\left[\gamma_T\sum_{t=1}^T[f(X_t) - f(x^*)]\right] \leq \mathbb{E}\left[\sum_{t=1}^T \gamma_t\langle\nabla f(X_t), X_t - x^*\rangle\right] \leq D_1 + \mathbb{E}\left[\sum_{t=1}^T \gamma_t^2\delta_t^2\right]$$

We prove this bound in the supplement, where we also show that $\mathbb{E}[\sum_{t=1}^{T} \gamma_t^2 \delta_t^2] = \mathcal{O}(\log T)$.

At this point the analysis between Theorems 6.8 and 6.9 branches out. First, in the case of Theorem 6.8, we show that the method's step-size is bounded from below as $\gamma_t \geq 1/\sqrt{(\delta_0^2 + 2G_\sigma^2)t}$; the guarantee (6.71) then follows by the boxing bound. Instead, in the case of Theorem 6.9, the analysis is more involved and relies crucially on the lower bound $\gamma_t \geq 1/(A + B\sigma\sqrt{t})$. The bound (6.72) then follows by combining this lower bound for γ_t with the regret boxing bound above. Therefore, we first provide the crucial lemma of almost sure boundedness of the residual.

Lemma 6.10. *Assume that f satisfies (RC) and X_t are the AdaMir iterates run with feedback of the form (SFO). Then, the sequence of the residuals δ_t^2 is bounded with probability 1. In particular, we have:*

$$\delta_t^2 \leq \tilde{G}^2 = \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2 \quad \text{for all } t = 1, 2, \dots \text{ almost surely} \tag{6.74}$$

Almost Sure Boundedness of the Residual Under (RC)

Proof. By working in the same spirit, we get that:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t \langle V_t, X_t - X_{t+1} \rangle \tag{6.75}$$

and by recalling that:

$$V_t = \nabla f(X_t) + U_t \tag{6.76}$$

we get with probability 1:

$$\begin{aligned} D(X_t, X_{t+1}) + D(X_{t+1}, X_t) &\leq \gamma_t \left[\langle \nabla f(X_t), X_t - X_{t+1} \rangle + \langle U_t, X_t - X_{t+1} \rangle\right] \\ &\leq \gamma_t \left[G\sqrt{2D(X_{t+1}, X_t)} + \|U_t\|_* \|X_t - X_{t+1}\|\right] \end{aligned} \tag{6.77}$$

with the second inequality being obtained by (RC). Now, by invoking the strong convexity assumption of K, the (LHS) of the above becomes:

$$\begin{aligned} \gamma_t \left[G\sqrt{2D(X_{t+1}, X_t)} + \|U_t\|_* \|X_t - X_{t+1}\|\right] \leq \gamma_t [&G\sqrt{2(D(X_{t+1}, X_t) + D(X_t, X_{t+1}))} \\ &+ \|U_t\|_* \sqrt{\frac{2}{K}(D(X_{t+1}, X_t) + D(X_t, X_{t+1}))}] \end{aligned} \tag{6.78}$$

which in turn yields:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t \sqrt{D(X_{t+1}, X_t) + D(X_t, X_{t+1})} \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\|U_t\|_*\right] \tag{6.79}$$

Therefore, we get:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t^2 \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\|U_t\|_*\right]^2 \tag{6.80}$$

and by *stochastic first-order oracle* (SFO) we get with probability 1:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t^2 \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2 \tag{6.81}$$

or equivalently,

$$\delta_t^2 = \frac{D(X_t, X_{t+1}) + D(X_{t+1}, X_t)}{\gamma_t^2} \leq \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2 \tag{6.82}$$

and the result follows. □

Armed with the above we are ready to provide the detailed of Theorem 6.8 and Theorem 6.9. In particular, we have

Stochastic Analysis Under (RC)

Proof of Theorem 6.8. By the second part of Proposition 4.3, we have:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle V_t, X_t - x^* \rangle + D(X_t, X_{t+1}) \\ &\leq D(x^*, X_t) - \gamma_t \langle V_t, X_t - x^* \rangle + D(X_{t+1}, X_t) + D(X_t, X_{t+1}) \\ &\leq D(x^*, X_t) - \gamma_t \langle V_t, X_t - x^* \rangle + \gamma_t^2 \delta_t^2 \end{aligned} \tag{6.83}$$

which yields after rearranging and summing $t = 1, \ldots, T$:

$$\sum_{t=1}^{T} \gamma_t \langle V_t, X_t - x^* \rangle \leq D(x^*, X_1) + \sum_{t=1}^{T} \gamma_t^2 \delta_t^2 \tag{6.84}$$

and by recalling that $V_t = \nabla f(X_t) + U_t$ and taking expectations on both sides we get:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle\right] \leq D(x^*, X_1) + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \langle U_t, X_t - x^* \rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t^2 \delta_t^2\right] \tag{6.85}$$

First, we shall the (LHS) from below. In particular, we have by convexity:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t (f(X_t) - f(x^*))\right] \tag{6.86}$$

Moreover, by denoting $\tilde{G}^2 = \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2$ we have with probability 1:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t (f(X_t) - f(x^*) &= \sum_{t=1}^{T} \frac{1}{\sqrt{\delta_0^2 + \sum_{s=1}^{t-1} \delta_s^2}} (f(X_t) - f(x^*) \\ &\geq \sum_{t=1}^{T} \frac{1}{\sqrt{\delta_0^2 + \tilde{G}^2 t}} (f(X_t) - f(x^*)) \\ &\geq \sum_{t=1}^{T} \frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2) t}} (f(X_t) - f(x^*) \end{aligned}$$

$$\geq \frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)T}} \sum_{t=1}^{T} (f(X_t) - f(x^*) \tag{6.87}$$

with the second inequality being obtained by Lemma 6.10. Hence, we get:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle\right] \geq \frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)T}} \mathbb{E}\left[\sum_{t=1}^{T} (f(X_t) - f(x^*))\right] \tag{6.88}$$

We now turn our attention towards to the (LHS). In particular, we shall bound each term individually from above.

- For the term $\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \langle U_t, X_t - x^* \rangle\right]$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t \langle U_t, X_t - x^* \rangle\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \langle U_t, X_t - x^* \rangle\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\gamma_t \langle U_t, X_t - x^* \rangle | \mathcal{F}_t\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \, \mathbb{E}\left[\langle U_t, X_t - x^* \rangle | \mathcal{F}_t\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\gamma_t \langle \mathbb{E}[U_t | \mathcal{F}_t], X_t - x^* \rangle\right] = 0
\end{aligned} \tag{6.89}$$

  with the third and the fourth equality being obtained by the fact that γ_t and X_t are $\mathcal{F}_t-$ measurable.

- For the term $\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t^2 \delta_t^2\right]$: By applying Lemma A.5 and Lemma 6.10, we have with probability 1:

$$\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 \leq 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \sum_{t=1}^{T} \frac{\delta_t^2}{\delta_0^2}) \leq 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{K\delta_0^2} T) \tag{6.90}$$

  Therefore we get:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t^2 \delta_t^2\right] \leq 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2} T) \tag{6.91}$$

Thus, combining all the above we obtain:

$$\frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)T}} \mathbb{E}\left[\sum_{t=1}^{T} (f(X_t) - f(x^*))\right] \leq D(x^*, X_1) + 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2} T) \tag{6.92}$$

and hence,

$$\mathbb{E}\left[\sum_{t=1}^{T} (f(X_t) - f(x^*))\right] \leq \sqrt{(\delta_0^2 + \tilde{G}^2)T}\left[D(x^*, X_1) + 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2})T)\right] \tag{6.93}$$

The result follows by dividing both sides by T. □

Stachastic Analysis Under (RS)

Proof of Theorem 6.9. By Proposition 3.3, we have:

$$\begin{aligned} f(X_{t+1}) &\leq f(X_t) + \langle \nabla f(X_t), X_{t+1} - X_t \rangle + \beta D(X_{t+1}, X_t) \\ &\leq f(X_t) + \langle \nabla f(X_t), X_{t+1} - X_t \rangle + \beta \left[D(X_{t+1}, X_t) + D(X_t, X_{t+1}) \right] \\ &= f(X_t) + \langle V_t, X_{t+1} - X_t \rangle + \langle U_t, X_t - X_{t+1} \rangle + \beta \gamma_t^2 \delta_t^2 \\ &\leq f(X_t) - \frac{1}{\gamma_t} \left[D(X_{t+1}, X_t) + D(X_t, X_{t+1}) \right] + \|U_t\|_* \|X_t - X_{t+1}\| + \beta \gamma_t^2 \delta_t^2 \\ &= f(X_t) - \gamma_t \delta_t^2 + \|U_t\|_* \|X_t - X_{t+1}\| + \beta \gamma_t^2 \delta_t^2 \end{aligned} \tag{6.94}$$

Now, since h is $K-$ strongly convex we have that:

$$\|X_t - X_{t+1}\| \leq \sqrt{\frac{2}{K} \left[D(X_{t+1}, X_t) + D(X_t, X_{t+1}) \right]} = \sqrt{\frac{2}{K}} \gamma_t \delta_t \tag{6.95}$$

and using the fact that the noise $\|U_t\|_* \leq \sigma$ almost surely, we have:

$$f(X_{t+1}) \leq f(X_t) - \gamma_t \delta_t^2 + \sqrt{\frac{2}{K}} \gamma_t \delta_t^2 + \beta \gamma_t^2 \delta_t^2 \tag{6.96}$$

Therefore, after rearranging and telescoping we get:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq 2 \left[f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{T} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2}) + \sigma \sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t \right] \tag{6.97}$$

Now, let us bound each term of the (RHS) of the above individually:

- For the term $\sum_{t=1}^{T} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2})$ we first set:

$$t_0 = \max\{1 \leq t \leq T : \gamma_t \geq \frac{1}{2\beta}\} \tag{6.98}$$

  Then, by decomposing the said sum we get:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2}) &= \sum_{t=1}^{t_0} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2}) + \sum_{t=t_0+1}^{T} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2}) \\ &\leq \sum_{t=1}^{t_0} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2}) \\ &\leq \beta \sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 \end{aligned} \tag{6.99}$$

  with the second inequality being obtained by the definition of t_0. Now, due to the fact that $\delta_t^2 \leq \tilde{G}^2$ almost surely (by invoking Lemma 6.10) we have:

$$\beta \sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 = \beta \sum_{t=1}^{t_0} \frac{\delta_t^2}{\delta_0^2 + \sum_{s=1}^{t-1} \delta_s^2}$$

$$\leq \beta \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{1}{\delta_0^2} \sum_{t=1}^{t_0-1} \delta_t^2) \right]$$
$$\leq \beta \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{1}{\delta_0^2}(\delta_0^2 + \sum_{t=1}^{t_0-1} \delta_t^2) \right]$$
$$\leq \beta \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{1}{\delta_0^2 \gamma_{t_0}^2} \right] \tag{6.100}$$

Therefore, by the definition of t_0 we finally get with probability 1:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 (\beta \gamma_t - \frac{1}{2}) \leq \beta \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{4\beta^2}{\delta_0^2} \right] \tag{6.101}$$

- For the term $\sigma\sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t$ we have:

$$\sigma\sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t = \sigma\sqrt{\frac{2}{K}} \sum_{t=1}^{T} \sqrt{\gamma_t^2 \delta_t^2} \leq \sigma\sqrt{\frac{2}{K}} \sqrt{T} \sqrt{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2} \tag{6.102}$$

Therefore, by working in the same spirit as above we get:

$$\sigma\sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t \leq \sigma\sqrt{\frac{2}{K}} \sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{1}{\delta_0^2} \sum_{t=1}^{T} \delta_t^2)}$$
$$\leq \sigma\sqrt{\frac{2}{K}} \sqrt{T} \sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2} T)} \tag{6.103}$$

On the other hand, we may the (LHS) from below as follows:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \geq \gamma_T \sum_{t=1}^{T} \delta_t^2 \geq \gamma_T \left[\delta_0^2 - \delta_0^2 + \sum_{t=1}^{T} \delta_t^2 \right] = \frac{\gamma_T}{\gamma_{T+1}^2} - \delta_0^2 \gamma_T = \frac{1}{\gamma_T} - \delta_0^2 \gamma_T \tag{6.104}$$

So, combining the above:

$$\frac{1}{\gamma_T} - \delta_0^2 \gamma_T \leq 2(f(X_1) - \min_{x \in \mathcal{X}} f(x) + \beta \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{4\beta^2}{\delta_0^2} \right]$$
$$+ \sigma\sqrt{\frac{2}{K}} \sqrt{T} \sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2} T)}) \tag{6.105}$$

which finally yields with probability 1:

$$\frac{1}{\gamma_T} \leq \delta_0 + 2(f(X_1) - \min_{x \in \mathcal{X}} f(x) + \beta \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{4\beta^2}{\delta_0^2} \right]$$
$$+ \sigma\sqrt{\frac{2}{K}} \sqrt{T} \sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2} T)}) \tag{6.106}$$

and hence with probability 1:

$$\gamma_T \geq \left[\delta_0 + 2(f(X_1) - \min_{x\in\mathcal{X}} f(x) + \beta\left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log\frac{4\beta^2}{\delta_0^2}\right] + \sigma\sqrt{\frac{2}{K}}\sqrt{T}\sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2}T))}\right]^{-1}$$

□

Therefore, by setting:

$$A = \delta_0 + 2(f(X_1) - \min_{x\in\mathcal{X}} f(x) + \beta\left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log\frac{4\beta^2}{\delta_0^2}\right] \tag{6.107}$$

and

$$B = \sigma\sqrt{\frac{2}{K}}\sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2}T))} \tag{6.108}$$

we get that:

$$\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\gamma_T\right] \geq \left(A + B\sqrt{T}\right)^{-1}\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\right] \tag{6.109}$$

Moreover, working in the same spirit as in Theorem 6.8 we have:

$$\left(A + B\sqrt{T}\right)^{-1}\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\gamma_T\right] \leq \left(D_1 + \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right]\right) \tag{6.110}$$

which in turn yields:

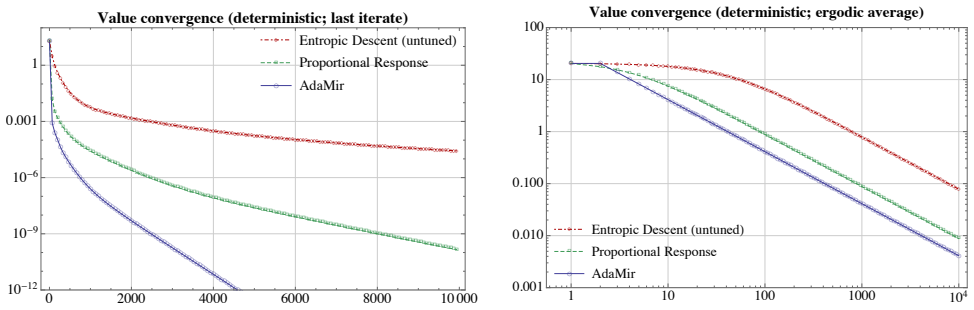
$$\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\right] \leq \left(D_1 + \mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right]\right)\left(A + B\sqrt{T}\right) \tag{6.111}$$

The result then follows by dividing both sides by T and by the fact that $\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right] = \mathcal{O}(\log T)$.

6.4 FISHER MARKETS: A CASE STUDY

6.4.1 *The Fisher market model*

We now proceed to illustrate the convergence properties of AdaMir in a Fisher equilibrium problem with linear utilities – both stochastic and deterministic. Following [94], a Fisher market consists of a set $\mathcal{N} = \{1, \dots, N\}$ of N *buyers* – or *players* – that seek to share a set $\mathcal{A} = \{1, \dots, n\}$ of n perfectly divisible goods (ad space, CPU/GPU runtime, bandwidth, etc.). The allocation mechanism for these goods follows a proportionally fair price-setting rule that is sometimes referred to as a *Kelly auction* [61]: each player $i = 1, \dots, N$ bids x_{ia} per unit of the a-th good, up the player's individual budget; for the sake of simplicity, we assume that this budget is equal to 1 for all players, so $\sum_{a=1}^{n} x_{ia} \leq 1$ for all $i = 1, \dots, N$. The price of the a-th good is then set to be the sum of the players' bids, i.e., $p_a = \sum_{i\in\mathcal{N}} x_{ia}$; then, each player gets a prorated fraction of each good, namely $w_{ia} = x_{ia}/p_a$.



(a) Last-iterate convergence **(b)** Ergodic convergence

Figure 6.1: The convergence speed of (EGD), (PR) and AdaMir in a stationary Fisher market.

Now, if the marginal utility of the i-th player per unit of the a-th good is θ_{ia}, the agent's total utility will be

$$u_i(x_i; x_{-i}) = \sum_{a \in \mathcal{A}} \theta_{ia} w_{ia} = \sum_{a \in \mathcal{A}} \frac{\theta_{ia} x_{ia}}{\sum_{j \in \mathcal{N}} x_{ja}}, \tag{6.112}$$

where $x_i = (x_{ia})_{a \in \mathcal{A}}$ denotes the bid profile of the i-th player, and we use the shorthand $(x_i; x_{-i}) = (x_1, \dots, x_i, \dots, x_N)$. A *Fisher equilibrium* is then reached when the players' prices bids follow a profile $x^* = (x_1^*, \dots, x_N^*)$ such that

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \tag{Eq}$$

for all $i \in \mathcal{N}$ and all $x_i = (x_{ia})_{a \in \mathcal{A}}$ such that $x_{ia} \geq 0$ and $\sum_{a \in \mathcal{A}} x_{ia} = 1$.[3]

As was observed by Shmyrev [107], the equilibrium problem (Eq) can be rewritten equivalently as

$$\begin{aligned} \text{minimize} \quad & F(x; \theta) \equiv \sum_{a \in \mathcal{A}} p_a \log p_a - \sum_{i \in \mathcal{N}} \sum_{a \in \mathcal{A}} x_{ia} \log \theta_{ia} \\ \text{subject to} \quad & p_a = \sum_{i \in \mathcal{N}} x_{ia}, \ \sum_{a \in \mathcal{A}} x_{ia} = 1, \text{ and } x_{ia} \geq 0 \text{ for all } a \in \mathcal{A}, i \in \mathcal{N}, \end{aligned} \tag{Opt}$$

with the standard continuity convention $0 \log 0 = 0$. In the above, the agents' marginal utilities are implicitly assumed fixed throughout the duration of the game. On the other hand, if these utilities fluctuate stochastically over time, the corresponding reformulation instead involves the *mean* objective

$$f(x) = \mathbb{E}[F(x; \omega)]. \tag{6.113}$$

Because of the logarithmic terms involved, F (and, a fortiori, f) cannot be Lipschitz continuous or smooth in the standard sense. However, as was shown by Birnbaum et al. [23], the problem satisfies (RS) over $\mathcal{X} = \{x \in \mathbb{R}_+^{Nn} : \sum_{a \in \mathcal{A}} x_{ia} = 1\}$ relative to the negative entropy function $h(x) = \sum_{ia} x_{ia} \log x_{ia}$. As a result, mirror descent methods based on this Bregman function are natural candidates for solving (6.113).

3 It is trivial to see that, in this market problem, all users would saturate their budget constraints at equilibrium, i.e., $\sum_{a \in \mathcal{A}} x_{ia} = 1$ for all $i \in \mathcal{N}$.

In more detail, following standard arguments [20], the general mirror descent template (MD) relative to h can be written as

$$[^{+}x_{ia}] = \frac{x_{ia}\exp(-\gamma g_{ia})}{\sum_{a'\in\mathcal{A}} x_{ia'}\exp(-\gamma g_{ia'})} \tag{6.114}$$

where the (stochastic) gradient vector $g \equiv g(x;\theta)$ is given in components by

$$g_{ia} = 1 + \log p_a - \log \theta_{ia}. \tag{6.115}$$

Explicitly, this leads to the entropic gradient descent algorithm

$$X_{ia,t+1} = \frac{X_{ia,t}(\theta_{ia}/p_a)^{\gamma_t}}{\sum_{a'\in\mathcal{A}} X_{ia',t}(\theta_{ia'}/p_{a'})^{\gamma_t}} \tag{EGD}$$

In particular, as a special case, the choice $\gamma = 1$ gives the *proportional response* (PR) algorithm of Wu and Zhang [114], namely

$$X_{ia,t+1} = \frac{\theta_{ia} w_{ia,t}}{\sum_{a'\in\mathcal{A}} \theta_{ia'} w_{ia',t}}, \tag{PR}$$

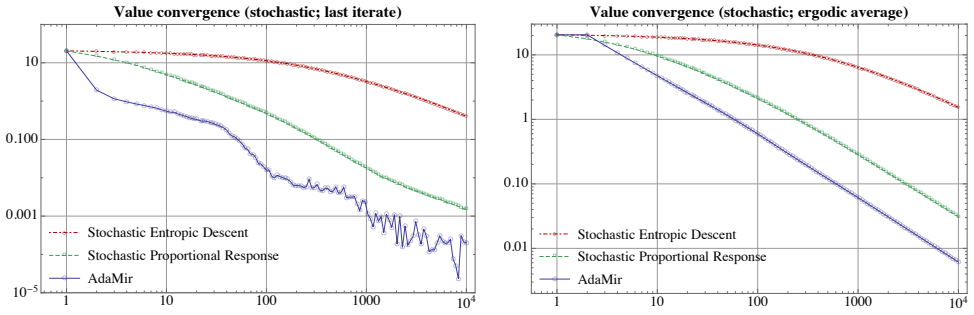
where $w_{ia,t} = X_{ia,t}/\sum_{j\in\mathcal{N}} X_{ja,t}$. As far as we aware, the PR algorithm is considered to be the most efficient method for solving *deterministic* Fisher equilibrium problems [23].

6.4.2 *Experimental validation and methodology*

For validation purposes, we ran a series of numerical experiments on a synthetic Fisher market model with $N = 50$ players sharing $n = 5$ goods, and utilities drawn uniformly at random from the interval $[2, 8]$. For stationary markets, the players' marginal utilities were drawn at the outset of the game and were kept fixed throughout; for stochastic models, the parameters were redrawn at each stage around the mean value of the stationary model (for consistency of comparisons). All experiments were run on a MacBook Pro with a 6-Core Intel i7 CPU clocking in at 2.6GHZ and 16 GB of DDR4 RAM at 2667 MHz. The Mathematica notebook used to generate the raw data and run the algorithms is included as part of the supplement (but not the entire sequence of random seed used in the stochastic case, as this would exceed the OpenReview upload limit).

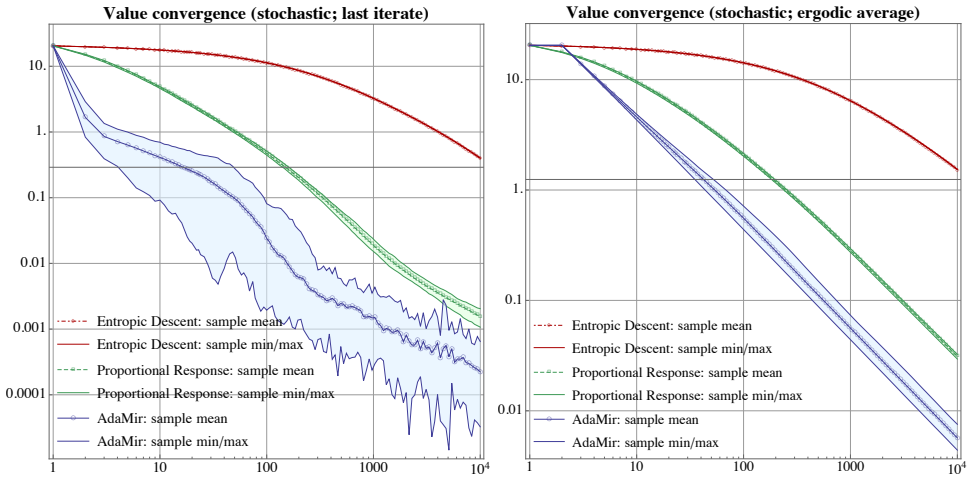
In each regime, we tested three algorithms, all initialized at the barycenter of $\mathcal{X}$: *a*) an untuned version of (EGD); *b*) the proportional response algorithm (PR); and *c*) AdaMir. For stationary markets, we ran the untuned version of (EGD) with a step-size of $\gamma = .1$; (PR) was ran "as is", and AdaMir was run with δ_0 determined by drawing a second initial condition from $\mathcal{X}$. In the stochastic case, following the theory of Lu [71] and Antonakopoulos et al. [7], the updates of (EGD) and (PR) were modulated by a $\sqrt{t}$ factor to maintain convergence; by contrast, AdaMir was run unchanged to test its adaptivity properties.

The results are reported in Figs. 6.1–6.3. For completeness, we plot the evolution of each method in terms of values of f, both for the "last iterate" X_t and the "ergodic average" $\bar{X}_t$. The results for the deterministic case are presented in Fig. 6.1. For stochastic market models, we present a sample realization in Fig. 6.2, and a statistical study over $S = 50$ sample realizations in Fig. 6.3. In all cases, AdaMir



(a) Last-iterate convergence **(b)** Ergodic convergence

Figure 6.2: The convergence speed of (EGD), (PR) and AdaMir in a stochastic Fisher market, with marginal utilities drawn i.i.d. at each epoch.



(a) Last-iterate convergence **(b)** Ergodic convergence

Figure 6.3: Statistics for the convergence speed of (EGD), (PR) and AdaMir in a stochastic Fisher market, with marginal utilities drawn i.i.d. at each epoch. The marked lines are the observed means from $S = 50$ realizations, whereas the shaded areas represent a 95% confidence interval.

outperforms both (EGD) and (PR), in terms of both last-iterate and time-average guarantees.

An interesting observation is that each method's last iterate exhibits faster convergence than its time-average, and the convergence speed of the methods' time-averaged trajectories is faster than our worst-case predictions. This is due to the specific properties of the Fisher market model under consideration: more often than not, players tend to allocate all of their budget to a single good, so almost all of the problem's inequality constraints are saturated at equilibrium. Geometrically, this means that the problem's solution lies in a low-dimensional face of $\mathcal{X}$, which is identified at a very fast rate, hence the observed accelerated rate of convergence. However, this is a specificity of the market model under consideration and should not be extrapolated to other convex problems – or other market equilibrium models to boot.

7

VARIATIONAL INEQUALITIES BEYOND LIPSCHITZ CONTINUITY

 # This section incorporates material from the papers [6, 8]

In this chapter we proceed by illustrating our contributions concerning the NoLips (VI) framework; in particular in what follows we consider operators that satisfy either (MB) or (MS). In a nutshell, Chapter 7 collects three types of results:

1. Establish methods that achieve optimal rates given that the optimizer has a prior knowledge on the regularity class of the associated operator.
2. We proceed by deriving adaptive methods relative to the smoothness modulus of the respective operator.
3. Finally, we provide our fully adaptive method for non-smooth/smooth and stochastic cases. More precisely, we provide a method which does not require any prior knowledge of Lipschitz conditions and/or the type of the oracle's feedback (deterministic or stochastic).

In what follows, we illustrate the above contributions in detail.

In Section 7.1 we start by presenting the non adaptive case for operators that transcend the typical Lipschitz regularity conditions. Having this in hand we show that the traditional optimal rates are recovered also for (VI) associated with operators with possible singularities.

Our next step is to explore various adaptivity aspects. In doing so, we start *given* that the associated operator satisfies the respective smoothness like condition. To that end, in Section 7.2 we derive an adaptive mirror-prox algorithm which attains the optimal $\mathcal{O}(1/T)$ rate of convergence in problems with possibly singular operators, without any prior knowledge of the degree of smoothness (the Bregman analogue of the Lipschitz constant).

Subsequently, in Section 7.3 we introduce a novel adaptive step-size policy with mirror prox as the underlying template. The combination of these ingredients will allow us to automatically exploit the geometry of the gradient data observed at earlier iterations to perform more informative extra-gradient steps in later ones. Thanks to this adaptation mechanism, the proposed method automatically detects whether the problem is smooth or not, without requiring any prior tuning by the optimizer. As a result, the algorithm simultaneously achieves order-optimal convergence rates, i.e., it converges to an $\mathcal{O}(1/T)$ rate for smooth problems, and $\mathcal{O}(1/\sqrt{T})$ for non-smooth ones. Importantly, these guarantees do not require any

	EG [63]	GRAAL [75]	GMP [109]	AMP [6]	BL [14]	ADAPROX [8]
PARAM. AGNOSTIC	×	✓	PARTIAL	✓	PARTIAL	✓
UNIVERSALITY	×	×	✓	×	✓	✓
UNBOUNDED	×	✓	×	×	×	✓
SINGULARITIES	×	×	×	✓	×	✓

Table 7.1: Overview of related adaptive methods for solving variational inequalities. For the purposes of this table, "parameter-agnostic" means that the method does not require prior knowledge of the parameters of the problem it was designed to solve (Lipschitz modulus, domain diameter, etc.); "rate interpolation" means that the algorithm's convergence rate is $\mathcal{O}(1/T)$ or $\mathcal{O}(1/\sqrt{T})$ in smooth / non-smooth problems respectively; "unbounded domain" is self-explanatory; and, finally, "singularities" means that the problem's defining vector field may blow up at a boundary point of the problem's domain.

of the standard boundedness or Lipschitz continuity conditions that are typically assumed in the literature.

That said, the set of results presented in Section 7.3 requires perfect oracle feedback. On that account, in Section 7.4 we present the full potency of our results by being able to treat also stochastic settings. In particular, we employ an adaptive learning rate combined with the Dual Extrapolation algorithmic template. This combination allow us to achieve optimal convergence rates for both deterministic and stochastic settings without any prior knowledge over the boundedness, smoothness and/or the level of noise.

7.1 NON ADAPTIVE CASE

We first start by presenting a family of *non-adaptive* methods and their respective rates for NoLips operators. To that end we will make some preliminary assumptions. In particular throughout this section, we assume that the following blanket assumptions hold:

Blanket Assumptions

Assumption 7.1. The solution set $\mathcal{X}^* \equiv \mathrm{Sol}(\mathcal{X}, A)$ of (VI) is nonempty.

Assumption 7.2. A is monotone and β-Bregman continuous, i.e.,

$$\|A(x) - A(x')\|_{x,*} \leq \beta\sqrt{2D(x,x')} \quad \text{for all} \quad x, x' \in \mathcal{X} \tag{7.1}$$

with D being the Bregman divergence defined in Definition 3.3.

In addition to the above, in terms of the oracle's feedback structure we assume that the optimizer gains access to an (SFO) mechanism where $U_t \in \mathcal{V}^*$ is an additive noise variable. The two cases of interest that we consider here are (*i*) when $U_t = 0$ for all t; and (*ii*) when U_t satisfies the statistical hypotheses:

Statistical Assumptions

a) *Zero-mean:* $\mathbb{E}[U_t \,|\, \mathcal{F}_t] = 0.$ (7.2a)

b) *Finite variance:* $\mathbb{E}[\|U_t\|_*^2 \,|\, \mathcal{F}_t] \leq \sigma^2.$ (7.2b)

with $\mathcal{F}_t$ denoting the history (natural filtration) of X_t. Finally, we assume for the moment that the smoothness parameter β is known a priori.

Having all this in hand, we can now extend the (optimal) standard convergence rates of the Euclidean setting for the general class of (7.1). Formally, this is stated by the following result.

NoLips Guarantees

Theorem 7.1 (Antonakopoulos et al. [6]). *Assume that A satisfies Assumptions 7.1 and 7.2, and let Gap_H denote the restricted gap function for the Bregman zone $\mathcal{C}_H = \{x \in \mathcal{X} : D(x, x_c) \leq H\}$. Suppose further that (MP) is run with an α-strongly convex Bregman function and oracle feedback of the form (SFO). Then, for all $H > 0$, the averaged sequence $\bar{X}_T = \sum_{t=1}^T \gamma_t X_{t+1/2} / \sum_{t=1}^T \gamma_t$ enjoys the following gap bounds:*

a) If $\sigma^2 = 0$ and the algorithm's step-size satisfies

$$0 < \gamma_{\min} \equiv \inf_t \gamma_t \leq \sup_t \gamma_t \equiv \gamma_{\max} \leq \sqrt{\alpha}/\beta, \tag{7.3}$$

we have

$$\mathrm{Gap}_H(\bar{X}_T) \leq \frac{H}{\gamma_{\min}} \frac{1}{T} \tag{7.4}$$

b) Otherwise, if $\sigma^2 > 0$ and $\gamma_t \leq \sqrt{\alpha/2}/\beta$, we have

$$\mathbb{E}[\mathrm{Gap}_H(\bar{X}_T)] = \mathcal{O}\left(\frac{H + \sigma^2 \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}\right) \tag{7.5}$$

In particular, if $\gamma_t \propto 1/\sqrt{T}$, we get $\mathbb{E}[\mathrm{Gap}_H(\bar{X}_t)] = \mathcal{O}(1/\sqrt{T})$.

For convenience we divide the proof of Theorem 7.1 into the deterministic and stochastic part. In doing so, the main ingredient of the proof of the deterministic case is the following energy inequality:

Method's Template Inequality

$$D(p, X_{t+1}) \leq D(p, X_t) - \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle - \left(1 - \frac{\beta^2 \gamma_t^2}{\alpha}\right) D(X_{t+\frac{1}{2}}, X_t).$$

(7.6) is obtained directly by the connection established by Proposition 4.4 between two prox-steps combined with the (7.1). Formally, we show the following result.

Proposition 7.2 (Antonakopoulos et al. [6]). *Assume that A satisfies Assumption 7.2 and (MP) is run with perfect oracle feedback. Then, for all $p \in \mathcal{X}$, we have:*

$$D(p, X_{t+1}) \leq D(p, X_t) - \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle - \left(1 - \frac{\beta^2 \gamma_t^2}{\alpha}\right) D(X_{t+\frac{1}{2}}, X_t).$$

Proof. By setting $x = X_t$, $y_1 = -\gamma_t A(X_t)$, $x_1^+ = X_{t+\frac{1}{2}}$, $y_2 = -\gamma_t A(X_{t+\frac{1}{2}})$ and $x_2^+ = X_{t+1}$ in Proposition 4.4, we readily obtain:

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &- \gamma_t \langle A(X_{t+\frac{1}{2}}) - A(X_t), X_{t+1} - X_{t+\frac{1}{2}} \rangle \\ &- D(X_{t+1}, X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}}, X_t). \end{aligned} \tag{7.6}$$

Proceeding line-by-line, the Fenchel-Young inequality applied to the function $\phi(x) = \|x\|^2_{X_{t+\frac{1}{2}}}$ further gives

$$\langle A(X_{t+\frac{1}{2}}) - A(X_t), X_{t+1} - X_{t+\frac{1}{2}} \rangle \leq \frac{\alpha}{2\gamma_t}\|X_{t+1} - X_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}}} + \frac{\gamma_t}{2\alpha}\|A(X_{t+\frac{1}{2}}) - A(X_t)\|_{X_{t+\frac{1}{2}},*}. \tag{7.7}$$

Thus, by substituting in (7.6), we get

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &+ \frac{\alpha}{2}\|X_{t+1} - X_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}}} + \frac{\gamma_t^2}{2\alpha}\|A(X_{t+\frac{1}{2}}) - A(X_t)\|^2_{X_{t+\frac{1}{2}},*} \\ &- D(X_{t+1}, X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}}, X_t). \end{aligned} \tag{7.8}$$

and hence, by Lemma 3.2, we obtain:

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &+ \frac{\gamma_t^2}{2\alpha}\|A(X_{t+\frac{1}{2}}) - A(X_t)\|^2_{X_{t+\frac{1}{2}},*} - D(X_{t+\frac{1}{2}}, X_t). \end{aligned} \tag{7.9}$$

However, the Bregman continuity of A also yields

$$\|A(X_{t+\frac{1}{2}}) - A(X_t)\|^2_{X_{t+\frac{1}{2}},*} \leq 2\beta^2 D(X_{t+\frac{1}{2}}, X_t) \tag{7.10}$$

so our claim follows by combining Eqs. (7.9) and (7.10). □

Having established the template inequality in Proposition 7.2 we are now in a position to illustrate the proof for the $\mathcal{O}(1/T)$ convergence rate of the restricted merit function (2.10) for deterministic problems. Formally, we have:

Proof of Theorem 7.1 - deterministic case. Fix some $p \in \mathcal{C}_H$. Since $\gamma_t \leq 1/\beta$ by assumption, a slight rearrangement of Proposition 7.2 readily yields:

$$\gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_t) - D(p, X_{t+1}) \tag{7.11}$$

Moreover, by the monotonicity of A, we also have:

$$\langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle. \tag{7.12}$$

Thus, combining the two inequalities above, we get

$$\gamma_t \langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_t) - D(p, X_{t+1}) \tag{7.13}$$

and, proceeding to telescope from $t = 1$ to T, we obtain:

$$\sum_{t=1}^{T} \gamma_t \langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_1) - D(p, X_{t+1}) \leq D(p, x_c) \tag{7.14}$$

Then, dividing by $\sum_{t=1}^{T}\gamma_t$ finally yields

$$\langle A(p), \bar{X}_T - p\rangle \leq \frac{D(p,x_c)}{\sum_{t=1}^{T}\gamma_t} \leq \frac{D(p,x_c)}{\gamma_{\min} T}, \tag{7.15}$$

so our result follows by taking the supremum over all $p \in \mathcal{X}$ such that $D(p,x_c) \leq H$ (i.e., over all $p \in \mathcal{C}_H$). □

We now turn our attention towards the stochastic part of Theorem 7.1. In a nutshell, we emphasize that the main building block for deriving this result is the inequality obtained in Proposition 4.4. More precisely, we have the following:

Proof of Theorem 7.1 - stochastic case. Working in the same spirit as for the deterministic case, let $x = X_t$, $y_1 = -\gamma_t V_t$, $x_1^+ = X_{t+\frac{1}{2}}$, $y_2 = -\gamma_t V_{t+\frac{1}{2}}$ and $x_2^+ = X_{t+1}$ in the first part of Proposition 4.4. We then get:

$$\begin{aligned} D(p,X_{t+1}) &\leq D(p,X_t) - \gamma_t\langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - p\rangle \\ &\quad + \left[\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t)\right] \\ &\leq D(p,X_t) - \gamma_t\langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle \\ &\quad - \gamma_t\xi_{t+\frac{1}{2}} + \left[\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t)\right] \end{aligned} \tag{7.16}$$

where we used the feedback decomposition $V_{t+\frac{1}{2}} = A(X_{t+\frac{1}{2}}) + U_{t+\frac{1}{2}}$ for $V_{t+\frac{1}{2}}$ and we set $\xi_{t+\frac{1}{2}} = \langle U_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - p\rangle$ in the last line. By the second part of Proposition 4.4, we also have

$$\begin{aligned} \gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t) &\leq \gamma_t\langle V_t - V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle \\ &\quad - D(X_{t+1},X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}},X_t) \end{aligned} \tag{7.17}$$

Now, by applying the Fenchel-Young inequality to the duality pairing in the above inequality, we get

$$\gamma_t\langle V_t - V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle \leq \frac{\gamma_t^2}{2\alpha}\|V_t - V_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}},*} + \frac{\alpha}{2}\|X_{t+1} - X_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}}}. \tag{7.18}$$

On the other hand, by the stochastic oracle assumption (SFO), we have:

$$\begin{aligned} \frac{\gamma_t^2}{2\alpha}\|V_t - V_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}},*} &\leq \frac{\gamma_t^2}{\alpha}\|A(X_t) - A(X_{t+\frac{1}{2}})\|^2_{X_{t+\frac{1}{2}},*} + \frac{\gamma_t^2}{\alpha}\|U_t - U_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}},*} \\ &\leq \frac{2\beta^2\gamma_t^2}{\alpha}D(X_{t+\frac{1}{2}},X_t) + \frac{\gamma_t^2}{\mu\alpha}\|U_t - U_{t+\frac{1}{2}}\|^2_*. \end{aligned} \tag{7.19}$$

where the last line follows from the Bregman continuity of A (Assumption 7.2) and the fact that $\|\cdot\|_x \geq \mu\|\cdot\|$ for some $\mu > 0$ and all $x \in \mathcal{X}$ (implying in turn that $\|\cdot\|_{x,*} \leq \mu^{-1}\|\cdot\|_*$ for all $x \in \mathcal{X}$). We thus get:

$$\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t) \leq \left(\frac{2\beta^2\gamma_t^2}{\alpha} - 1\right)D(X_{t+\frac{1}{2}},X_t) + \frac{\gamma_t^2}{\mu\alpha}\|U_t - U_{t+\frac{1}{2}}\|^2_* \tag{7.20}$$

Since $\gamma_t^2 \leq \alpha/(2\beta^2)$ by assumption, substituting (7.20) in (7.16) and rearranging yields

$$\gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_t) - D(p, X_{t+1}) - \gamma_t \xi_{t+\frac{1}{2}} + \frac{\gamma_t^2}{\mu\alpha} \|U_t - U_{t+\frac{1}{2}}\|_*^2 \tag{7.21}$$

which in turn yields:

$$\gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_t) - D(p, X_{t+1})) - \gamma_t \xi_{t+\frac{1}{2}} + \frac{2\gamma_t^2}{\mu\alpha} \left[\|U_t\|_*^2 + \|U_{t+\frac{1}{2}}\|_*^2 \right]. \tag{7.22}$$

In order to bound $\xi_{t+\frac{1}{2}}$, we will need to introduce the auxilliary process

$$Z_{t+1} = \arg\min_{x \in \mathcal{X}} \{ \langle U_{t+\frac{1}{2}}, Z_t - x \rangle + \frac{\mu}{\gamma_t} D(x, Z_t) \} \tag{7.23}$$

with $Z_1 = x_c$. We then have

$$-\gamma_t \xi_{t+1} = \gamma_t \langle U_{t+\frac{1}{2}}, p - X_{t+\frac{1}{2}} \rangle = \gamma_t \langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}} \rangle + \gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle \tag{7.24}$$

In order to bound the term which depends on p, we have the following:

$$\begin{aligned} \gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle &= \gamma_t \langle U_{t+\frac{1}{2}}, p - Z_{t+1} \rangle + \gamma_t \langle U_{t+\frac{1}{2}}, Z_{t+1} - Z_t \rangle \\ &\leq \mu \langle \nabla h(Z_{t+1}) - \nabla h(Z_t), p - Z_{t+1} \rangle \\ &\quad + \frac{\gamma_t^2}{2\alpha} \|U_{t+\frac{1}{2}}\|_{*,Z_t}^2 + \frac{\alpha}{2} \|Z_{t+1} - Z_t\|_{Z_t}^2 \end{aligned} \tag{7.25}$$

and so,

$$\gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle \leq \mu \langle \nabla h(Z_{t+1}) - \nabla h(Z_t), p - Z_{t+1} \rangle + \frac{\gamma_t^2}{2\mu\alpha} \|U_{t+\frac{1}{2}}\|_*^2 + \frac{\alpha\mu}{2} \|Z_{t+1} - Z_t\|^2. \tag{7.26}$$

Hence, by the three-point identity, we obtain:

$$\begin{aligned} \gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle &\leq \mu [D(p, Z_t) - D(p, Z_{t+1})] - \mu D(Z_{t+1}, Z_t) \\ &\quad + \frac{\gamma_t^2}{2\mu\alpha} \|U_{t+\frac{1}{2}}\|_*^2 + \frac{\alpha\mu}{2} \|Z_{t+1} - Z_t\|^2 \\ &\leq \mu [D(p, Z_t) - D(p, Z_{t+1})] + \frac{\gamma_t^2}{2\mu\alpha} \|U_{t+\frac{1}{2}}\|_*^2 \end{aligned} \tag{7.27}$$

where the last inequality is a consequence of the strong convexity of h. Thus, combining all these with the fact that A is monotone, we can telescope and obtain

$$\sum_{t=1}^{T} \gamma_t \langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq (1 + \mu) D(p, x_c)$$

$$+\sum_{t=1}^{T}\gamma_t\langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}}\rangle$$
$$+\frac{1}{\mu\alpha}\sum_{t=1}^{T}\gamma_t^2\left[2\|U_t\|_*^2+\frac{5}{2}\|U_{t+\frac{1}{2}}\|_*^2\right].$$

Hence, after dividing by $\sum_{t=1}^{T}\gamma_t$ and taking the supremum over $p\in\mathcal{C}_H$, by setting $\lambda_t = \langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}}\rangle$ we get:

$$\mathrm{Gap}_H(\bar{X}_T)\leq\frac{(1+\mu)H+\sum_{t=1}^{T}\gamma_t\lambda_t+\frac{1}{\mu\alpha}\sum_{t=1}^{T}\gamma_t^2\left[2\|U_t\|_*^2+\frac{5}{2}\|U_{t+\frac{1}{2}}\|_*^2\right]}{\sum_{t=1}^{T}\gamma_t}. \tag{7.28}$$

Since $\mathbb{E}[\langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}}\rangle] = \mathbb{E}[\mathbb{E}[\langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}}\rangle]\,|\,\mathcal{F}_{t+\frac{1}{2}}] = 0$, taking expectations yields

$$\mathbb{E}[\mathrm{Gap}_H(\bar{X}_t)]\leq\frac{(1+\mu)D+\frac{9\sigma^2}{2\mu\alpha}\sum_{t=1}^{T}\gamma_t^2}{\sum_{t=1}^{T}\gamma_t}, \tag{7.29}$$

which proves our claim. Finally, the RHS of this last inequality is $\tilde{\mathcal{O}}(1/T^{1/2})$ if $\gamma_t\propto 1/\sqrt{t}$, so the $\tilde{\mathcal{O}}(1/\sqrt{T})$ result follows. □

Theorem 7.1 relies crucially on prior knowledge of the following key factors:

1. That the associated operator satisfies the respective Bregman smoothness regularity condition.
2. A fortiori, in order to properly tune the method's step-size policy the optimizer needs to be able to estimate the precise (Bregman) Lipschitz constant.

As a prelude of the analysis to come the following section will be focusing on relaxing different aspects of these elements.

7.2 ADAPTIVITY TO THE SMOOTHNESS MODULUS

As we already mentioned, a crucial assumption underlying the analysis of the previous section is that the optimizer must know in advance – or be otherwise able to estimate – the Bregman constant β. In practice, this can be difficult to achieve, so it is important to be able to run (MP) with an *adaptive* step-size policy. Therefore our first step towards adaptivity is to design methods for solving Bregman smooth (VI)'s where the respective "smoothness" parameter is unknown a priori.

Estimator of Smoothness Constant

Our starting point is the observation that, with perfect oracle feedback, one can estimate β by setting

$$\beta_t = \frac{\|A(X_{t+1/2}) - A(X_t)\|_{X_{t+1/2},*}}{\sqrt{2D(X_{t+1/2}, X_t)}} \tag{7.30}$$

whenever $X_{t+1/2}\neq X_t$; obviously, if A is β-Bregman continuous, we have $\beta_t\leq\beta$.[1] However, the fact that the Bregman constant is being *under*-estimated means that a

1 In a Euclidean setting, similar ideas can be found in, e.g., [24, 75]. We ignore the origins of this technique.

step-size policy of the form $\gamma_t \propto \sqrt{\alpha}/\beta_t$ would *over*-estimate the inverse Bregman constant $1/\beta$, so the resulting step-size policy would have no reason to satisfy (7.3).

To overcome this obstacle, we introduce the following comparison mechanism: first, at each $t = 1, 2, \ldots$, we use the estimation (7.30) to test the step-size $\bar{\gamma}_t = \sqrt{\alpha}/\beta_t$. Then, to avoid the growth phenomenon outlined above, we shrink $\bar{\gamma}_t$ by a constant factor of θ and, to avoid running into vanishing step-size issues, we take the previous step-size employed if the shrunk one would be smaller. Formally, we consider the adaptive step-size policy:

Step-Size Adaptive to the Lipschitz Constant

$$\gamma_{t+1} = \begin{cases} \min\{\gamma_t, \theta\sqrt{\alpha}/\beta_t\} & \text{if } X_t \neq X_{t+1/2}, \\ \gamma_t & \text{otherwise,} \end{cases} \tag{7.31}$$

Guarantees under Adaptivity to the Lipschitz Constant

with β_t defined as in (7.30) and $\theta \in (0, 1)$ chosen arbitrarily.

Theorem 7.3 (Antonakopoulos et al. [6]). *Assume that the monotone operator A satisfies Assumptions 7.1 and 7.2, and (MP) is run with perfect oracle feedback and the adaptive step-size policy (7.31). Then, with notation as in Theorem 7.1, the algorithm's ergodic average*

$$\bar{X}_T = \sum_{t=1}^{T} \gamma_t X_{t+1/2} \Big/ \sum_{t=1}^{T} \gamma_t \tag{7.32}$$

enjoys the gap bound

$$\mathrm{Gap}_H(\bar{X}_T) = \mathcal{O}(1/T). \tag{7.33}$$

Proof. We begin with an induction argument to show that the adaptive step-size policy $\gamma_{t+1} = \min\{\gamma_t, \theta\sqrt{\alpha}/\beta_t\}$ is lower bounded as

$$\gamma_t \geq \min\{\gamma_1, \theta\sqrt{\alpha}/\beta\}. \tag{7.34}$$

Indeed, assuming this bound for γ_t, we have either *a*) $\gamma_{t+1} = \gamma_t \geq \theta\sqrt{\alpha}/\beta$ by the inductive assumption; or *b*) $\gamma_{t+1} = \theta\sqrt{\alpha}/\beta_t \geq \theta\sqrt{\alpha}/\beta$ by the fact that β_t is an under-estimate of β. Thus, with β_t (weakly) decreasing, it follows that γ_t converges to some well-defined limit value $\gamma_\infty \geq \theta\sqrt{\alpha}/\beta < \sqrt{\alpha}/\beta$.

To proceed, given that $\beta_t \leq \beta$, working in the same spirit as we did to obtain the basic energy inequality (7.6) in the previous section, we get:

$$D(p, X_{t+1}) \leq D(p, X_t) + \gamma_t \langle g_{t+1/2}, X_{t+1/2} - p\rangle - \left(1 - \theta^2 \frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+1/2}, X_t) \tag{7.35}$$

leading to the estimate

$$\begin{aligned} \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - p\rangle \leq{}& D(p, X_t) - D(p, X_{t+1}) \\ &- \left(1 - \theta^2 \frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+1/2}, X_t). \end{aligned} \tag{7.36}$$

Since γ_t converges, it follows that $\lim_{t\to\infty} \gamma_t^2/\gamma_{t+1}^2 = 1$, so we get

$$\lim_{t\to\infty}\left(1-\theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) = 1-\theta^2 > 0, \tag{7.37}$$

implying in turn that

$$\left(1-\theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+1/2}, X_t) > 0 \tag{7.38}$$

for all t greater than some (finite) t_0. Accordingly, summing and telescoping as in the analysis of the previous section, we get

$$\begin{aligned}\sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}), X_{t+1/2}-p\rangle &\leq D(p,x_c) + \sum_{t=1}^{t_0}\left(1-\theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right)D(X_{t+1/2},X_t)\\ &< +\infty\end{aligned}$$

whenever $T > t_0$. Our result then follows by dividing both sides of this last inequality by $\sum_{t=1}^{T}\gamma_t$ and recalling the fact that $\gamma_t \geq \theta\sqrt{\alpha}/\beta > 0$ for all t. □

As we mentioned the contributions of Theorem 7.3 hinge on the fact that the optimizer knows in advance that associate operator satisfies (7.1). In what follows, we shall tackle this drawback by developing methods that are agnostic to the respective regularity condition at hand.

7.3 THE DETERMINISTIC CASE

Moving forward we define the appropriate adaptive step-size policy that will allow the (MP) method to exhibit "regime-agnostic" optimal convergence rates, i.e., adjust optimally its performance without any prior knowledge of the underlying regularity condition. Our starting point for designing such methods is by considering first the deterministic case. More precisely, throughout this section we assume the following blanket conditions:

1. $\mathcal{X}$ is a regular Finsler space (cf. Section 3.2.2).
2. Regarding the NoLips condition we will assume that the respective operator satisfies either (MB) or (MS)
3. The associated regularizer h is in the sense of Definition 3.3.

With all this is in place, the deterministic version (MP) method, defined by the following recursion:

$$\begin{aligned}X_{t+1/2} &= P_{X_t}(-\gamma_t V_t)\\ X_{t+1} &= P_{X_t}(-\gamma_t V_{t+1/2})\end{aligned} \tag{7.39}$$

can be adapted to our current setting as follows:

Universal Step-Size (Deterministic)

$$\gamma_t = \frac{1}{\sqrt{1+\sum_{j=1}^{t-1}\|V_{j+1/2}-V_j\|^2_{X_{j+1/2},*}}} \tag{Adapt}$$

with $V_t = A(X_t)$, $t = 1, 1/2, \ldots$. In words, this method builds on the template of (MP) by replacing the global norm with a dual Finsler norm evaluated at the algorithm's leading state $X_{j+1/2}$ combined with the respective adaptive step-size policy. We conclude this section by providing an intuitive explanation for the step-size (Adapt). In particular, under (MB) we get that:

Intuition Behind Universality

$$\|V_j - V_{j+1/2}\|^2_{X_{j+1/2},*} \approx \text{"constant"}$$

and hence $\gamma_t \propto 1/\sqrt{t}$. Hence, we have that:

$$\sum_{t=1}^{T} \gamma_t = \Omega(\sqrt{T}) \tag{7.40}$$

On the other hand under (MS) we show that γ_t stabilizes to some strictly positive value which in turn yields:

$$\sum_{t=1}^{T} \gamma_t = \Omega(T) \tag{7.41}$$

and thus leads to a faster convergence rate. In the forthcoming analysis, we shall explain this behaviour of the adaptive step-size in detail.

7.3.1 *Optimal rate interpolation*

Universality Guarantees (Deterministic)

With all this in hand, our main result for our method can be stated as follows:

Theorem 7.4 (Antonakopoulos et al. [8])**.** *Suppose A is a monotone operator, let $\mathcal{C}$ be a compact neighborhood of a solution of* (VI)*, and set $H = \sup_{x\in\mathcal{C}} D(x, X_1)$ Then,* (MP) *run with the adaptive step-size* (Adapt) *enjoys the guarantees:*

1. *If A satisfies* (MB)*:*

$$\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T) = \mathcal{O}\left(\frac{H + G^3(1+1/K)^2 + \log(1+4G^2(1+2/K)^2T)}{\sqrt{T}}\right). \tag{7.42}$$

2. *If A satisfies* (MS)*:*

$$\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T) = \mathcal{O}(H/T). \tag{7.43}$$

In a nutshell we mention here that its key element is the determination of the asymptotic behavior of the adaptive step-size policy γ_t in the non-smooth and smooth regimes, i.e., under (MB) and (MS) respectively. At a very high level, (MB) guarantees that the difference sequence $\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}$ is bounded, which implies in turn that $\sum_{t=1}^{T} \gamma_t = \Omega(\sqrt{T})$ and eventually yields the bound (7.42) for the algorithm's ergodic average $\bar{X}_T$. This is accomplished formally in the following lemma.

Boundness of the Residual Under (MB)

Lemma 7.5 (Antonakopoulos et al. [8])**.** *Suppose that the monotone operator A satisfies* (MB)*. Then, the sequence $\|A(X_{t+1/2}) - A(X_t)\|^2_{X_t,*}$ is bounded. In particular, the following inequality holds:*

$$\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq C^2 \tag{7.44}$$

with $C = 2G + \beta\frac{4G}{K}$.

Proof. It suffices to show that: $\|A(X_{t+1/2}) - A(X_t)\|_{X_{t+1/2},*}$ is bounded. More precisely, by the triangle inequality we have:

$$\|A(X_{t+1/2}) - A(X_t)\|_{X_{t+1/2},*} \leq \|A(X_{t+1/2})\|_{X_{t+1/2},*} + \|A(X_t)\|_{X_{t+1/2},*} \tag{7.45}$$

We shall bound the (RHS) part of (7.45) term by term. In particular, we have:

- For the first term $\|A(X_{t+1/2})\|_{X_{t+1/2},*}$ we readily get by (MB):

$$\|A(X_{t+1/2})\|_{X_{t+1/2},*} \leq G \tag{7.46}$$

- For the second term $\|A(X_t)\|_{X_{t+1/2},*}$, we have:

$$\begin{aligned}\|A(X_t)\|_{X_{t+1/2},*} &\leq \|A(X_t)\|_{X_t,*} + \beta\left[\|X_t - X_{t+1/2}\|_{X_t} + \|X_t - X_{t+1/2}\|_{X_{t+1/2}}\right] \\ &\leq G + \beta\left[\|X_t - X_{t+1/2}\|_{X_t} + \|X_t - X_{t+1/2}\|_{X_{t+1/2}}\right] \end{aligned} \tag{7.47}$$

Therefore, it suffices to show that the quantity $\|X_t - X_{t+1/2}\|_{X_t} + \|X_t - X_{t+1/2}\|_{X_{t+1/2}}$ is bounded from above. Indeed, we have:

$$\begin{aligned} D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) &= \langle \nabla h(X_t) - \nabla h(X_{t+1/2}), X_t - X_{t+1/2}\rangle \\ &\leq \gamma_t \langle A(X_t), X_t - X_{t+1/2}\rangle \\ &\leq G\gamma_t \|X_t - X_{t+1/2}\|_{X_t} \end{aligned}$$

where the last inequality is obtained by (MB). Moreover, by Definition 3.3 we get:

$$\begin{aligned} D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) &\leq \gamma_t G\sqrt{\frac{2}{K} D(X_{t+1/2}, X_t)} \\ &\leq G\sqrt{\frac{2}{K}\left[D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t)\right]} \end{aligned}$$

which yields

$$D(X_t, X_{t+1/2}) + D(X_t, X_{t+1/2}) \leq \frac{2G^2}{K} \tag{7.48}$$

Hence, by the local strong convexity in Definition 3.3 of h, we get:

$$\frac{K}{2}\left[\|X_t - X_{t+1/2}\|^2_{X_t} + \|X_t - X_{t+1/2}\|^2_{X_{t+1/2}}\right] \leq \frac{2G^2}{K} \tag{7.49}$$

which in turn implies that:

$$\|X_t - X_{t+1/2}\|_{X_t} \leq \frac{2G}{K} \text{ and } \|X_t - X_{t+1/2}\|_{X_{t+1/2}} \leq \frac{2G}{K} \tag{7.50}$$

and so,

$$\|X_t - X_{t+1/2}\|_{X_t} + \|X_t - X_{t+1/2}\|_{X_{t+1/2}} \leq \frac{4G}{K} \tag{7.51}$$

Moreover, by combining (7.47) and (7.51) we get:

$$\|A(X_t)\|_{X_{t+1/2},*} \leq G + \beta\frac{4G}{K} \tag{7.52}$$

Summarizing, (7.45) combined with (7.47) and (7.52) yields:

$$\|A(X_{t+1/2}) - A(X_t)\|_{X_{t+1/2},*} \leq 2G + \beta\frac{4G}{K} \tag{7.53}$$

and hence the result follows. □

Summability of the Residual Under (MS)

On the other hand, if (MS) kicks in, we have the following finer result:

Lemma 7.6 (Antonakopoulos et al. [8]). *Assume the monotone operator A satisfies (MS). Then,*

1. *γ_t decreases monotonically to a strictly positive limit $\gamma_\infty = \lim_{t\to\infty}\gamma_t > 0$;*
2. *The sequence $\|A(X_{t+1/2}) - A(X)\|_{X_{t+1/2},*}$ is square summable: in particular,i.e.,*

$$\sum_{t=1}^{\infty}\|A(X_{t+1/2}) - A(X)\|^2_{X_{t+1/2},*} = 1/\gamma_\infty^2 - 1. \tag{7.54}$$

Proof. Since γ_t is decreasing and bounded from below ($\gamma_t \geq 0$), then we readily obtain that its limit exists and more precisely we have:

$$\lim_{t\to+\infty}\gamma_t = \inf_{t\in\mathbb{N}}\gamma_t = \gamma_\infty \geq 0 \tag{7.55}$$

We now assume that $\gamma_\infty = 0$. Then, by recalling (7.6):

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t\langle A(X_{t+1/2}), X_{t+1/2} - p\rangle \\ &+ \gamma_t\langle A(X_{t+1/2}) - A(X_t), X_{t+1} - X_{t+1/2}\rangle \\ &- D(X_{t+1/2}, X_t) - D(X_{t+1}, X_{t+1/2}) \end{aligned} \tag{7.56}$$

By rearranging the above and telescoping $t = 1, \dots, T$ we get:

$$\begin{aligned} \sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}), X_{t+1/2} - p\rangle \leq D(p, X_1) \\ + \sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}) - A(X_t), X_{t+1} - X_{t+1/2}\rangle \\ - \sum_{t=1}^{T}D(X_{t+1/2}, X_t) - \sum_{t=1}^{T}D(X_{t+1}, X_{t+1/2}) \end{aligned} \tag{7.57}$$

whereas, by applying Fenchel-Young inequality to the above we readily get:

$$\sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}), X_{t+1/2} - p\rangle \leq D(p, X_1)$$

$$+\frac{1}{2K}\sum_{t=1}^{T}\gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}+\frac{K}{2}\sum_{t=1}^{T}\|X_{t+1}-X_{t+1/2}\|^2_{X_{t+1/2}}$$
$$-\sum_{t=1}^{T}D(X_{t+1/2},X_t)-\sum_{t=1}^{T}D(X_{t+1},X_{t+1/2}) \quad (7.58)$$

and by considering that the local-strong convexity of Definition 3.3:

$$\frac{K}{2}\sum_{t=1}^{T}\|X_{t+1}-X_{t+1/2}\|^2_{X_{t+1/2}}-\sum_{t=1}^{T}D(X_{t+1},X_{t+1/2})\leq 0 \quad (7.59)$$

we finally obtain:

$$\sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}),X_{t+1/2}-p\rangle\leq D(p,X_1)+\frac{1}{2K}\sum_{t=1}^{T}\gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}$$
$$-\sum_{t=1}^{T}D(X_{t+1/2},X_t) \quad (7.60)$$

Therefore, by the definition (MS) we have:

$$\sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}),X_{t+1/2}-p\rangle\leq D(p,X_1)+\frac{1}{2K}\sum_{t=1}^{T}\gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}$$
$$-\frac{K}{2\beta^2}\sum_{t=1}^{T}\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*} \quad (7.61)$$

which becomes:

$$\sum_{t=1}^{T}\gamma_t\langle A(X_{t+1/2}),X_{t+1/2}-p\rangle\leq D(p,X_1)$$
$$+\sum_{t=1}^{T}\left[\frac{\gamma_t^2}{2K}-\frac{K}{4\beta^2}\right]\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}-\frac{K}{4\beta^2}\sum_{t=1}^{T}\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*} \quad (7.62)$$

Now, by setting $p=x^*$ with x^* being a solution of (VI) and using the fact that $\langle A(X_{t+1/2}),X_{t+1/2}-x^*\rangle\geq 0$ and $D(x^*,X_1)\leq D'$ (by the compatibility of h), we obtain:

$$\frac{K}{4\beta^2}\sum_{t=1}^{T}\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}\leq D'+\sum_{t=1}^{T}\left[\frac{\gamma_t^2}{2K}-\frac{K}{4\beta^2}\right]\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*} \quad (7.63)$$

Moreover, by observing that the quantity $\left[\frac{\gamma_t^2}{2K}-\frac{K}{4\beta^2}\right]\leq 0$, whenever $\gamma_t\leq\sqrt{2}K/2\beta$ and since we assumed that $\gamma_t\to 0$, there exists some $t_0\in\mathbb{N}$ such that:

$$\left[\frac{\gamma_t^2}{2K}-\frac{K}{4\beta^2}\right]\leq 0 \text{ for all } t\geq t_0 \quad (7.64)$$

Therefore, (7.63) becomes:

$$\frac{1}{\gamma_{T+1}^2} - 1 = \sum_{t=1}^{T} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq D' + \sum_{t=1}^{t_0} \left[\frac{\gamma_t^2}{2K} - \frac{K}{4\beta^2}\right] \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \tag{7.65}$$

In addition, since $1/\gamma_{T+1} \to +\infty$, by the fact that $\gamma_t \to 0$, this yields that:

$$+\infty \leq D' + \sum_{t=1}^{t_0} \left[\frac{\gamma_t^2}{2K} - \frac{K}{4\beta^2}\right] \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \tag{7.66}$$

which is a contradiction. Hence, we get that:

$$\lim_{t \to +\infty} \gamma_t = \inf_{t \in \mathbb{N}} \gamma_t = \gamma_\infty > 0 \tag{7.67}$$

In order to prove our second claim, we first recall the definition of γ_t:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}} \tag{7.68}$$

whereas by developing and rearranging we have:

$$\sum_{j=1}^{t-1} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} = \frac{1}{\gamma_t^2} - 1 \tag{7.69}$$

Hence, by taking limits on both sides we get:

$$\begin{aligned} \sum_{t=1}^{+\infty} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} &= \lim_{t \to +\infty} \sum_{j=1}^{t-1} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \\ &= \frac{1}{\gamma_\infty^2} - 1 \end{aligned}$$

where $0 \leq \frac{1}{\gamma_\infty^2} - 1 < +\infty$, since $0 < \gamma_\infty \leq 1$ and therefore the result follows. □

By means of this lemma, it follows that $\sum_{t=1}^{T} \gamma_t \geq \gamma_\infty T = \Omega(T)$; hence it ultimately follows that (MP) run with (Adapt) enjoys an $\mathcal{O}(1/T)$ rate of convergence under (MS).

Proof of Theorem 7.4. By recalling (7.6) we have:

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - p \rangle \\ &+ \gamma_t \langle A(X_{t+1/2}) - A(X_t), X_{t+1} - X_{t+1/2} \rangle - D(X_{t+1/2}, X_t) - D(X_{t+1}, X_{t+1/2}) \end{aligned} \tag{7.70}$$

We start our analysis rearranging (7.6). In particular, by telescoping $t = 1, \ldots, T$ we get:

$$\sum_{t=1}^{T} \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - p \rangle \leq D(p, X_1) + \sum_{t=1}^{T} \gamma_t \langle A(X_{t+1/2}) - A(X_t), X_{t+1} - X_{t+1/2} \rangle$$

$$- \sum_{t=1}^{T} D(X_{t+1/2}, X_t) - \sum_{t=1}^{T} D(X_{t+1}, X_{t+1/2}) \quad (7.71)$$

On the other hand, since A is monotone, we readily get:

$$\gamma_t \langle A(p), X_{t+1/2} - p \rangle \leq \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - p \rangle \quad (7.72)$$

Thus, combining (7.72) and (7.71), dividing by $\sum_{t=1}^{T} \gamma_t$ and setting

$$\bar{X}_T = \left[\sum_{t=1}^{T} \gamma_t \right]^{-1} \sum_{t=1}^{T} \gamma_t X_{t+1/2} \quad (7.73)$$

we get:

$$\langle A(p), \bar{X}_T - p \rangle \leq \left[\sum_{t=1}^{T} \gamma_t \right]^{-1} (D(p, X_1) + \sum_{t=1}^{T} \gamma_t \langle A(X_{t+1/2}) - A(X_t), X_{t+1} - X_{t+1/2} \rangle - \sum_{t=1}^{T} D(X_{t+1/2}, X_t) - \sum_{t=1}^{T} D(X_{t+1}, X_{t+1/2})) \quad (7.74)$$

whereas, by applying Fenchel-Young inequality to the above we readily get:

$$\langle A(p), \bar{X}_T - p \rangle \leq \left[\sum_{t=1}^{T} \gamma_t \right]^{-1} (D(p, X_1) + \frac{1}{2K} \sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} + \frac{K}{2} \sum_{t=1}^{T} \|X_{t+1} - X_{t+1/2}\|^2_{X_{t+1/2}} - \sum_{t=1}^{T} D(X_{t+1/2}, X_t) - \sum_{t=1}^{T} D(X_{t+1}, X_{t+1/2})) \quad (7.75)$$

Thus, if $\mathcal{C}$ is a compact neighbourhood of the solution set $\mathcal{X}^*$, considering that by Definition 3.3:

$$\frac{K}{2} \sum_{t=1}^{T} \|X_{t+1} - X_{t+1/2}\|^2_{X_{t+1/2}} - \sum_{t=1}^{T} D(X_{t+1}, X_{t+1/2}) \leq 0 \quad (7.76)$$

and taking suprema on both sides, yields:

$$\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T) \leq \left[\sum_{t=1}^{T} \gamma_t \right]^{-1} (\sup_{p \in \mathcal{C}} D(p, X_1) + \frac{1}{2K} \sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} - \sum_{t=1}^{T} D(X_{t+1/2}, X_t)) \quad (7.77)$$

1. Case 1: Convergence under (MB): Therefore, in order to determine the convergence speed of $\overline{X}_T$ under (MB), we shall examine the asymptotic behaviour of each term of the nominator on the (RHS) of (7.87). In particular, we have the following: *Analysis Under (MB)*

   - For the first term: we readily get by the compactness of $\mathcal{C}$,

   $$\sup_{p \in \mathcal{C}} D(p, X_1) \leq D' \text{ for some constant } D' > 0. \quad (7.78)$$

by the Definition 3.3 of the regularizer h.

- For the second term: $\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}$, we have:

$$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} = \sum_{t=1}^{T} (\gamma_t^2 - \gamma_{t+1}^2) \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} + \sum_{t=1}^{T} \gamma_{t+1}^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \quad (7.79)$$

Since, γ_t is non-increasing and therefore $(\gamma_t^2 - \gamma_{t+1}^2 \geq 0)$, and $\gamma_t \leq 1$ the above becomes:

$$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq C^2 + \sum_{t=1}^{T} \gamma_{t+1} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \quad (7.80)$$

and by the definition of γ_t we get:

$$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} = C^2 + \sum_{t=1}^{T} \frac{\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}{1 + \sum_{j=1}^{t} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}} \quad (7.81)$$

and finally,

$$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq C^2 + 1 + \log(1 + \sum_{t=1}^{T} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}) \quad (7.82)$$

with the last inequality being obtained by Lemma A.2 which combined with (MB) yields:

$$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq C^2 + 1 + \log(1 + C^2 T) \quad (7.83)$$

Finally, for $\sum_{t=1}^{T} \gamma_t$, we have the following lower-bound

$$\sum_{t=1}^{T} \gamma_t = \sum_{t=1}^{T} \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}} \geq \sum_{t=1}^{T} \frac{1}{\sqrt{1 + tC^2}} \quad (7.84)$$

which yields:

$$\sum_{t=1}^{T} \gamma_t = \Omega(\sqrt{T}) \quad \text{and} \quad \sum_{t=1}^{T} \gamma_t \to +\infty \quad (7.85)$$

Now, by combining (7.78), (7.83) and (7.85) we readily get that under (MB) we get that:

$$\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T) = \mathcal{O}(1/\sqrt{T}). \quad (7.86)$$

2. Case 2: Convergence under (MS) We now suppose that A satisfies (MS) condition. By applying Lemma 7.6 along with : *Analysis Under (MS)*

$$\sum_{t=1}^{T} \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - p \rangle \leq D(p, X_1) + \frac{1}{2K} \sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} - \sum_{t=1}^{T} D(X_{t+1/2}, X_t) \quad (7.87)$$

by examining the asymptotic behaviour term by term, we get:

- For the first term $D(x^*, X_1)$, since $x^* \in \operatorname{dom} A = \operatorname{dom} h$ and $X_1 \in \operatorname{dom} \partial h$, we have:
$$D(x^*, X_1) < +\infty \quad (7.88)$$
- For the second term $\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}$ we have:
$$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq \sum_{t=1}^{T} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \quad (7.89)$$
and by applying Lemma 7.6 we have:
$$\sum_{t=1}^{T} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq \frac{1}{\gamma_\infty^2} - 1 \quad (7.90)$$
with $\gamma_\infty = \inf_t \gamma_t > 0$.

Finally, by applying Lemma 7.6 once more by considering $\gamma_\infty = \inf_{t \in \mathbb{N}} \gamma_t > 0$ we have:
$$\sum_{t=1}^{T} \gamma_t \geq \gamma_\infty \sum_{t=1}^{T} 1 = \gamma_\infty T \quad (7.91)$$
which yields:
$$\sum_{t=1}^{T} \gamma_t = \Omega(T) \quad (7.92)$$
and the result follows.

□

Having established optimal convergence rate interpolation guarantees for the ergodic average of the (MP) iterates, a natural question that arises what the asymptotic behaviour of the iterates themselves,i.e., before any average occurs. This problem is treated in the next section.

7.3.2 *Trajectory convergence*

Throughout this section we will provide a trajectory convergence result that governs the *actual* iterates of the adaptive (MP) algorithm. Formally, we have the following:

Theorem 7.7 (Antonakopoulos et al. [8]). *Suppose that $\langle A(x), x - x^* \rangle < 0$ whenever x^* is a solution of (VI) and x is not. If, A satisfies (MB) or (MS), the iterates X_t of (MP) run with the adaptive step-size (Adapt) converge to a solution of (VI).*

Last Iterate Convergence

The importance of this result is that, in many practical applications (especially in non-monotone problems), it is more common to harvest the "last iterate" of the method (X_t) rather than its ergodic average $(\bar{X}_T)$; as such, Theorem 7.7 provides a certain justification for this design choice.

Structurally, the first step is to show that X_t visits any neighborhood of a solution point $x^* \in \mathcal{X}^*$ infinitely often (this is where the coherence assumption $\langle A(x), x - x^* \rangle$ is used). The second is to use this trapping property in conjunction with a suitable "energy inequality" to establish convergence via the use of a quasi-Fejér technique as in [36].

Vanishing Residuals

Lemma 7.8 (Antonakopoulos et al. [8]). *Suppose that A satisfies (MB) (respectively (MS)) and $X_t, X_{t+1/2}$ are the iterates of (MP) run with the adaptive step-size (Adapt). Then, the following hold:*

1. $\|X_{t+1/2} - X_t\| \to 0$ *while* $t \to +\infty$
2. $\max\{D(X_{t+1/2}, X_t), D(X_t, X_{t+1/2})\} \leq \frac{2G^2}{K}\gamma_t^2$

Proof. For the proof of the first claim, we shall treat the cases of (MB) and (MS) individually.

1. Under (MB) condition: Since γ_t is decreasing and bounded from below, then we readily obtain that its limit exists and more precisely:

$$\lim_{t\to+\infty} \gamma_t = \gamma_\infty \geq 0 \tag{7.93}$$

We shall distinguish two individual cases:

- $\gamma_\infty > 0$*:* By recalling the definition of the adaptive step-size:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|A(X_{j+1/2}) - A(X_j)\|^2_{X_{j+1/2},*}}} \tag{7.94}$$

whereas by rearranging and developing we have:

$$\sum_{j=1}^{t-1} \|A(X_{j+1/2}) - A(X_j)\|^2_{X_{j+1/2},*} = \frac{1}{\gamma_t^2} - 1 \tag{7.95}$$

Therefore, by taking limits on both sides:

$$\sum_{t=1}^{+\infty} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} = \lim_{t\to+\infty} \frac{1}{\gamma_t^2} - 1 = \frac{1}{\gamma_\infty^2} - 1 \geq 0 \tag{7.96}$$

Hence, by recalling (7.6) we have:

$$\begin{aligned}\sum_{t=1}^{T} D(X_{t+1/2}, X_t) &\leq D(x^*, X_1) + \sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \\ &\leq D(x^*, X_1) + \sum_{t=1}^{T} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}\end{aligned}$$

which in turn by (7.96) yields:

$$\sum_{t=1}^{+\infty} D(X_{t+1/2}, X_t) < +\infty \tag{7.97}$$

and hence $D(X_{t+1/2}, X_t) \to 0$. Moreover, by considering Definition 3.3:

$$\frac{K}{2}\|X_{t+1/2} - X_t\|_{X_t}^2 \leq D(X_{t+1/2}, X_t) \tag{7.98}$$

Now, by recalling $\mu\|\cdot\| \leq \|\cdot\|_x$, we get:

$$\|X_{t+1/2} - X_t\|^2 \leq \frac{1}{\mu^2}\|X_{t+1/2} - X_t\|_{X_t}^2 \tag{7.99}$$

and the result follows.

- $\gamma_\infty = 0$*:* By the prox-step, we get:

$$\begin{aligned}\langle \nabla h(X_t) - \nabla h(X_{t+1/2}), X_t - X_{t+1/2}\rangle &\leq \gamma_t \langle A(X_t), X_t - X_{t+1/2}\rangle \\ &\leq \gamma_t \|A(X_t)\|_{X_t,*}\|X_t - X_{t+1/2}\|_{X_t}\end{aligned} \tag{7.100}$$

On the other hand, we have:

$$\langle \nabla h(X_t) - \nabla h(X_{t+1/2}), X_t - X_{t+1/2}\rangle = D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) \tag{7.101}$$

Thus, we get by Definition 3.3:

$$\begin{aligned}D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) &\leq \gamma_t \|A(X_t)\|_{X_t,*}\|X_t - X_{t+1/2}\|_{X_t} \\ &\leq \gamma_t G\sqrt{\frac{2}{K}\left[D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t)\right]}\end{aligned}$$

where the last inequality is obtained due to (MB). This in turn yields:

$$D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) \leq \frac{2G^2}{K}\gamma_t^2 \tag{7.102}$$

So, a fortiori we have:

$$D(X_t, X_{t+1/2}) \leq \frac{2G^2}{K}\gamma_t^2 \tag{7.103}$$

Moreover, by Definition 3.3:

$$\frac{K}{2}\|X_{t+1/2} - X_t\|_{X_{t+1/2}}^2 \leq D(X_t, X_{t+1/2}) \leq \frac{2G^2}{K}\gamma_t^2 \tag{7.104}$$

Now, by recalling $\mu\|\cdot\| \leq \|\cdot\|_x$, we get:

$$\|X_{t+1/2} - X_t\|^2 \leq \frac{1}{\mu^2}\|X_{t+1/2} - X_t\|_{X_t}^2 \tag{7.105}$$

and the result follows since we assumed that $\gamma_t \to 0$.

2. Under (MS) condition: Following similar reasoning as above, we have:

$$\sum_{t=1}^{T} D(X_{t+1/2}, X_t) \leq D(x^*, X_1) + \sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}$$
$$\leq D(x^*, X_1) + \sum_{t=1}^{T} \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}$$

which by taking limits on both sides and by applying Lemma 7.6 we get that:

$$\sum_{t=1}^{+\infty} D(X_{t+1/2}, X_t) < +\infty \tag{7.106}$$

Therefore, $D(X_{t+1/2}, X_t) \to 0$, whereas by applying Definition 3.3 we obtain:

$$\frac{K}{2}\|X_{t+1/2} - X_t\|^2_{X_t} \leq D(X_{t+1/2}, X_t) \tag{7.107}$$

Now, by recalling $\mu\|\cdot\| \leq \|\cdot\|_x$, we get:

$$\|X_{t+1/2} - X_t\|^2 \leq \frac{1}{\mu^2}\|X_{t+1/2} - X_t\|^2_{X_t} \tag{7.108}$$

and the result follows.

On the other hand, for the second claim, we have by the prox-step:

$$D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) \leq \gamma_t \langle A(X_t), X_{t+1/2} - X_t\rangle$$
$$\leq \gamma_t G \|X_{t+1/2} - X_t\|_{X_t}$$

Therefore, by following the same reasoning with the first claim, we get:

$$D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) \leq \frac{2G^2}{K}\gamma_t^2 \tag{7.109}$$

and hence since $D(\cdot, \cdot) \geq 0$, we have:

$$D(X_{t+1/2}, X_t) \leq \frac{2G^2}{K}\gamma_t^2 \text{ and } D(X_t, X_{t+1/2}) \leq \frac{2G^2}{K}\gamma_t^2 \tag{7.110}$$

and so the result follows □

Extracting a Convergent Sub-sequence

Remark 7.1. We shall point out that (1) in Lemma 7.8 establishes the convergence with respect to the global ambient reference norm of $\mathbb{R}^n$.

Proposition 7.9 (Antonakopoulos et al. [8]). *Suppose that A satisfies* (MB) *(respectively* (MS)*). Then, the iterates $X_t, X_{t+1/2}$ of* (MP) *run with the adaptive step-size* (Adapt) *possess convergent subsequences towards the equilibrium set $\mathcal{X}^*$.*

Proof. By Lemma 7.8, it suffices to show that $X_{t+1/2}$ possesses such a subsequence. Assume to the contrary that it does not. That implies that:

$$\liminf_t \operatorname{dist}(X_{t+1/2}, \mathcal{X}^*) = \delta > 0 \tag{7.111}$$

which in turn yields,

$$\liminf_t \langle A(X_{t+1/2}), X_{t+1/2} - x^* \rangle = c > 0 \tag{7.112}$$

Now, by setting $p = x^*$ for some $x^* \in \mathcal{X}^*$ in (7.6), we get:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - x^* \rangle + \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2}} \\ &\leq D(x^*, X_t) - c\gamma_t + \gamma_t^2 \|A(X_{t+1/2}) - A(X)\|^2_{X_{t+1/2}} \end{aligned}$$

whereas by telescoping $t = 1, \ldots, T$ we obtain:

$$D(x^*, X_T) \leq D(x^*, X_1) - \sum_{t=1}^{T} \gamma_t \left[c - \frac{\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}{\sum_{t=1}^{T} \gamma_t} \right] \tag{7.113}$$

Having this established this general setting, we shall examine the asymptotic behaviour term by term for each regularity case individually, which in both cases shall lead to a contradiction.

1. Under (MB) condition:

   - For the first term: $\sum_{t=1}^{T} \gamma_t$, we have by (7.85) that:

   $$\sum_{t=1}^{T} \gamma_t \to +\infty \quad \text{and} \quad \sum_{t=1}^{T} \gamma_t = \Omega(\sqrt{T}) \tag{7.114}$$

   - For the second term $\dfrac{\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}{\sum_{t=1}^{T} \gamma_t}$, we first examine the denominator. In particular, by the definition of (Adapt) we get:

   $$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} = \sum_{t=1}^{T} \frac{\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}{1 + \sum_{j=1}^{t-1} \|A(X_{j+1/2}) - A(X_j)\|^2_{X_{j+1/2},*}} \tag{7.115}$$

   which by recalling (7.83) we obtain:

   $$\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} = \mathcal{O}(\log T) \tag{7.116}$$

   So, by combining (7.114) and (7.116) we readily obtain:

   $$\frac{\sum_{t=1}^{T} \gamma_t^2 \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}}{\sum_{t=1}^{T} \gamma_t} \to 0 \text{ while } T \to +\infty \tag{7.117}$$

   Therefore, by letting $T \to +\infty$, the inequality (7.113) yields $D(x^*, X_T) \to -\infty$, contradiction.

2. Under (MS) condition: Examining the asymptotic behavior of (7.113) term by term under (MS) we get the following:

- For $\sum_{t=1}^{T}\gamma_t$, (MS) guarantees by (7.92):

$$\sum_{t=1}^{T}\gamma_t = \Omega(T) \;\text{ and }\; \sum_{t=1}^{T}\gamma_t \to +\infty \tag{7.118}$$

- For $\frac{\sum_{t=1}^{T}\gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2,*}}}{\sum_{t=1}^{T}\gamma_t}$, (7.6) guarantees:

$$\sum_{t=1}^{T}\gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2,*}} = \mathcal{O}(1) \tag{7.119}$$

  which combined with (7.92) gives us:

$$\frac{\sum_{t=1}^{T}\gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2,*}}}{\sum_{t=1}^{T}\gamma_t} \to 0 \tag{7.120}$$

Therefore, y letting $T \to +\infty$, the inequality (7.113) yields that $D(x^*, X_T) \to -\infty$, a contradiction.

□

Having all this at hand, we are finally in the position to prove the main result of this section; namely the convergence of the actual iterates of the method. For that we will need an intermediate lemma that shall allow us to pass from a convergent subsequence to global convergence (see also [36], [98]).

Quasi-Fejer Sequences

Lemma 7.10. *Let $\chi \in (0,1]$, $(\alpha_t)_{t\in\mathbb{N}}$, $(\beta_t)_{t\in\mathbb{N}}$ non-negative sequences and $(\varepsilon_t)_{t\in\mathbb{N}} \in l^1(\mathbb{N})$ such that $t = 1, 2, \ldots$:*

$$\alpha_{t+1} \leq \chi\alpha_t - \beta_t + \varepsilon_t \tag{7.121}$$

Then, α_t converges.

Proof. First, one shows that $\alpha_{t\in\mathbb{N}}$ is a bounded sequence. Indeed, one can derive directly that:

$$\alpha_{t+1} \leq \chi^{t+1}\alpha_0 + \sum_{k=0}^{t}\chi^{t-k}\varepsilon_k \tag{7.122}$$

Hence, $(\alpha_t)_{t\in\mathbb{N}}$ lies in $[0, \alpha_0 + \varepsilon]$, with $\varepsilon = \sum_{t=0}^{+\infty}\varepsilon_t$. Now, one is able to extract a convergent subsequence $(\alpha_{k_t})_{t\in\mathbb{N}}$, let say $\lim_{t\to+\infty}\alpha_{k_t} = \alpha \in [0, \alpha_0 + \varepsilon]$ and fix $\delta > 0$. Then, one can find some t_0 such that $\alpha_{k_{t_0}} - \alpha < \frac{\delta}{2}$ and $\sum_{m>t_{k_{t_0}}}\varepsilon_m < \frac{\delta}{2}$. That said, we have:

$$0 \leq \alpha_t \leq \alpha_{k_{t_0}} + \sum_{m>t_{k_{t_0}}}\varepsilon_m < \frac{\delta}{2} + \alpha + \frac{\delta}{2} = \alpha + \delta \tag{7.123}$$

Hence, $\limsup_t \alpha_t \leq \liminf_t \alpha_t + \delta$. Since, δ is chosen arbitrarily the result follows.

□

Proof of Theorem 7.7. Once more, we shall treat each regularity class individually.

1. Under (MB) condition: For the (MB), by denoting $\lim_{t\to+\infty}\gamma_t = \gamma_\infty$ case we shall consider two cases for the asymptotic behaviour of the step-size γ_t.

   - $\gamma_\infty > 0$*:* By recalling the definition of γ_t:

   $$\gamma_t = \frac{1}{\sqrt{1+\sum_{j=1}^{t-1}\|A(X_{j+1/2})-A(X_j)\|^2_{X_{j+1/2}}}} \tag{7.124}$$

   whereas by rearranging we get:

   $$\sum_{j=1}^{t-1}\|A(X_{j+1/1})-A(X_j)\|^2_{X_{j+1/2}} = \frac{1}{\gamma_t^2}-1 \tag{7.125}$$

   and hence:

   $$\sum_{t=1}^{+\infty}\|A(X_{t+1/1})-A(X_t)\|^2_{X_{t+1/2}} = \frac{1}{\gamma_\infty^2}-1 < +\infty \tag{7.126}$$

   Therefore, by recalling (7.6), we have for solution of (VI), $x^* \in \mathcal{X}$

   $$\begin{aligned} D(x^*, X_{t+1}) \leq D(x^*, X_t) &- \gamma_t\langle A(X_{t+1/2}), X_{t+1/2}-x^*\rangle \\ &+ \gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{t+1/2,*} \end{aligned} \tag{7.127}$$

   which enables us to directly apply Lemma 7.10 for $\alpha_t = D(x^*, X_t)$, $\beta_t = \gamma_t\langle A(X_{t+1/2}), X_{t+1/2}-x^*\rangle$ and $\varepsilon_t = \gamma_t^2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}$.

   - $\gamma_\infty = 0$*:* Fix an equilibrium $x^* \in \mathcal{X}^*$ and consider the "Bregman zone":

   $$D_\varepsilon = \{x \in \mathcal{X} : D(x^*, x) < \varepsilon\} \tag{7.128}$$

   By the assumption for the regularizer h, it follows that there exists some $\delta > 0$ such that:

   $$B_\delta = \{x \in \mathcal{X} : \|x^*-x\| < \delta\} \tag{7.129}$$

   is contained in D_ε. Hence, by regularity assumption for the (2.10), it follows that:

   $$\langle A(x), x-x^*\rangle \geq c > 0 \text{ for some } c \equiv c(\varepsilon) > 0 \text{ and for all } x \notin D_\varepsilon, \tag{7.130}$$

   in particular, for all $x \in D_{2\varepsilon}\setminus D_\varepsilon$. Assume now that x^* is a limit point of X_t, i.e., $X_t \in D_{2\varepsilon}$ for infinitely many $t \in \mathbb{N}$. Now, by the prox-step, we get:

   $$\gamma_t\langle A(X_t), X_t-x^*\rangle \leq \langle \nabla h(X_t)-\nabla h(X_{t+1/2}), X_t-x^*\rangle \tag{7.131}$$

   whereas by Lemma 4.2 and after rearranging we get:

   $$\begin{aligned} D(x^*, X_{t+1/2}) &\leq D(x^*, X_t) - \gamma_t\langle A(X_t), X_t-x^*\rangle + D(X_t, X_{t+1/2}) \\ &\leq D(x^*, X_t) - \gamma_t\langle A(X_t), X_t-x^*\rangle + \max\{D(X_t, X_{t+1/2}), D(X_t, X_{t+1/2})\} \end{aligned}$$

Therefore, by Lemma 7.8 we obtain:

$$D(x^*, X_{t+1/2}) \leq D(x^*, X_t) - \gamma_t \langle A(X_t), X_t - x^* \rangle + \frac{2G^2}{K}\gamma_t^2 \tag{7.132}$$

We consider two cases:

a) $X_t \in D_{2\varepsilon} \setminus D_\varepsilon$: Then, $\langle A(X_t), X_t - x^* \rangle \geq c > 0$. So,

$$D(x^*, X_{t+1/2}) \leq D(x^*, X_t) - c\gamma_t + \frac{2G^2}{K}\gamma_t^2 \tag{7.133}$$

Now, provided that $\frac{2G^2\gamma_t^2}{K} \leq c\gamma_t$ or equivalently $\gamma_t \leq \frac{cK}{2G^2}$. we get: $D(x^*, X_{t+1/2}) \leq 2\varepsilon$.

b) $X_t \in D_\varepsilon$: Then, in this case we have:

$$D(x^*, X_{t+1/2}) \leq D(x^*, X_t) + \frac{2G^2}{K}\gamma_t^2 \tag{7.134}$$

Again, provided that $\frac{2G^2}{K}\gamma_t^2 \leq \varepsilon$ or equivalently $\gamma_t \leq \frac{\sqrt{2\varepsilon K}}{2G}$ we get $D(x^*, X_{t+1/2}) \leq 2\varepsilon$

Therefore, by summarizing the above we get that if $\gamma_t \leq \min\{\frac{\sqrt{2\varepsilon K}}{2G}, \frac{cK}{2G^2}\}$, we have that $X_{t+1/2} \in D_{2\varepsilon}$ whenever $X_t \in D_{2\varepsilon}$. Going further, due to Proposition 4.4 by setting $p = x^*$, $x_1 = X_{t+1/2}$, $x_2^+ = X_{t+1}$, $x = X_t$, $w_1 = -\gamma_t A(X_{t+1/2})$ and $w_2 = -\gamma_t A(X_{t+1/2})$ we get:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - x^* \rangle \\ &+ \gamma_t \langle A(X_{t+1/2}) - A(X_t), X_{t+1} - X_{t+1/2} \rangle - D(X_{t+1}, X_{t+1/2}) - D(X_{t+1/2}, X_t) \end{aligned} \tag{7.135}$$

whereas by applying Fenchel's inequality we obtain:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - x^* \rangle \\ &+ \frac{\gamma_t^2}{2K}\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} + \frac{K}{2}\|X_{t+1} - X_{t+1/2}\|^2_{X_{t+1/2}} \\ &- D(X_{t+1}, X_{t+1/2}) - D(X_{t+1/2}, X_t) \end{aligned} \tag{7.136}$$

Now, since $\frac{K}{2}\|X_{t+1} - X_{t+1/2}\|^2_{X_{t+1/2}} - D(X_{t+1}, X_{t+1/2}) \leq 0$ by Definition 3.3 we get:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - x^* \rangle \\ &+ \frac{\gamma_t^2}{2K}\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \end{aligned} \tag{7.137}$$

which, in turn, by (7.53) the above yields:

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) - \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - x^* \rangle + \frac{C^2}{2K}\gamma_t^2 \tag{7.138}$$

with $C = 2G + \beta\frac{4G}{K}$. Recall that $X_{t+1/2} \in D_{2\varepsilon}$ by our previous claim. We now consider the following two cases:

a) $X_{t+1/2} \in D_{2\varepsilon} \setminus D_\varepsilon$: In this case: $\langle A(X_{t+1/2}), X_{t+1/2} - x^*\rangle \geq c > 0$, so,

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) - c\gamma_t + \frac{C^2}{2K}\gamma_t^2 \tag{7.139}$$

which holds provided that $\frac{C^2\gamma_t^2}{2K} \leq c\gamma_t$ or equivalently $\gamma_t \leq \frac{2cK}{C^2}$,

b) $X_{t+1/2} \in D_\varepsilon$: First recall that:

$$\begin{aligned} D(X_{t+1/2}, X_{t+1}) + D(X_{t+1}, X_{t+1/2}) &\leq \frac{2\gamma_t^2}{K}\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \\ &\leq \frac{2\gamma_t^2}{K}C^2 \end{aligned}$$

Therefore, we get that:

$$\|X_{t+1} - X_{t+1/2}\|^2 \leq \frac{4\mu^2C^2}{K^2}\gamma_t^2 \tag{7.140}$$

Now, let us define the following:

$$D_\varepsilon(\alpha) = \max\{D(x^*, x) : \operatorname{dist}(x, D_\varepsilon(x^*)) < \alpha\} \tag{7.141}$$

Clearly, $D_\varepsilon(\alpha)$ is continuous relative to α and $\lim_{\alpha\to 0^+} D_\varepsilon(\alpha) = \varepsilon$. Therefore, we have:

$$D_\varepsilon(\alpha) \leq \varepsilon \quad \text{for all } \alpha \leq \alpha^* \text{ with } \alpha^* \text{ sufficiently small.} \tag{7.142}$$

Moreover, due to (7.140), we conclude that $D(x^*, X_{t+1}) \leq 2\varepsilon$, provided that $\gamma_t \leq \frac{\alpha^*}{2\mu C}K$.

We conclude that $X_{t+1} \in U_{2\varepsilon}$ provided that $X_t \in D_{2\varepsilon}$ and

$$\gamma_t \leq \min\{\frac{2cK}{G^2}, \frac{\sqrt{2\varepsilon K}}{2G}, \frac{\alpha^*}{2\mu C}K\} \tag{7.143}$$

Since, $\gamma_t \to 0$ and $X_t \in D_{2\varepsilon}$ infinitely often (due to Proposition 7.9) we conclude that $X_t \in D_{2\varepsilon}$ for all sufficiently large t. With $\varepsilon > 0$ being arbitrary, the result follows.

2. Under (MS) condition: By plugging in $\alpha_t = D(x^*, X_t)$, $\beta_t = \gamma_t\langle A(X_{t+1/2}, X_{t+1/2} - x^*\rangle$ and $\varepsilon_t = \gamma_t^2\|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*}$ in Lemma 7.10 and combine it with Lemma 7.6, we get $\inf_{x^*\in\mathcal{X}^*}\|x^*, X_t\|$ converges. Thus, the result follows by applying Proposition 7.9.

□

Having established the last iterate convergence of the adaptive method we proceed to evaluate numerically the performance of the method.

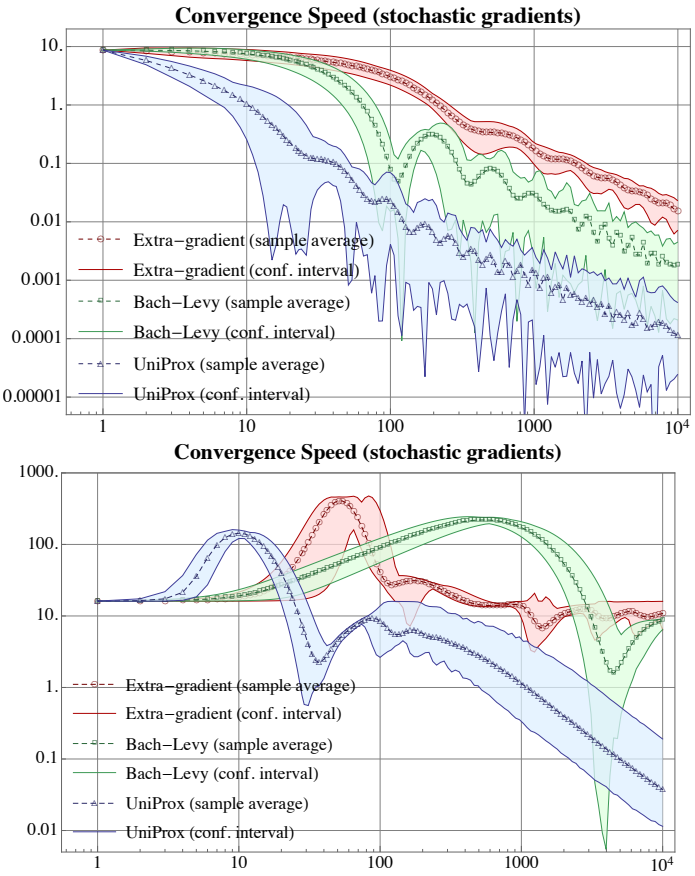


Figure 7.1: Numerical comparison between the extra-gradient (EG), Bach–Levy (BL) and AdaProx algorithms (red circles, green squares and blue triangles respectively). The figure on the left shows the methods' convergence in a 100×100 bilinear game; the one on the right shows the methods' convergence in a non-convex/non-concave covariance learning problem. In both cases, the parameters of the EG and BL algorithms have been tuned with a grid search (AdaProx has no parameters to tune). All curves have been averaged over $S = 100$ sample runs, and the 95% confidence interval is indicated by the shaded area.

7.3.3 *Numerical evaluation*

We conclude in this section with a numerical illustration of the convergence properties of AdaProx in two different settings: *a*) bilinear min-max games; and *b*) a simple Wasserstein GAN in the spirit of Daskalakis et al. [37] with the aim of learning an unknown covariance matrix.

BILINEAR MIN-MAX GAMES. For our first set of experiments, we consider a min-max game of the form of the form $\Phi(x_1, x_2) = (x_1 - x_1^*)^\top A(x_2 - x_2^*)$ with $x_1, x_2 \in \mathbb{R}^{100}$ and $A \in \mathbb{R}^{100} \times \mathbb{R}^{100}$ (drawn i.i.d. component-wise from a standard Gaussian). To test the convergence of AdaProx beyond the "full gradient" framework, we ran the algorithm with stochastic gradient signals of the form $V_t = A(X_t) + U_t$ where U_t is drawn i.i.d. from a centered Gaussian distribution with unit covariance matrix. We then plotted in Fig. 7.1 the squared gradient norm $\|A(\bar{X}_T)\|^2$ of the method's ergodic average $\bar{X}_T$ after T iterations (so values closer to zero are better). For

benchmarking purposes, we also ran the extra-gradient (EG) and Bach–Levy (BL) algorithms [14] with the same random seed for the simulated gradient noise. The step-size parameter of the EG algorithm was chosen as $\gamma_t = 0.025/\sqrt{t}$, whereas the BL algorithm was run with diameter and gradient bound estimation parameters $D_0 = .5$ and $M_0 = 2.5$ respectively (both determined after a hyper-parameter search since the only *theoretically* allowable values are $D_0 = M_0 = \infty$; interestingly, very large values for D_0 and M_0 did not yield good results). The experiment was repeated $S = 100$ times, and AdaProx gave consistently faster rates.

COVARIANCE MATRIX LEARNING. Going a step further, consider the covariance learning game

$$\Phi(x_1, x_2) = \mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}[x^\top x_1 x] - \mathbb{E}_{z\sim\mathcal{N}(0,I)}[z^\top x_1^\top x_2 x_1 z], \qquad x_1, x_2 \in \mathbb{R}^n \times \mathbb{R}^n. \tag{7.144}$$

The goal here is to generate data drawn from a centered Gaussian distribution with unknown covariance Σ; in particular, this model follows the Wasserstein GAN formulation of Daskalakis et al. [37] with generator and discriminator respectively given by $G(z) = x_1 z$ and $D(x) = x^\top x_2 x$ (no clipping). For the experiments, we took $n = 100$, a mini-batch of $m = 128$ samples per update, and we ran the EG, BL and AdaProx algorithms as above, tracing the square norm of A as a measure of convergence. Since the problem is non-monotone, there are several disjoint equilibrium components so the algorithms' behavior is considerably more erratic; however, after this initial warm-up phase, AdaProx again gave the faster convergence rates.

7.4 UNIVERSALITY IN THE PRESENCE OF NOISE

In order to derive our general universality result, we change gears from the (MP) template. In particular, we shall adopt a primal-dual approach; more precisely, our focal point is that of the *dual extrapolation* template presented in (DualX). We recall that the said method is defined by the following recursion:

Universal Dual Extrapolation

$$\begin{aligned} X_{t+1/2} &= P_{X_t}(-\gamma_t V_t) \\ Y_{t+1} &= Y_t - V_{t+1/2} \\ X_{t+1} &= Q(\gamma_{t+1} Y_{t+1}) \end{aligned}$$

Throughout this section, given the dual extrapolation method run once more with the adaptive learning rate (Adapt):

Universal Step-Size

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|V_{j+1/2} - V_j\|^2_{X_{j+1/2},*}}} \tag{Adapt}$$

In addition, we assume that the optimizer has access to a first order oracle of the form (SFO) which satisfies the following statistical assumptions:

Blanket Assumptions

Assumption 7.3. 1. *Zero-mean noise*, i.e.,

$$\mathbb{E}[U_t | \mathcal{F}_t] = 0 \quad \text{for all } t = 1, 2, \dots. \tag{7.145a}$$

2. *Boundedness with probability* 1, i.e., there exists some $\sigma^2 > 0$ such that

$$\|U_t\|_*^2 \leq \sigma^2 \quad \text{almost surely for all } t = 1, 2, \dots \tag{7.145b}$$

Furthermore, concerning the regularity conditions assumed for the associated operators and the ambient space the same assumptions hold as in Section 7.3, i.e.,

1. $\mathcal{X}$ is a regular Finsler space (cf. Section 3.2.2).
2. The respective generalizations of the standard Lipschitz regularity are: given a family of local norms $\|\cdot\|_x$ with $x \in \mathcal{X}$, the respective monotone operators under study will be that satisfying (MB) and/or (MS)
3. The associated regularizer h is a Bregman-Finsler function, i.e., satisfies Definition 3.3.

Having all this at hand, we are in position to present the main result of this section; namely we present optimal convergence rate guarantees for both deterministic and stochastic settings. Formally, we have the following theorem.

Universality Guarantees (Stochastic & Deterministic)

Theorem 7.11 (Antonakopoulos and Mertikopoulos [5])**.** *Assume that $X_{t+1/2}, X_t$ are the* (DualX) *iterates run with the adaptive step-size policy* (Adapt) *and a* (SFO) *satisfying* (7.145b)*. Then, the following hold:*

1. *If A satisfies* (MB)*, then,*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{7.146}$$

2. *If A satisfies* (MS)*, then,*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] = \mathcal{O}\left(\frac{A}{T} + \frac{B\sigma}{\sqrt{T}}\right) \tag{7.147}$$

In order to prove Theorem 7.11 we will use extensively a key template which connects iterates after the. respective prox and mirror steps. In what follows we will illustrate this in a detailed manner.

7.4.1 *Template inequalities*

The proof of Theorem 7.11 hinges again on a primal-dual type template inequality which involves *Fenchel couplings* instead of Bregman divergences as in (MP) setting. Namely, we seek to prove an inequality of the form:

$$\sum_{t=1}^{T} \langle V_{t+1/2}, X_{t+1/2} - x \rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \sum_{t=1}^{T} \langle V_{t+1/2} - V_t, Q(\gamma_t Y_{t+1}) - X_{t+1/2} \rangle - \sum_{t=1}^{T} \frac{1}{\gamma_t} D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \sum_{t=1}^{T} \frac{1}{\gamma_t} D(X_{t+1/2}, X_t) \tag{7.148}$$

Template Inequality for Universality

In doing so we will need the following result.

Lemma 7.12 (Antonakopoulos and Mertikopoulos [5]). *If $X_{t+1/2}, X_t$ are the iterates of (DualX) run with a decreasing learning rate γ_t, then the following inequality holds for all $x \in \mathcal{X}$:*

$$\frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) \leq \frac{1}{\gamma_t}F(x,\gamma_t Y_t) - \langle V_{t+1/2}, X_{t+1/2} - x\rangle$$
$$+ \langle V_{t+1/2} - V_t, Q(\gamma_{tY_{t+1}}) - X_{t+1/2}\rangle - \frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \frac{1}{\gamma_t}D(X_{t+1/2}, X_t) \tag{7.149}$$

Proof. For all $x \in \mathcal{X}$ we have:

$$\begin{aligned}\langle V_{t+1/2}, Q(\gamma_t Y_{t+1}) - x\rangle &= \frac{1}{\gamma_t}\langle \gamma_t Y_t - \gamma_t Y_{t+1}, Q(\gamma_t Y_{t+1}) - x\rangle \\ &= \frac{1}{\gamma_t}F(x,\gamma_t Y_t) - \frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) - \frac{1}{\gamma_t}F(Q(\gamma_t Y_{t+1}),\gamma_t Y_t)\end{aligned}$$

Therefore, by rearranging we get:

$$\begin{aligned}\frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) &= \frac{1}{\gamma_t}F(x,\gamma_t Y_t) - \langle V_{t+1/2}, Q(\gamma_t Y_{t+1}) - x\rangle - \frac{1}{\gamma_t}F(Q(\gamma_t Y_{t+1}),\gamma_t Y_t) \\ &= \frac{1}{\gamma t}F(x,\gamma_{tY_t}) - \langle V_{t+1/2}, X_{t+1/2} - x\rangle + \langle V_{t+1/2}, Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle \\ &\quad - \frac{1}{\gamma_t}F(Q(\gamma_t Y_{t+1}),\gamma_t Y_t)\end{aligned}$$

and since $F(Q(\gamma_t Y_{t+1}),\gamma_t Y_t) \geq D(Q(\gamma_t Y_{t+1}), X_t)$ the above becomes:

$$\frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) \leq \frac{1}{\gamma_t}F(x,\gamma_t Y_t) - \langle V_{t+1/2}, X_{t+1/2} - x\rangle + \langle V_{t+1/2}), Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle$$
$$- \frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_t) \tag{7.150}$$

On the other hand, by the prox-step we have:

$$\begin{aligned}\langle V_t, X_{t+1/2} - x\rangle &\leq \frac{1}{\gamma_t}\langle \nabla h(X_t) - \nabla h(X_{t+1/2}), X_{t+1/2} - x\rangle \\ &= \frac{1}{\gamma_t}\langle \nabla h(X_{t+1/2}) - \nabla h(X_t), x - X_{t+1/2}\rangle \\ &= \frac{1}{\gamma_t}D(x, X_t) - \frac{1}{\gamma_t}D(x, X_{t+1/2}) - \frac{1}{\gamma_t}D(X_{t+1/2}, X_t)\end{aligned}$$

where the last equality is obtained by Lemma 4.2. Hence, by rearranging and setting $x = Q(\gamma_t Y_{t+1})$ we get:

$$\frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) + \frac{1}{\gamma_t}D(X_{t+1/2}, X_t) + \langle V_t, X_{t+1/2} - Q(\gamma_t Y_{t+1})\rangle \leq \frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_t) \tag{7.151}$$

Thus, by combining the above inequalities we obtain:

$$\frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) \leq \frac{1}{\gamma_t}F(x,\gamma_t Y_t) - \langle V_{t+1/2}, X_{t+1/2} - x\rangle + \langle V_{t+1/2}, Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle + \langle V_t, X_{t+1/2} - Q(\gamma_t Y_{t+1})\rangle - \frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \frac{1}{\gamma_t}D(X_{t+1/2}, X_t) \tag{7.152}$$

So, finally we have:

$$\frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) \leq \frac{1}{\gamma_t}F(x,\gamma_t Y_t) - \langle V_{t+1/2}, X_{t+1/2} - x\rangle + \langle V_{t+1/2} - V_t, Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle - \frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \frac{1}{\gamma_t}D(X_{t+1/2}, X_t) \tag{7.153}$$

and the result follows. □

Now armed Lemma 7.12, we are in the position to prove our main "energy" inequality (7.148).

Regret Inequality

Proposition 7.13. *Assume that $X_{t+1/2}, X_t$ are the iterates of (DualX) run with a decreasing learning rate γ_t. Then, for all $x \in \mathcal{X}$, the following "regret" estimation holds:*

$$\sum_{t=1}^{T}\langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \sum_{t=1}^{T}\langle V_{t+1/2} - V_t, Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle - \sum_{t=1}^{T}\frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \sum_{t=1}^{T}\frac{1}{\gamma_t}D(X_{t+1/2}, X_t) \tag{7.154}$$

Proof. By setting $E_t = \frac{1}{\gamma_t}F(x,\gamma_t Y_t)$, we have:

$$\begin{aligned} E_{t+1} - E_t &= \frac{1}{\gamma_{t+1}}F(x,\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}F(x,\gamma_t Y_t) \\ &= \left[\frac{1}{\gamma_{t+1}}F(x,\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1})\right] + \left[\frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) - \frac{1}{\gamma_t}F(x,\gamma_t Y_t)\right] \end{aligned}$$

Now let us deal with each bracket individually.

- For the first term $\left[\frac{1}{\gamma_{t+1}}F(x,\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1})\right]$ we have:

  $$\frac{1}{\gamma_{t+1}}F(x,\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) = \frac{1}{\gamma_{t+1}}h(x) + \frac{1}{\gamma_{t+1}}h^*(\gamma_{t+1}Y_{t+1}) - \langle Y_{t+1}, x\rangle - \frac{1}{\gamma_t}h(x) - \frac{1}{\gamma_t}h^*(\gamma_t Y_{t+1}) + \langle Y_{t+1}, x\rangle$$

  and hence we have:

  $$\frac{1}{\gamma_{t+1}}F(x,\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}F(x,\gamma_t Y_{t+1}) = \left[\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right]h(x) + \frac{1}{\gamma_{t+1}}h^*(\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}h^*(\gamma_t Y_{t+1}) \tag{7.155}$$

Now, considering the function:

$$\phi(\gamma) = \frac{1}{\gamma}\left[h^*(\gamma w) + \min h\right] \tag{7.156}$$

By taking the derivative (with respect to γ) we get:

$$\begin{aligned}\phi'(\gamma) &= \frac{1}{\gamma}\langle w, Q(\gamma w)\rangle - \frac{1}{\gamma^2}\left[h^*(\gamma w) + \min h\right] \\ &= \frac{1}{\gamma^2}\left[\langle \gamma w, Q(\gamma w)\rangle - h^*(\gamma w) + \min h\right] \\ &= \frac{1}{\gamma^2}\left[h(Q(\gamma w)) - \min h\right] \\ &\geq 0\end{aligned}$$

Therefore, we get that ϕ is non-decreasing function and since $\gamma_{t+1} \leq \gamma_t$ we have $\phi(\gamma_{t+1}) \leq \phi(\gamma_t)$, i.e.,:

$$\frac{1}{\gamma_{t+1}}\left[h^*(\gamma_{t+1}Y_{t+1}) + \min h\right] \leq \frac{1}{\gamma_t}\left[h^*(\gamma_t Y_{t+1}) + \min h\right] \tag{7.157}$$

where after rearranging we obtain:

$$\frac{1}{\gamma_{t+1}}h^*(\gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}h^*(\gamma_t Y_{t+1}) \leq \left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t+1}}\right]\min h \tag{7.158}$$

Hence, by combining all the above we get:

$$\frac{1}{\gamma_{t+1}}F(x, \gamma_{t+1}Y_{t+1}) - \frac{1}{\gamma_t}F(x, \gamma_t Y_{t+1}) \leq \left[\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right]h(x) - \min h \tag{7.159}$$

- For the second term $\left[\frac{1}{\gamma_t}F(x, \gamma_t Y_{t+1}) - \frac{1}{\gamma_t}F(x, \gamma_t Y_t)\right]$ we readily get by Lemma 7.12:

$$\begin{aligned}&\frac{1}{\gamma_t}F(x, \gamma_t Y_{t+1}) - \frac{1}{\gamma_t}F(x, \gamma_t Y_t) \leq \langle V_{t+1/2}, X_{t+1/2} - x\rangle \\ &+ \langle V_{t+1/2} - V_t, Q(\gamma_{tY_{t+1}}) - X_{t+1/2}\rangle - \frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \frac{1}{\gamma_t}D(X_{t+1/2}, X_t)\end{aligned} \tag{7.160}$$

Hence, combining all this and after telescoping through $t = 1, \ldots, T$ we get:

$$\begin{aligned}&\sum_{t=1}^{T}\langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{1}{\gamma_1}F(x, Y_1) + \left[\frac{1}{\gamma_{T+1}} - \frac{1}{\gamma_1}\right](h(x) - \min h) \\ &+ \sum_{t=1}^{T}\langle V_{t+1/2} - V_t, Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle - \sum_{t=1}^{T}\frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \sum_{t=1}^{T}\frac{1}{\gamma_t}D(X_{t+1/2}, X_t)\end{aligned} \tag{7.161}$$

Moreover, by setting $Y_1 = 0$ we get:

$$\frac{1}{\gamma_1}F(x, Y_1) = \frac{1}{\gamma_1}h(x) + \frac{1}{\gamma_1}h^*(0) = \frac{1}{\gamma_1}\left[h(x) - \min h\right] \tag{7.162}$$

So, finally we get:

$$\sum_{t=1}^{T}\langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \sum_{t=1}^{T}\langle V_{t+1/2} - V_t, Q(\gamma_t Y_{t+1}) - X_{t+1/2}\rangle - \sum_{t=1}^{T}\frac{1}{\gamma_t}D(Q(\gamma_t Y_{t+1}), X_{t+1/2}) - \sum_{t=1}^{T}\frac{1}{\gamma_t}D(X_{t+1/2}, X_t) \tag{7.163}$$

and the result follows. □

7.4.2 *Optimal rate interpolation analysis*

In order to proceed to the particular analysis of Theorem 7.11 we shall some additional results. The first concerns the almost sure boundedness of $\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}$ which is vital for establishing the stochastic rates under (MB). Formally, we. have the following result.

Almost Sure Boundedness of the Residual

Lemma 7.14 (Antonakopoulos and Mertikopoulos [5])**.** *Assume that $X_t, X_{t+1/2}$ are the iterates of (DualX) run with a non-increasing step-size γ_t. Moreover, assume that the oracle satisfies the mean square boundedness condition:*

$$\mathbb{E}\left[\|V(x;\omega)\|^2_{x,*}\right] \leq G^2 \tag{7.164}$$

Then, the sequence $\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}$ is bounded almost surely. In particular, the following inequality holds with probability 1:

$$\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*} \leq 4\left[G^2 + \mu^2\sigma^2\right] + 2\left[\sqrt{2G^2 + 2\mu^2\sigma^2} + \frac{2\beta\gamma_1}{K}\left[2G^2 + 2\mu^2\sigma^2\right]\right]^2 \tag{7.165}$$

Proof. We have:

$$\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*} \leq 2\|V_{t+1/2}\|^2_{X_{t+1/2},*} + 2\|V_t\|^2_{X_{t+1/2},*} \tag{7.166}$$

Now, let us bound each term of the above individually. For the term $\|V_{t+1/2}\|^2_{X_{t+1/2},*}$ we have:

$$\begin{aligned}\|V_{t+1/2}\|^2_{X_{t+1/2},*} &= \|A(X_{t+1/2}) + U_{t+1/2}\|^2_{X_{t+1/2},*}\\ &\leq 2\|A(X_{t+1/2})\|^2_{X_{t+1/2},*} + 2\|U_{t+1/2}\|^2_{X_{t+1/2},*}\\ &\leq 2\|A(X_{t+1/2})\|^2_{X_{t+1/2},*} + 2\mu^2\|U_{t+1/2}\|^2_*\\ &\leq 2\|A(X_{t+1/2})\|^2_{X_{t+1/2},*} + 2\mu^2\sigma^2\end{aligned}$$

Moreover, we have:

$$\|A(X_{t+1/2})\|^2_{X_{t+1/2},*} = \|\mathbb{E}\left[A(X_{t+1/2})|\mathcal{F}_{t+1/2}\right]\|^2_{X_{t+1/2},*}$$

$$
\begin{aligned}
&= \|\mathbb{E}\left[A(X_{t+1/2})|\mathcal{F}_{t+1/2}\right] + \mathbb{E}\left[U_{t+1/2}|\mathcal{F}_{t+1/2}\right]\|^2_{X_{t+1/2},*} \\
&= \|\mathbb{E}\left[V_{t+1/2}|\mathcal{F}_{t+1/2}\right]\|^2_{X_{t+1/2},*} \\
&\leq \mathbb{E}\left[\|V_{t+1/2}\|^2_{X_{t+1/2},*}|\mathcal{F}_{t+1/2}\right]
\end{aligned}
$$

with the last inequality being obtained by applying Jensen's inequality. Hence, since the oracle satisfies the mean square boundedness condition we get:

$$\|A(X_{t+1/2})\|^2_{X_{t+1/2},*} \leq G^2 \tag{7.167}$$

Therefore, summarizing: $\|V_{t+1/2}\|^2_{X_{t+1/2},*}$ is upper bounded almost surely by:

$$\|V_{t+1/2}\|^2_{X_{t+1/2},*} \leq 2G^2 + 2\mu^2\sigma^2 \tag{7.168}$$

Now, we turn our attention to $\|V_t\|^2_{X_{t+1/2},*}$, recalling the regularity of $\mathcal{X}$ we have for some $\beta \geq 0$ such that:

$$\frac{\|V_t\|_{X_{t+/2},*}}{\|V_t\|_{X_t,*}} \leq 1 + \beta\|X_t - X_{t+1/2}\|_{X_t} \tag{7.169}$$

or equivalently:

$$\|V_t\|_{X_{t+1/2},*} \leq \|V_t\|_{X_t,*} + \beta\|V_t\|_{X_t,*}\|X_t - X_{t+1/2}\|_{X_t} \tag{7.170}$$

Now, by the definition of (DualX) we have:

$$
\begin{aligned}
\gamma_t\langle V_t, X_t - X_{t+1/2}\rangle &\geq D(X_t, X_{t+1/2}) + D(X_{t+1/2}, X_t) \\
&\geq \frac{K}{2}\|X_t - X_{t+1/2}\|^2_{X_t}
\end{aligned}
$$

with the last inequality being obtained by applying Definition 3.3. So, by applying Cauchy-Shwartz inequality on the (LHS) we get:

$$
\begin{aligned}
\frac{K}{2}\|X_t - X_{t+1/2}\|^2_{X_t} &\leq \gamma_t\langle V_t, X_t - X_{t+1/2}\rangle \\
&\leq \gamma_t\|V_t\|_{X_t,*}\|X_{t+1/2} - X_t\|_{X_t}
\end{aligned}
$$

and so,

$$\|X_t - X_{t+1/2}\|_{X_t} \leq \frac{2\beta\gamma_1}{K}\|V_t\|_{X_t,*} \tag{7.171}$$

So, combining (7.170) and (7.171) we get:

$$\|V_t\|_{X_{t+1/2},*} \leq \|V_t\|_{X_t,*} + \frac{2\beta\gamma_1}{K}\|V_t\|^2_{X_t,*} \tag{7.172}$$

Thus, what is left is to upper bound $\|V_t\|_{t,*}$. Working in the same spirit as above, we get that:

$$\|V_t\|^2_{X_t,*} \leq 2G^2 + 2\mu^2\sigma^2 \tag{7.173}$$

Hence, combining (7.172) with (7.173) we get:

$$\|V_t\|_{X_{t+1/2},*} \leq \sqrt{2G^2 + 2\mu^2\sigma^2} + \frac{2\beta\gamma_1}{K}\left[2G^2 + 2\mu^2\sigma^2\right] \tag{7.174}$$

Finally, by combining (7.166) with (7.168) and (7.174) we get:

$$\|V_t - V_{t+1/2}\|^2_{X_{t+1/2}} \leq 4\left[G^2 + \mu^2\sigma^2\right] + 2\left[\sqrt{2G^2 + 2\mu^2\sigma^2} + \frac{2\beta\gamma_1}{K}\left[2G^2 + 2\mu^2\sigma^2\right]\right]^2 \tag{7.175}$$

and the proof is complete. □

Moving forward, the second crucial ingredient for establishing our main result we will need a result for martingale differences, introduced by [14, 57].

Proposition 7.15 (Bach and Levy [14], Kakade [57]). *Let $\mathcal{C} \subset \mathbb{R}^n$ and $h : \mathcal{X} \to \mathbb{R}$ be a Bregman function. Also assume that for all $x \in \mathcal{C}$ we have:*

$$h(x) - \min_{x\in\mathcal{C}} h(x) \leq \frac{1}{2}D^2 \tag{7.176}$$

Martingale Difference Estimation

Then, for any martingale difference sequence $(\zeta_t) \in \mathbb{R}^n$ and any random vector x defined over $\mathcal{C}$, we have:

$$\mathbb{E}\left[\langle \sum_{t=1}^{T} \zeta_t, x\rangle\right] \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\|\zeta_t\|_*^2\right]} \tag{7.177}$$

Armed with these tools we are in position to prove Theorem 7.11.

Proof of Theorem 7.11. For the sake of convenience we shall present the analysis under (MB) and (MS) separately.

Analysis Under (MB)

1. For the (MB) case we have the following: By recalling Proposition 7.13 we have for all $x \in \mathcal{X}$:

$$\begin{aligned}\sum_{t=1}^{T}\langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \frac{1}{2K}\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*} \\ + \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|Q(\gamma_t Y_{t+1}) - X_{t+1/2}\|^2_{X_{t+1/2}} - \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t} \\ - \sum_{t=1}^{T} D(Q(\gamma_t Y_{t+1}), X_{t+1/2})\end{aligned} \tag{7.178}$$

Thus by applying Definition 3.3 the above becomes:

$$\begin{aligned}\sum_{t=1}^{T}\langle V_{t+1/2}, X_{t+1/2} - x\rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \frac{1}{2K}\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*} \\ - \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t}\end{aligned} \tag{7.179}$$

Now by the definition of the oracle's feedback, we have:

$$\sum_{t=1}^{T}\langle A(X_{t+1/2}), X_{t+1/2} - x\rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \frac{1}{2K}\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}$$

$$+\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle - \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t} \quad (7.180)$$

Moreover, since A is monotone we have for all $x \in \mathcal{X}$:

$$\langle A(x), X_{t+1/2} - x\rangle \leq \langle A(X_{t+1/2}), X_{t+1/2} - x\rangle \quad (7.181)$$

which in turn yields:

$$\sum_{t=1}^{T}\langle A(x), X_{t+1/2} - x\rangle \leq \frac{h(x) - \min h}{\gamma_{T+1}} + \frac{1}{2K}\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}$$
$$+\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle - \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t} \quad (7.182)$$

and so, by dividing both sides by T and exploiting convexity we have:

$$\langle A(x), \overline{X}_T - x\rangle \leq \frac{1}{T}\Bigg(\frac{h(x) - \min h}{\gamma_{T+1}} + \frac{1}{2K}\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}$$
$$+\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle - \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t}\Bigg) \quad (7.183)$$

Now, by considering a compact neighbourhood $\mathcal{C}$ of a solution and taking suprema we have:

$$\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T) \leq \frac{1}{T}\Bigg(\frac{D}{\gamma_{T+1}} + \frac{1}{2K}\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}$$
$$+\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle - \frac{K}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t}\Bigg) \quad (7.184)$$

where after taking expectations on both sides we get:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{1}{T}\Bigg(D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}\right]$$
$$+\sup_{x\in\mathcal{C}}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] - \frac{K}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\gamma_t}\|X_{t+1/2} - X_t\|^2_{X_t}\right]\Bigg) \quad (7.185)$$

Now, we shall bound from above each (RHS) term individually:

- For the term $D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]$ we have:

$$D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] = D\,\mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T}\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}}\right]$$
$$\leq D\sqrt{1 + \sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}\right]}$$

and hence by applying Lemma 7.14 we get:

$$D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \leq D\sqrt{1+C^2T} \tag{7.186}$$

with C^2 being the constant obtained in Lemma 7.14.

- For the term $\frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\right]$:

$$\begin{aligned}\frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\right] &= \frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}(\gamma_t-\gamma_{t+1})\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\right]\\ &\quad+\frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t+1}\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\right]\end{aligned} \tag{7.187}$$

Now, we have by applying Lemma 7.14:

$$\frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}(\gamma_t-\gamma_{t+1})\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\right] \leq \frac{1}{2K}C^2 \tag{7.188}$$

- For the term $\sup_{x\in\mathcal{C}}\,\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x-X_{t+1/2}\rangle\right]$ we have:

$$\begin{aligned}\sup_{x\in\mathcal{C}}\,\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x-X_{t+1/2}\rangle\right] &= \sup_{x\in\mathcal{C}}\,\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x\rangle\right]\\ &\quad-\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, X_{t+1/2}\rangle\right]\end{aligned} \tag{7.189}$$

For the second term of the expression we have:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, X_{t+1/2}\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\langle\mathbb{E}\left[U_{t+1/2}|\mathcal{F}_{t+1/2}\right], X_{t+1/2}\rangle\right]\\ &= 0\end{aligned}$$

with the last equality being obtained by the zero mean assumption for the noise U_t for $t = 1, 1/2, \ldots$ Now, the tricky part is dealing with the first term; at this point we shall apply Proposition 7.15. In particular, we have:

$$\begin{aligned}\sup_{x\in\mathcal{C}}\,\mathbb{E}\left[\sum_{t=1}^{T}\langle U_t, x\rangle\right] &= \max_{x\in\mathcal{C}}\,\mathbb{E}\left[\langle\sum_{t=1}^{T}U_t, x\rangle\right]\\ &\leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|U_{t+1/2}\|^2_*\right]}\\ &\leq \frac{D\sigma\sqrt{T}}{2}\end{aligned}$$

with the last inequality obtained by the (almost sure) boundedness of the noise.

Therefore, summarizing we get that:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{7.190}$$

and hence the first result follows.

2. Now we turn our attention towards the (MS) case. In particular, working in the same spirit as in (MB) we have: *Analysis Under (MS)*

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{1}{T}\Bigg(D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\right] \\ + \sup_{x\in\mathcal{C}}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] - \frac{K}{2}\,\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\beta^2\gamma_t}\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}\right]\Bigg)\end{aligned} \tag{7.191}$$

Now, set:

$$B_t^2 = \min\{\|A(X_t)-A(X_{t+1/2})\|^2_{X_{t+1/2},*}, \|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\} \tag{7.192}$$

and the respective auxiliary learning rate:

$$\tilde{\gamma}_t = \frac{1}{\sqrt{1+\sum_{j=1}^{t-1}B_j^2}} \tag{7.193}$$

By definition of B_t^2, we have $\frac{1}{\tilde{\gamma}_t} \leq \frac{1}{\gamma_t}$ and hence:

$$-\frac{1}{\gamma_t}\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*} \leq -\frac{1}{\tilde{\gamma}_t}B_t^2 \tag{7.194}$$

Moreover by denoting $\xi_t = [V_{t+1/2}-V_t] - [A(X_{t+1/2})-A(X_t)]$ we have:

$$\begin{aligned}\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*} &\leq 2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*} + 2\|\xi_t\|^2_{X_t,*} \\ &\leq 2\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*} + 2\mu\|\xi_t\|^2_*\end{aligned}$$

Moreover, we have:

$$\begin{aligned}\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*} \leq B_t^2 + \|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*} \\ - \min\{\|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}, \|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*}\}\end{aligned} \tag{7.195}$$

which in turn yields:

$$\begin{aligned}\|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*} &\leq B_t^2 + \max\{0, \|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*} - \|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}\} \\ &\leq B_t^2 + B_t^2 + 2\mu\|\xi_t\|^2_* \\ &= 2B_t^2 + 2\|\xi_t\|^2_* \\ &\leq B_t^2 + \max\{0, \|V_{t+1/2}-V_t\|^2_{X_{t+1/2},*} - \|A(X_{t+1/2})-A(X_t)\|^2_{X_{t+1/2},*}\} \\ &\leq B_t^2 + B_t^2 + 2\mu\|\xi_t\|^2_* \\ &= 2B_t^2 + 2\mu\|\xi_t\|^2_*\end{aligned}$$

with the last inequality being obtained by the fact that if

$$\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*} \geq \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \tag{7.196}$$

then it yields:

$$\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*} - \|A(X_{t+1/2}) - A(X_t)\|^2_{X_{t+1/2},*} \leq B_t^2 + 2\mu\|\xi_t\|_*^2 \tag{7.197}$$

Having all this at hand, we revisit (7.199). In particular, we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}\right] \leq C^2 + \mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T}\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}}\right] \tag{7.198}$$

with C^2 denoting the constant derived in Lemma 7.14. Now by combining the inequalities we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\|V_{t+1/2} - V_t\|^2_{X_{t+1/2},*}\right] &\leq C^2 + \mathbb{E}\left[\sqrt{1 + 2\sum_{t=1}^{T}B_t^2 + 2\sum_{t=1}^{T}\alpha_t^2\|\xi_t\|_*^2}\right] \\
&\leq C^2 + \mathbb{E}\left[\sqrt{1 + 2\sum_{t=1}^{T}B_t^2} + \sqrt{2\mu\sum_{t=1}^{T}\|\xi_t\|_*^2}\right] \\
&\leq C^2 + \mathbb{E}\left[\sum_{t=1}^{T}\tilde{\gamma}_t B_t^2 + 2\sqrt{2\sum_{t=1}^{T}\mu\|\xi_t\|_*^2}\right]
\end{aligned}$$

So, (7.199) becomes:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq \frac{1}{T}\Bigg(D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \sup_{x\in\mathcal{C}}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] \\
+ \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{2K}\tilde{\gamma}_t - \frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right] + \frac{1}{K}\,\mathbb{E}\left[\sqrt{2\mu\sum_{t=1}^{T}\|\xi_t\|_*^2}\right] + C^2\Bigg)
\end{aligned} \tag{7.199}$$

Now let us bound each term individually:

- For the term $D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]$ we have by Lemma 7.14:

$$D\,\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \leq D\sqrt{1 + C^2T} \tag{7.200}$$

- For the term $\sup_{x\in\mathcal{C}}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right]$ working in the same spirit as in we have:

$$\sup_{x\in\mathcal{C}}\mathbb{E}\left[\sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] \leq \frac{D\sigma\sqrt{T}}{2} \tag{7.201}$$

- For the term $\frac{1}{K}\,\mathbb{E}\left[\sqrt{2\mu\sum_{t=1}^{T}\|\xi_t\|_*^2}\right]$ we have by Jensen's inequality:

$$\frac{1}{K}\,\mathbb{E}\left[\sqrt{2\mu\sum_{t=1}^{T}\|\xi_t\|_*^2}\right] \leq \frac{1}{K}\sqrt{2\mu\sum_{t=1}^{T}\mathbb{E}\left[\|\xi_t\|_*^2\right]} \tag{7.202}$$

Moreover we have that:

$$\begin{aligned}\mathbb{E}[\|\xi_t\|_*^2] &= \mathbb{E}[\|U_{t+1/2}-U_t\|_*^2]\\ &\leq 2\,\mathbb{E}[\|\xi_{t+1/2}\|_*^2]+2\,\mathbb{E}[\|\xi_t\|_*^2]\\ &\leq 4\sigma^2\end{aligned}$$

which in turn implies:

$$\frac{1}{K}\,\mathbb{E}\left[\sqrt{2\mu\sum_{t=1}^{T}\|\xi_t\|_*^2}\right] \leq \frac{2\sigma}{K}\sqrt{2\mu T} \tag{7.203}$$

- For the term $\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{2K}\tilde{\gamma}_t-\frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right]$ we have: First we set:

$$t_0=\max\{1\leq t\leq T:\tilde{\gamma}_t\geq\frac{K}{\beta}\} \tag{7.204}$$

This yields:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{2K}\tilde{\gamma}_t-\frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right] &= \mathbb{E}\left[\sum_{t=1}^{t_0}\left(\frac{1}{2K}\tilde{\gamma}_t-\frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right]+\mathbb{E}\left[\sum_{t=t_0+1}^{T}\left(\frac{1}{2K}\tilde{\gamma}_t-\frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right]\\ &\leq \mathbb{E}\left[\sum_{t=1}^{t_0}\left(\frac{1}{2K}\tilde{\gamma}_t-\frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right]\end{aligned}$$

with the last inequality being obtained by the definition of t_0. Moreover, we have:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{2K}\tilde{\gamma}_t-\frac{K}{2\beta^2\tilde{\gamma}_t}\right)B_t^2\right] &\leq \frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{t_0}\tilde{\gamma}_t B_t^2\right]\\ &= \frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{t_0}(\tilde{\gamma}_t-\tilde{\gamma}_{t+1})B_t^2\right]+\frac{1}{2K}\,\mathbb{E}\left[\sum_{t=1}^{t_0}\tilde{\gamma}_{t+1}B_t^2\right]\\ &\leq \frac{C^2}{2K}+\frac{1}{2K}\,\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{t_0}B_t^2}\right]\\ &= \frac{C^2}{2K}+\frac{1}{2K}\,\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{t_0-1}B_t^2}\right]+\frac{1}{2K}\,\mathbb{E}\left[B_{t_0}\right]\\ &\leq \frac{C^2}{2K}+\frac{C}{2K}+\frac{1}{2K}\,\mathbb{E}\left[\frac{1}{\tilde{\gamma}_{t_0}}\right]\end{aligned}$$

$$\leq \frac{C^2}{2K} + \frac{C}{2K} + \frac{\beta}{2K^2}$$

Hence summarizing the above bounds, we conclude that:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{(X)}_T)\right] = \mathcal{O}\left(\frac{A}{T} + \frac{B\sigma}{\sqrt{T}}\right) \tag{7.205}$$

and so the result follows.

□

8

PERSPECTIVES

We conclude this thesis by providing some research perspectives for future work. These directions concern the two key optimization scenarios treated in this thesis. We shall illustrate these for each framework individually.

8.1 MINIMIZATION SETTINGS

Our theoretical analysis confirms that Mirror Descent methods concurrently achieve optimal rates of convergence in relatively continuous and relatively smooth problems, both stochastic or deterministic, constrained or unconstrained, and without requiring any prior knowledge of the problem's smoothness/continuity parameters. These appealing properties open the door to the following future research directions:

- *NoLips Acceleration:*

  One important question that remains is whether the $\mathcal{O}(1/T)$ rate can be improved to $\mathcal{O}(1/T^2)$ for relative smooth problems. Assuming boundedness, we know that this is possible in the Euclidean case: AcceleGrad and UnixGrad already achieve an accelerated rate [60, 67]. On the other hand, for problems that are h-smooth in the sense of [19], the very recent paper of [38] showed that the $\mathcal{O}(1/T)$ rate is, in general, unimprovable.

  One idea is to substitute relative smoothness by (MS) conditions. In particular, as we discussed above, one may show that metrically smooth problems are also h-smooth for a suitable choice of h, suggesting that the $\mathcal{O}(1/T)$ rate may also be optimal in this problem class. We conjecture that this indeed the case; at the same time, there is no evidence to suggest that an accelerated rate cannot be obtained for real-world singular problems like D-optimal design or PIP.

- *Adaptation between smooth and relative smooth objectives:*

  Another linked question is whether we could design a method which is able to adapt its performance optimally between the classes of smooth and relative smooth functions. More precisely, is it possible to have a method which exhibits a generic rate of $\mathcal{O}(1/T)$ for relative smooth objectives and automatically adjusts its performance to $\mathcal{O}(1/T^2)$ if the respective function is smooth in the ordinary sense. An intuitive approach is to examine whether a Bregman based variant of AcceleGrad can satisfy this type of adaptivity.

We defer both these questions to future work.

8.2 VARIATIONAL INEQUALITY SETTING

Our main goal in Chapter 7 was to design a universal, regime-agnostic first-order method for variational inequality problems with possibly unbounded domains and/or divergent operators for both deterministic and stochastic settings. By leveraging a suitable Finsler regularity framework and a compatible Bregman toolkit, adaptive Mirror-Prox/ Dual Extrapolation algorithms achieve the above desiderata, and their rates interpolates sharply between $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T)$ for metrically bounded/ stochastic and smooth problems respectively. This leaves open several questions such as:

- (VI) *to* (Opt) *adaptation:*

  As we already described throughout there is a convergence rate discrepancy between min-min and min-max problems: if the underlying operator is smooth, it is possible to achieve a $\mathcal{O}(1/T^2)$ value convergence rate for (Opt); by contrast for (VI) the best attainable rate is $\mathcal{O}(1/T)$. However, in order to apply the appropriate algorithm the optimizer should know in advance whether the associated operator is a smooth gradient field or not. Therefore, a natural open question which arises in this context is whether it is possible to design an accelerated regime-agnostic method that achieves a $\mathcal{O}(1/T)$ rate for general variational inequalities – i.e., those for which A is not a gradient field – and a $\mathcal{O}(1/T^2)$ when the underlying operator is a gradient. The key difficulty in order to establish such a method is to be able to provide simultaneously:

  1. an acceleration mechanism which is activated whenever the operator is a gradient field.

  2. An extra-gradient type template to ensure optimal rates for (VI)

  An obvious candidate seems to be UniXGrad; however it remains unclear whether it exhibits such a behavior.

- *Last iterate convergence rates for adaptive methods convergence rate for* (VI)*:*

  As we already mentioned the best attainable rate for (VI) relative to the restricted merit function is $\mathcal{O}(1/T)$; as we discussed this rate is achieved by the ergodic and/or time average of extra-gradient methods. This rises the question of what is the asymptotic behaviour of the last iterate of (MP) method, i.e., before any type of average occurs. In a recent paper, Golowich et al. [45] showed that the last iterate of (MP) is actually slower, if run with a constant step-size $\leq 1/\beta$, showing that it exhibits a rate of the order $\mathcal{O}(1/\sqrt{T})$. Furthermore, Yoon and Ryu [117] showed that $\mathcal{O}(1/T)$ can be recovered for the last iterate of a (MP) variants which incorporates a so-called anchoring mechanism. That said, their still requires a prior knowledge of the Lipschitz constant. Thus, the question of what is performance of (MP) run with an adaptive step-size remains open. More precisely, can an adaptive step-size ensure that the last iterate of (MP) exhibits optimal speed of convergence? We defer this question for future work.

BIBLIOGRAPHY

[1] Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *COLT '08: Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

[2] Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[3] Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004.

[4] Kimon Antonakopoulos and Panayotis Mertikopoulos. Adaptive first-order methods revisited: Convex minimization without lipschitz requirements. In *NeurIPS '21: Proceedings of the 35rd International Conference on Neural Information Processing Systems*, 2021.

[5] Kimon Antonakopoulos and Panayotis Mertikopoulos. Universal methods for variational inequalities with divergent operators. in preparation, 2021.

[6] Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox algorithm for variational inequalities with singular operators. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[7] Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach. In *ICLR '20: Proceedings of the 2020 International Conference on Learning Representations*, 2020.

[8] Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[9] Hedy Attouch, Jérôme Bolte, Patrick Redont, and Marc Teboulle. Singular Riemannian barrier methods and gradient-projection dynamical systems for constrained optimization. *Optimization*, 53(5-6):435–454, October 2004.

[10] Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[11] Alfred Auslender. *Optimisation: Méthodes numériques*. Masson, 1976.

[12] Alfred Auslender and Marc Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. Springer Monographs in Mathematics. Springer-Verlag, New York, NY, 2003.

[13] Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.

[14] Francis Bach and Kfir Yehuda Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *COLT '19: Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

[15] Jean-Bernarnd Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel Journal of Mathematics*, 26:137–150, 1977.

[16] Maximilian Balandat, Walid Krichene, Claire Tomlin, and Alexandre Bayen. Minimizing regret on reflexive Banach spaces and Nash equilibria in continuous zero-sum games. In *NIPS '16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[17] David Dai-Wai Bao, Shiing-Shen Chern, and Zhongmin Shen. *An Introduction to Riemann-Finsler Geometry*. Number 200 in Graduate Texts in Mathematics. Springer-Verlag, New York, NY, 2000.

[18] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

[19] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, May 2017.

[20] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

[22] Mario Bertero, Patrizia Boccacci, Gabriele Desiderà, and Giuseppe Vicidomini. Image deblurring with Poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, November 2009.

[23] Benjamin Birnbaum, Nikhil R. Devanur, and Lin Xiao. Distributed algorithms via gradient descent for Fisher markets. In *EC' 11: Proceedings of the 12th ACM Conference on Electronic Commerce*, 2011.

[24] Radu Ioan Boţ, Ernö Robert Csetnek, and Phan Tu Vuong. The forward-backward-forward method from continuous and discrete perspective for pseudo-monotone variational inequalities in Hilbert spaces. https://arxiv.org/abs/1808.08084, 2018.

[25] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[26] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

[27] Mario Bravo and Panayotis Mertikopoulos. On the robustness of learning in games with stochastically perturbed payoff observations. *Games and Economic Behavior*, 103, John Nash Memorial issue:41–66, May 2017.

[28] Mario Bravo, David S. Leslie, and Panayotis Mertikopoulos. Bandit learning in concave N-person games. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[29] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[30] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

[31] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5 (1):1–122, 2012.

[32] Yair Censor and Arnold Lent. An iterative row action method for internal convex programming. *Journal of Optimization Theory and Applications*, 34:321–353, 1981.

[33] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[34] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993.

[35] Shiing-Shen Chern. Finsler geometry is just Riemannian geometry without the quadratic restriction. *Notices of the American Mathematical Society*, 43(9):959–963, 1996.

[36] Patrick L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In Dan Butnariu, Yair Censor, and Simeon Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, pages 115–152. Elsevier, New York, NY, USA, 2001.

[37] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[38] Radu-Alexandru Dragomir, Adrien B. Taylor, Alexandre d' Aspremont, and Jérôme Bolte. Optimal complexity and certification of Bregman first-order merthods. https://arxiv.org/pdf/1911.08510.pdf, November 2019.

[39] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12: 2121–2159, 2011.

[40] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

[41] Lampros Flokas, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[42] Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *COLT '14: Proceedings of the 27th Annual Conference on Learning Theory*, 2014.

[43] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[44] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezehski, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[45] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT '20: Proceedings of the 33rd Annual Conference on Learning Theory*, 2020.

[46] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

[47] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.

[48] Filip Hanzely and Peter Richtárik. Fastest rates for stochastic mirror descent methods. https://arxiv.org/abs/1803.07374, March 2018.

[49] Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. https://arxiv.org/abs/1808.03045, 2018.

[50] Niao He, Zaid Harchaoui, Yichen Wang, and Le Song. Fast and simple optimization for Poisson likelihood models. https://arxiv.org/abs/1608.01264, 2016.

[51] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

[52] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19:*

Proceedings of the 33rd International Conference on Neural Information Processing Systems, pages 6936–6946, 2019.

[53] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[54] Chonghai Hu, Weike Pan, and James T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS' 09: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[55] Houyuan Jiang and Huifu Xu. Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Trans. Autom. Control*, 53(6):1462–1475, July 2008.

[56] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[57] S. Kakade. Lecture notes in multivariate analysis, dimensionality reduction, and spectral methods. http://stat.wharton.upenn.edu/~skakade/courses/stat991_mult/lectures/MatrixConcen.pdf, 2010.

[58] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13:1865–1890, 2012.

[59] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. https://arxiv.org/abs/1608.04636, 2016.

[60] Ali Kavis, Kfir Yehuda Levy, Francis Bach, and Volkan Cevher. UnixGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[61] Frank P. Kelly, Aman K. Maulloo, and David K. H. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252, March 1998.

[62] Leonard Kleinrock. *Queueing Systems*, volume 1: Theory. John Wiley & Sons, New York, NY, 1975.

[63] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

[64] Walid Krichene. *Continuous and discrete dynamics for online learning and convex optimization*. PhD thesis, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2016.

[65] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, June 2012.

[66] John M. Lee. *Riemannian Manifolds: an Introduction to Curvature*. Number 176 in Graduate Texts in Mathematics. Springer, 1997.

[67] Kfir Yehuda Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[68] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[69] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[70] Lennart Ljung. Strong convergence of a stochastic approximation algorithm. *Annals of Statistics*, 6(3):680–696, 1978.

[71] Haihao Lu. "Relative-continuity" for non-Lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1 (4):288–303, June 2019.

[72] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively-smooth convex optimization by first-order methods and applications. *SIAM Journal on Optimization*, 28(1): 333–354, 2018.

[73] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[74] Yura Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[75] Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 2019.

[76] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *COLT '10: Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

[77] Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, November 2016.

[78] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, January 2018.

[79] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

[80] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[81] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[82] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. https://arxiv.org/pdf/1906.01115.pdf, 2019.

[83] Arkadi Semen Nemirovski. On optimality of Krylov's information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.

[84] Arkadi Semen Nemirovski. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

[85] Arkadi Semen Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[86] Arkadi Semen Nemirovski and David Berkovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983.

[87] Arkadi Semen Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[88] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269(543-547), 1983.

[89] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[90] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[91] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[92] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[93] M. B. Nevel'son and Rafail Z. Khasminskii. *Stochastic Approximation and Recursive Estimation*. American Mathematical Society, Providence, RI, 1976.

[94] Noam Nisan, Tim Roughgarden, Éva Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[95] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 2019. URL https://doi.org/10.1007/s10107-019-01420-0.

[96] Georgios Piliouras and Jeff S. Shamma. Optimization despite chaos: Convex relaxations to complex limit sets via Poincaré recurrence. In *SODA '14: Proceedings of the 25th annual ACM-SIAM Symposium on Discrete Algorithms*, 2014.

[97] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

[98] Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987.

[99] Leonid Denisovich Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[100] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013.

[101] Herbert Robbins and David Sigmund. A convergence theorem for nonnegative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, NY, 1971.

[102] Ralph Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[103] Gesualdo Scutari, Francisco Facchinei, Daniel Pérez Palomar, and Jong-Shi Pang. Convex optimization, game theory, and variational inequality theory in multiuser communication systems. *IEEE Signal Process. Mag.*, 27(3):35–49, May 2010.

[104] Siavash Mirshams Shahshahani. *A New Mathematical Framework for the Study of Linkage and Selection*. Number 211 in Memoirs of the American Mathematical Society. American Mathematical Society, Providence, RI, 1979.

[105] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[106] Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and Fenchel duality. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.

[107] Vadim Ivanovich Shmyrev. An algorithm for finding equilibrium in the linear exchange model with fixed budgets. *Journal of Applied and Industrial Mathematics*, 3:505–518, 2009.

[108] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. MIT Press, Cambridge, MA, USA, 2012.

[109] Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Mohammad Alkousa, and Alexander Titov. Generalized mirror prox for monotone variational inequalities: Universality and inexact oracle. https://arxiv.org/abs/1806.05140, 2018.

[110] Fedor Stonyakin, Alexander Gasnikov, Alexander Tyurin, Dmitry Pasechnyuk, Artem Agafonov, Pavel Dvurechensky, Darina Dvinskikh, Alexey Kroshnin, and Victorya Piskunova. Inexact model: A framework for optimization and variational inequalities. https://arxiv.org/abs/1902.00990, 2019.

[111] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170:67–96, 2018.

[112] Rachel Ward, Xiaoxia Wu, and Léon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. In *ICML '19: Proceedings of the 36th International Conference on Machine Learning*, 2019.

[113] John Glen Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers, Part II*, volume 1, pages 325–378, 1952.

[114] Fang Wu and Li Zhang. Proportional response dynamics leads to market equilibrium. In *STOC '07: Proceedings of the 39th annual ACM symposium on the Theory of Computing*, 2007.

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

[116] Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[117] TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *ICML'21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[118] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: Better rates under weaker conditions. https://arxiv.org/abs/1303.4645, 2013.

[119] Yihan Zhou, Victor S. Portella, Mark Schmidt, and Nicholas J. A. Harvey. Regret bounds without Lipschitz continuity: Online learning with relative Lipschitz losses. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[120] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen P. Boyd, and Peter W. Glynn. Stochastic mirror descent for variationally coherent optimization problems. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[121] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen P. Boyd, and Peter W. Glynn. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.

[122] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML '03: Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

APPENDIX

A

LEMMAS ON NUMERICAL INEQUALITIES

In this appendix, we provide some necessary inequalities on numerical sequences that we require for the convergence rate analysis of the previous sections. Most of the lemmas presented below already exist in the literature, and go as far back as Auer et al. [10] and McMahan and Streeter [76]; when appropriate, we note next to each lemma the references with the statement closest to the precise version we are using in our analysis. These lemmas can also be proved by the general methodology outlined in Gaillard et al. [42, Lem. 14], so we only provide a proof for two ancillary results that would otherwise require some more menial bookkeeping.

Lemma A.1 (76, 67)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sqrt{\sum_{t=1}^{T} \alpha_t} \leq \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T} \alpha_t} \tag{A.1}$$

Lemma A.2 (67)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sum_{t=1}^{T} \frac{\alpha_t}{1 + \sum_{i=1}^{t} \alpha_i} \leq 1 + \log(1 + \sum_{t=1}^{T} \alpha_t) \tag{A.2}$$

Lemma A.3. *Let $b_1, \dots, b_t$ a sequence of non-negative numbers with $b_1 > 0$. Then, the following inequality holds:*

$$\sum_{t=1}^{T} \frac{b_t}{\sum_{i=1}^{t} b_i} \leq 2 + \log\left(\frac{\sum_{t=1}^{T} b_t}{b_1}\right) \tag{A.3}$$

Proof. It is directly obtained by applying Lemma A.2 for the sequence $\alpha_t = b_t / b_1$. □

The following set of inequalities are due to [14]. For completeness, we provide a sketch of their proof.

Lemma A.4 (14)**.** *For all non-negative numbers: $\alpha_1, \dots \alpha_t \in [0, \alpha]$, $\alpha_0 \geq 0$, the following inequality holds:*

$$\sqrt{\alpha_0 + \sum_{t=1}^{T-1} \alpha_i} - \sqrt{\alpha_0} \leq \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\alpha_0 + \sum_{i=1}^{t-1} \alpha_j}} \leq \frac{2\alpha}{\sqrt{\alpha_0}} + 3\sqrt{\alpha} + 3\sqrt{\alpha_0 + \sum_{t=1}^{T-1} \alpha_t} \tag{A.4}$$

Lemma A.5. *For all non-negative numbers: $\alpha_1, \dots \alpha_t \in [0, \alpha]$, $\alpha_0 \geq 0$, we have:*

$$\sum_{t=1}^{T} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} \leq 2 + \frac{4\alpha}{\alpha_0} + 2\log\left(1 + \sum_{t=1}^{T-1} \frac{\alpha_t}{\alpha_0}\right) \tag{A.5}$$

Proof. Let us denote

$$T_0 = \min\left\{t \in [T] : \textstyle\sum_{j=1}^{t-1} \alpha_j \geq \alpha\right\} \tag{A.6}$$

Then, dividing the sum by T_0, we get:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} &\leq \sum_{t=1}^{T_0-1} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} + \sum_{t=T_0}^{T} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} \\
&\leq \frac{1}{\alpha_0} \sum_{t=1}^{T_0-1} \alpha_t + \sum_{t=T_0}^{T} \frac{\alpha_t}{1/2\alpha_0 + 1/2\alpha + 1/2\sum_{j=1}^{t-1} \alpha_j} \\
&\leq \frac{\alpha}{\alpha_0} + 2 \sum_{t=T_0}^{T} \frac{\alpha_i/\alpha_0}{1 + \sum_{j=T_0}^{t} \alpha_j/\alpha_0} \\
&\leq \frac{2\alpha}{\alpha_0} + 2 + 2\log\left(1 + \sum_{t=T_0}^{T} \alpha_i/\alpha_0\right) \\
&\leq \frac{2\alpha}{\alpha_0} + 2 + 2\log\left(1 + \sum_{t=1}^{T} \alpha_i/\alpha_0\right)
\end{aligned} \tag{A.7}$$

where we used the fact that $\sum_{j=1}^{T_0-2} \alpha_j \leq \alpha$ as well as for all $t \geq T_0, \sum_{j=1}^{t-1} \alpha_j \geq \alpha$ (both follow from the definition of T_0) and Lemma A.2. □

COLOPHON

This manuscript was typeset with LaTeX 2ε using Hermann Zapf's Palatino type face (the actual Type 1 PostScript fonts used were URW Palladio L and FPL). The monospaced text (hyperlinks, etc.) was typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera" (with Type 1 PostScript fonts by Malte Rosenau and Ulrich Dirr).

The typographic style of this dissertation was inspired by the authoritative genius of Bringhurst's *Elements of Typographic Style*, ported to LaTeX by André Miede, the original designer of the `classicthesis` template. Any unsightly deviations from these works should be attributed solely to the author's (not always successful) efforts to conform to the awkward A4 paper size.

Adaptive Methods for Optimization without Lipschitz Requirements

Grenoble, April 12, 2022

Kimon Antonakopoulos