Adaptive First-Order Methods Revisited: Convex Optimization without Lipschitz Requirements

Kimon Antonakopoulos **Panayotis Mertikopoulos**
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG 38000 Grenoble, France
kimon.antonakopoulos@inria.fr panayotis.mertikopoulos@imag.fr

Abstract

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, motivated by a recent flurry of activity on non-Lipschitz (NoLips) optimization, we consider problems that are continuous or smooth relative to a reference Bregman function – as opposed to a global, ambient norm (Euclidean or otherwise). These conditions encompass a wide range of problems with singular objective, such as Fisher markets, Poisson tomography, D-design, and the like. 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.

1 Introduction

Owing to their wide applicability and flexibility, first-order methods continue to occupy the forefront of research in learning theory and continuous optimization. Their analysis typically revolves around two basic regularity conditions for the problem at hand: (*i*) Lipschitz continuity of the problem's objective, and / or (*ii*) 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 [30, 31, 43]. 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 classes [12, 28, 31].

This disparity in convergence rates has led to a surge of interest in adaptive methods that can seamlessly interpolate between these different regimes. Two state-of-the-art methods of this type are the ACCELEGRAD and UNIXGRAD algorithms of Levy et al. [23] and Kavis et al. [21]: both algorithms simultaneously achieve an $\mathcal{O}(1/\sqrt{T})$ value convergence rate in non-smooth problems, an $\mathcal{O}(1/T^2)$ rate in smooth problems, and an $\mathcal{O}(1/\sqrt{T})$ average rate if run with bounded, unbiased stochastic gradients (the smoothness does not affect the rate in this case). In this regard, UNIXGRAD and ACCELEGRAD both achieve a "best of all worlds" guarantee which makes them ideal as off-the-shelf solutions for applications where the problem class is not known in advance – e.g., as in online traffic routing, game theory, etc.

At the same time, there have been considerable efforts in the literature to account for problems that do not adhere to these Lipschitz regularity requirements – such as Fisher markets, quantum tomography, D-design, Poisson deconvolution / inverse problems, and many other examples [5, 8, 9, 25, 26, 40]. The reason that the Lipschitz framework fails in this case is that, even when the problem's domain is bounded, the objective function exhibits singularities at the boundary, so it cannot be Lipschitz

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

continuous or smooth. As a result, no matter how small we pick the step-size of a standard gradient method (adaptive or otherwise), the existence of domains with singular gradients can – and typically *does* – lead to catastrophic oscillations (especially in the stochastic case).

A first breakthrough in this area was provided by Birnbaum et al. [8] and, independently, Bauschke et al. [5] and Lu et al. [26], who considered a "Lipschitz-like" gradient continuity condition for problems with singularities.[1] At around the same time, Lu [25] and Teboulle [40] introduced a "relative continuity" condition which plays the same role for Lipschitz continuity. Collectively, instead of using a global norm as a metric, these conditions employ a Bregman divergence as a measure of distance, and they replace gradient descent with *mirror descent* [12, 28].

In these extended problem classes, *non-adaptive* mirror descent methods achieve an $\mathcal{O}(1/\sqrt{T})$ value convergence rate in relatively continuous problems [1, 25], an $\mathcal{O}(1/T)$ rate in relatively smooth problems [5, 8], and an $\mathcal{O}(1/\sqrt{T})$ average rate if run with stochastic gradients in relatively continuous problems [1, 25]. Importantly, the $\mathcal{O}(1/T)$ rate for relatively smooth problems *does not match* the $\mathcal{O}(1/T^2)$ rate for standard Lipschitz smooth problems: in fact, even though [19] proposed a tentative path towards faster convergence in certain non-Lipschitz problems, Dragomir et al. [16] recently established an $\Omega(1/T)$ lower bound for problems that are relatively-but-not-Lipschitz smooth.

Our contributions. Our aim in this paper is to provide an *adaptive, parameter-agnostic* method that simultaneously achieves order-optimal rates in the above "non-Lipschitz" framework. By design, the proposed method – which we call *adaptive mirror descent* (AdaMir) – has the following desirable properties:

1. When run with perfect gradients, the trajectory of queried points converges, and the method's rate of convergence in terms of function values interpolates between $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T)$ for relatively continuous and relatively smooth problems respectively.
2. When run with stochastic gradients, the method attains an $\mathcal{O}(1/\sqrt{T})$ average rate of convergence.
3. The method applies to both constrained and unconstrained problems, without requiring a finite Bregman diameter or knowledge of a compact domain containing a solution.

The enabling apparatus for these properties is an adaptive step-size policy in the spirit of [4, 17, 27]. However, a major difference – and technical difficulty – is that the relevant definitions cannot be stated in terms of global norms, because the variation of non-Lipschitz function explodes at the boundary of the problem's domain (put differently, gradients may be unbounded even over bounded domains). For this reason, our policy relies on the aggregation of a suitable sequence of "Bregman residuals" that stabilizes seamlessly when approaching a smooth solution, thus enabling the method to achieve faster convergence rates.

Related work. Beyond the references cited above, problems with singular objectives were treated in a recent series of papers in the context of online and stochastic optimization [1, 44]; however, the proposed methods are *not* adaptive, and *they do not interpolate* between different problem classes.

In the context of adaptive methods, the widely used AdaGrad algorithm of Duchi et al. [17] and McMahan and Streeter [27] was recently shown to interpolate between an $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T)$ rate of convergence [23, 24]. More precisely, Li and Orabona [24] showed that a specific, "one-lag-behind" variant of AdaGrad with prior knowledge of the smoothness modulus achieves an $\mathcal{O}(1/T)$ rate in smooth, unconstrained problems; concurrently, Levy et al. [23] obtained the same rate in a parameter-agnostic context. In both cases, AdaGrad achieves an $\mathcal{O}(1/\sqrt{T})$ rate of convergence in stochastic problems (though with somewhat different assumptions for the randomness).

In terms of rate optimality for Lipschitz smooth problems, AdaGrad is outperformed by AcceleGrad [23] and UnixGrad [21]: these methods both achieve an $\mathcal{O}(1/T^2)$ rate in Lipschitz smooth problems, and an $\mathcal{O}(1/\sqrt{T})$ in stochastic problems with bounded gradient noise. By employing an efficient line-search step, the *universal primal gradient descent* (UPGD) algorithm of Nesterov [33] achieves order-optimal guarantees in the wider class of Hölder continuous problems (which includes the Lipschitz continuous and smooth cases as extreme cases); however, UPGD does not cover stochastic problems or problems with relatively continuous / smooth objectives.

[1] This condition was first examined by Birnbaum et al. [8] in the context of Fisher markets. The analysis of Bauschke et al. [5] and Lu et al. [26] is much more general, but several ideas are already present in [8].

		Constr. / Uncon.	Stoch. (L)	Order-optimal	RC	RS	Stoch. (R)
AdaGrad	[17]	✓/✓	✓	×	×	×	×
AcceleGrad	[23]	×/ ✓	✓	✓	×	×	×
UnixGrad	[21]	✓/ ×	✓	✓	×	×	×
UPGD	[33]	✓/ ✓	×	✓	×	×	×
GMP	[39]	✓/ ✓	×	partial	×	$1/T$	×
AdaProx	[2]	✓/ ✓	×	✓	partial	partial	×
AdaMir	[ours]	✓/ ✓	✓	✓	$1/\sqrt{T}$	$1/T$	$1/\sqrt{T}$

Table 1: Overview of related work. For the purposes of this table, (L) refers to "Lipschitz" and (R) to "relative" continuity or smoothness respectively. "Order-optimal" means that the algorithm attains the best rates for the worst instance in the class it was designed for (see cited papers for the details). Logarithmic factors are ignored throughout; we also note that the $\mathcal{O}(1/T)$ rate in the RS column is, in general, unimprovable [16].

As far as we are aware, the closest works to our own are the *generalized mirror-prox* (GMP) algorithm of [39] and the AdaProx method of [1], both designed for variational inequality problems. The GMP algorithm can achieve interpolation between different classes of Hölder continuous problems and can adapt to the problem's relative smoothness modulus, but it does not otherwise interpolate between the relatively smooth and relatively continuous classes. Moreover, GMP requires knowledge of a "domain of interest" containing a solution of the problem; in this regard, it is similar to AcceleGrad [23] (though it does not require an extra projection step). The recently proposed AdaProx method of Antonakopoulos et al. [2] also achieves a similar interpolation result under a set of assumptions that are closely related – but not equivalent – to the relatively continuous/smooth setting of our paper. In any case, neither of these papers covers the stochastic case; to the best of our knowledge, AdaMir is the first method in the literature capable of adaptiving to relatively continuous / smooth objectives, even under uncertainty. For convenience, we detail these related works in Table 1 above.

2 Problem setup and preliminaries

Problem statement. Throughout the sequel, we will focus on convex minimization problems of the general form

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

In the above, $\mathcal{X}$ is a convex subset of a normed d-dimensional space $\mathcal{V} \cong \mathbb{R}^d$, and $f \colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is a proper lower semi-continuous (l.s.c.) convex function with $\operatorname{dom} f = \{x \in \mathcal{V} : f(x) < \infty\} = \mathcal{X}$. Compared to standard formulations, we stress that *$\mathcal{X}$ is not assumed to be compact, bounded, or even closed*. This lack of closedness will be an important feature for our analysis because we are interested in objectives that may develop singularities at the boundary of their domain; for a class of relevant examples of this type, see [1, 5, 7–9, 26, 44] and references therein.

To formalize our assumptions for (Opt), we will write $\partial f(x)$ for the *subdifferential* of f at x, and $\mathcal{X}_\circ \equiv \operatorname{dom} \partial f = \{x \in \mathcal{X} : \partial f(x) \neq \varnothing\}$ for the *domain of subdifferentiability* of f. Formally, elements of ∂f will be called subgradients, and we will treat them throughout as elements of the dual space $\mathcal{V}^*$ of $\mathcal{V}$. By standard results, we have $\operatorname{ri} \mathcal{X} \subseteq \mathcal{X}_\circ \subseteq \mathcal{X}$, and any solution x^* of (Opt) belongs to $\mathcal{X}_\circ$ [36, Chap. 26]; to avoid trivialities, we will make the following blanket assumption.

Assumption 1. The solution set $\mathcal{X}^* \equiv \arg\min f \subseteq \mathcal{X}_\circ$ of (Opt) is nonempty.

Two further assumptions that are standard in the literature (but which we relax in the sequel) are:

1. *Lipschitz continuity:* there exists some $G > 0$ such that

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

2. *Lipschitz smoothness:* there exists some $L > 0$ such that

$$f(x') \leq f(x) + \langle v, x' - x \rangle + \frac{L}{2}\|x' - x\|^2 \quad \text{for all } x, x' \in \mathcal{X} \text{ and all } v \in \partial f(x). \tag{LS}$$

Remark. For posterity, we note that the smoothness requirement (LS) *does not* imply that $\partial f(x)$ is a singleton. The reason for this more general definition is that we want to concurrently treat problems with smooth and non-smooth objectives, and also feasible domains that are contained in lower-dimensional subspaces of $\mathcal{V}$.[2] We also note that we will be mainly interested in cases where the above requirements all *fail* because f and/or its derivatives blow up at the boundary of $\mathcal{X}$. By this token, we will not treat (LC)/(LS) as "blanket assumptions"; we discuss this in detail in the sequel.

The oracle model. From an algorithmic point of view, we aim to solve (Opt) by using iterative methods that require access to a *stochastic first-order oracle* (SFO) [31]. 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 $g(x;\omega) \in \mathcal{V}^*$ 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:

$$g_t \equiv g(X_t;\omega_t) = \nabla f(X_t) + U_t \tag{1}$$

with $U_t \in \mathcal{V}^*$ denoting the "measurement error" relative to some selection $\nabla f(X_t)$ of $\partial f(X_t)$. 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, g_t and U_t are not. Finally, we will also make the statistical assumption that

$$\mathbb{E}[U_t \mid \mathcal{F}_t] = 0 \quad \text{and} \quad \|U_t\|_*^2 \leq \sigma^2 \quad \text{for all } t = 1, 2, \ldots \tag{SFO}$$

This assumption is standard in the analysis of parameter-agnostic adaptive methods, cf. [4, 21, 23, 41] and references therein. For concreteness, we will refer to the case $\sigma = 0$ as deterministic – since, in that case, $U_t = 0$ for all t. Otherwise, if $\liminf_t \|U_t\|_* > 0$, the noise will be called *persistent* and the model will be called *stochastic*.

3 Relative Lipschitz continuity and relative Lipschitz smoothness

3.1. Bregman functions. We now proceed to describe a flexible template extending the standard Lipschitz continuity and Lipschitz smoothness conditions – (LC) and (LS) – to functions that are possibly singular at the boundary points of $\mathcal{X}$. The main idea of this extension revolves around the *non-Lipschitz* (NoLips) framework that was first studied by Birnbaum et al. [8] and then rediscovered independently by Bauschke et al. [5] and Lu et al. [26]. The key notion in this setting is that of a suitable "reference" *Bregman function* which provides a geometry-adapted measure of divergence on $\mathcal{X}$. This is defined as follows:

Definition 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. $\operatorname{dom} \partial h \subseteq \mathcal{X} \subseteq \operatorname{dom} h$.
2. The subdifferential of h admits a continuous selection $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom} \partial h$.
3. h is strongly convex, i.e., there exists some $K > 0$ such that

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

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

$$D(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x \rangle. \tag{3}$$

Remark. The notion of a Bregman function was first introduced by Bregman [11]. Our definition follows [20, 29, 32] and leads to the smoothest presentation, but there are variant definitions where h is not necessarily assumed strongly convex, cf. [1, 13, 14] and references therein.

Some standard examples of Bregman functions are as follows:

[2] For example, the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ with $f(x_1, 0) = x_1$ and $f(x_1, x_2) = \infty$ for $x_2 \neq 0$ is perfectly smooth on its domain ($x_2 = 0$); however, $\partial f(x_1, 0) = \{(1, v_2) : v_2 \in \mathbb{R}\}$, and this set is never a singleton.

- **Euclidean regularizer:** Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^d$ 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 $x, x' \in \mathcal{X}$
- **Entropic regularizer:** Let $\mathcal{X} = \{x \in \mathbb{R}_+^d : \sum_{i=1}^d x_i = 1\}$ be the unit simplex of $\mathbb{R}^d$ 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 $x' \in \mathcal{X}$ $x \in \operatorname{ri} \mathcal{X}$.
- **Log-barrier:** Let $\mathcal{X} = \mathbb{R}_{++}^d$ denote the (open) positive orthant of $\mathbb{R}^d$. Then, the *log-barrier* on $\mathcal{X}$ is defined as $h(x) = -\sum_{i=1}^d \log x_i$ for all $x \in \mathbb{R}_{++}^d$. The corresponding divergence is known as the *Itakura-Saito divergence* and is given by $D(x, x') = \sum_{i=1}^d (x_i/x_i' - \log(x_i/x_i') - 1)$ [14].

3.2. Relative continuity. With this background in hand, we proceed to discuss how to extend the Lipschitz regularity assumptions of Section 2 to account for problems with singular objective functions. We begin with the notion of *relative continuity* (RC), as introduced by Lu [25] and extended further in a recent paper by Zhou et al. [44]:

Definition 2. A convex l.s.c. function $f \colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is said to be *relatively continuous* if there exists some $G > 0$ such that

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

In the literature, there have been several extensions of (LC) to problems with singular objectives. Below we discuss some of these variants and how they can be integrated in the setting of Definition 2.

► **Example 1** (W$[f, h]$ continuity)**.** This notion intends to single out sufficient conditions for the convergence of "proximal-like" methods like mirror descent. Specifically, following Teboulle [40], f is said to be W$[f, h]$-continuous relative to h on $\mathcal{X}$ (read: "f is weakly h-continuous") if there exists some $G > 0$ such that, for all $t > 0$, 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 $\Delta = [\langle \nabla f(x), x - x' \rangle]^2 - 2G^2 D(x', x)$, so it is immediate to check that (RC) holds.

► **Example 2** (Riemann–Lipschitz continuity)**.** Concurrently to the above, Antonakopoulos et al. [1] introduced a Riemann-Lipschitz continuity condition extending (LC) as follows. Let $\|\cdot\|_x$ be a family of local norms on $\mathcal{X}$ (possibly induced by an appropriate Riemannian metric), and let $\|v\|_{x,*} = \max_{\|x'\|_x \leq 1} \langle v, x' \rangle$ denote the corresponding dual norm. Then, f is *Riemann–Lipschitz continuous* relative to $\|\cdot\|_x$ if there exists some $G > 0$ such that:

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

As we show in the paper's supplement, (RLC) $\Longrightarrow$ (RC) so (RC) is more general in this regard.

3.3. Relative smoothness. As discussed above, the notion of *relative smoothness* (RS) was introduced by [8] and independently rediscovered by [5, 26]. It is defined as follows:

Definition 3. A convex l.s.c. function $f \colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is said to be *relatively smooth* if there exists some $L > 0$ such that

$$Lh - f \quad \text{is convex.} \tag{RS}$$

The main motivation behind this elegant definition is the following variational characterizations:

Proposition 1. *The following statements are equivalent:*

1. *f satisfies* (RS).
2. *f satisfies the inequality* $f(x) \leq f(x') + \langle \nabla f(x'), x - x' \rangle + LD(x, x')$,
3. *f satisfies the inequality* $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \leq L\left[D(x, x') + D(x', x)\right]$.

A close variant of Proposition 1 appears in [5, 8, 26], so we do not prove it here. Instead, we discuss below a different extension of (LS) that turns out to be a special case of (RS).

► **Example 3** (Metric smoothness)**.** Similar in spirit to (RLC), Antonakopoulos et al. [2] introduced an extension of (LS) that replaces the global norm $\|\cdot\|$ with a local norm $\|\cdot\|_x$, $x \in \mathcal{X}_\circ$. In particular, given such a norm, we say that f is *metrically smooth* (relative to $\|\cdot\|_x$) if

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

An observation that seems to have been overlooked by [2] is that (MS) $\Longrightarrow$ (RS), so (RS) is more general. We prove this observation in the appendix.

3.4. More examples. Some concrete examples of optimization problems satisfying (RC), (RS) or both (but not their Euclidean counterparts) are Fisher markets [8, 38], Poisson inverse problems [1, 5], support vector machines [25, 44], D-design [9, 26], etc. Because of space constraints, we do not detail these examples here; however, we provide an in-depth presentation of a Fisher market model in the appendix, along with a series of numerical experiments used to validate the analysis to come.

4 Adaptive mirror descent

We are now in a position to define the proposed *adaptive mirror descent* (ADAMIR) method. In abstract recursive form, ADAMIR follows the basic mirror descent template

$$x^+ = P_x(-\gamma g), \tag{MD}$$

where P is a generalized Bregman proximal operator induced by h (see below for the detailed definition), g is a search direction determined by a (sub)gradient of f at x, and $\gamma > 0$ is a step-size parameter. We discuss these elements in detail below, starting with the prox-mapping P.

4.1. The prox-mapping. Given a Bregman function h, its induced *prox-mapping* is defined as

$$P_x(v) = \arg\min_{x' \in \mathcal{X}} \{\langle v, x - x'\rangle + D(x', x)\} \quad \text{for all } x \in \operatorname{dom} \partial h,\ v \in \mathcal{V}^*, \tag{4}$$

where $D(x', x)$ denotes the Bregman divergence of h. Of course, in order for (4) to be well-defined, the $\arg\min$ must be attained in $\operatorname{dom} \partial h$. Indeed, we have:

Proposition 2. *The recursion* (MD) *satisfies* $x^+ \in \operatorname{dom} \partial h$ *for all* $x \in \operatorname{dom} \partial h$ *and all* $g \in \mathcal{V}^*$.

To streamline our discussion, we postpone the proof of Proposition 2 to the supplement. For now, we only note that it implies that the abstract recursion (MD) is *well-posed*, i.e., it can be iterated for all $t = 1, 2, \dots$ to generate a sequence $X_t \in \mathcal{X}$.

4.2. The method's step-size. The next important element of (MD) is the method's step-size. In the unconstrained case, a popular adaptive choice is the so-called "inverse-sum-of-squares" policy

$$\gamma_t = 1 \Big/ \sqrt{\textstyle\sum_{s=1}^t \|\nabla f(X_s)\|_*^2}, \tag{5}$$

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 (5) 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 the (5) 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* [31] that can be defined here as

$$\delta = \|x^+ - x\| \big/ \gamma. \tag{6}$$

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) 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

$$\delta = \sqrt{D(x, x^+) + D(x^+, x)}/\gamma. \tag{7}$$

Obviously, when $h(x) = (1/2)\|x\|_2^2$, we readily recover the definition of the Euclidean gradient mapping (6). 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.

4.3. The ADAMIR algorithm. We are finally in a position to put everything together and define the *adaptive mirror descent* (ADAMIR) 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

$$X_{t+1} = P_{X_t}(-\gamma_t g_t) \tag{8}$$

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

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

In the sequel, we will use the term "ADAMIR" to refer interchangeably to the update $X_t \leftarrow X_{t+1}$ and the specific step-size policy used within. The convergence properties of ADAMIR are discussed in detail in the next two sections (in both deterministic and stochastic problems); in the supplement, we also perform a numerical validation of the method in the context of a Fisher market model.

5 Deterministic analysis and results

We are now in a position to state our main convergence results for ADAMIR. 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.

5.1. Ergodic convergence and rate interpolation. We begin with the convergence of the method's "time-averaged" state, i.e., $\bar{X}_T = (1/T)\sum_{t=1}^T X_t$.

Theorem 1. *Let X_t, $t = 1, 2, \dots$, 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:*

$$f(\bar{X}_T) - \min f \leq \frac{\sqrt{2}G\big[D_1 + 8G^2/\delta_0^2 + 2\log(1 + 2G^2T/\delta_0^2)\big]}{\sqrt{T}} + \frac{3\sqrt{2}G + 4G^2/\delta_0^2}{T}. \tag{9}$$

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

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

$$f(\bar{X}_T) - \min f \leq \left[f(X_1) - \min f + \left(2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{4L^2}{\delta_0^2}\right)L\right]^2 \frac{D_1}{T}. \tag{10}$$

Theorem 1 shows that, up to logarithmic factors, ADAMIR achieves the min-max optimal bounds for functions in the *RC* $\cup$ *RS* oracle complexity class.[3] The key element of the proof (which we detail in the supplement), is the following regret bound:

Proposition 3. *With notation as in Theorem 1,* ADAMIR *enjoys the regret bound*

$$\sum_{t=1}^T [f(X_t) - f(x^*)] \leq \frac{D_1}{\gamma_T} + \frac{\sum_{t=1}^T \gamma_t^2 \delta_t^2}{\gamma_T} + \sum_{t=1}^T \gamma_t \delta_t^2. \tag{11}$$

[3] We recall here that, in contrast to (LS), the $\mathcal{O}(1/T)$ rate is optimal in (RS), cf. Dragomir et al. [16].

The proof of Proposition 3 hinges on the specific definition of AdaMir's step-size, and the exact functional form of the regret bound (11) plays a crucial role in the sequel. Specifically, under the regularity conditions (RC) and (RS), we respectively obtain the following key lemmas:

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

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

As we explain below, the boundedness estimate of Lemma 1 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 2. This property of AdaMir 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.[4]

Armed with these two lemmas, we obtain the following series of estimates:

1. Under (RC), the terms in the RHS of (11) 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 (11) can be bounded as $\mathcal{O}(1)$, so we obtain an $\mathcal{O}(1/T)$ convergence rate for $\bar{X}_T$.

For the details of these calculations (including the explicit constants and logarithmic terms that appear in the statement of Theorem 1), we refer the reader to the supplement.

5.2. 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. The formal statement is as follows.

Theorem 2. *Suppose that* f *satisfies* (RC) *or* (RS)*. Then* X_t *converges to* $\arg\min f$.

The main idea of the proof (which we detail in the appendix) consists of two steps. The first key step is to show that, under (RC) $\cup$ (RS), the iterates of AdaMir have $\liminf f(X_t) = \min f$; we show this in Proposition C.1. 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 [10, 15]. 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; we provide the relevant details in the supplement.

Non-convex objectives. We close this section with two remarks on non-convex objectives. First, Theorem 2 applies verbatim to non-convex objectives f satisfying the "secant condition" [10, 45]

$$\inf\{\langle \nabla f(x), x - x^* \rangle : x^* \in \arg\min f, x \in \mathcal{K}\} > 0 \tag{12}$$

for every closed subset $\mathcal{K}$ of $\mathcal{X}$ that is separated by neighborhoods from $\arg\min f$. In the supplement, our results have all been derived based on this more general condition (it is straightforward to verify that (SI) always holds for convex functions).

Even more generally, Lemma 2 also allows us to derive results for general non-convex problems. Indeed, the proof of Proposition 1 shows that $\min_{1\leq t\leq T} \delta_t^2 = \mathcal{O}(1/T)$ *without* requiring any properties on f other than relative smoothness. As a result, we conclude that the "best iterate" of the method – i.e., the iterate with the least residual – decays as $\mathcal{O}(1/T)$. This fact partially generalizes a similar result obtained in [24, 41] 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.

[4] In more detail, Levy et al. [23], Li and Orabona [24] and Kavis et al. [21] 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.

6 Stochastic analysis

In this last section, we focus on the stochastic case ($\sigma > 0$). Our main results here are as follows.

Theorem 3. *Let X_t, $t = 1, 2, \dots$, 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*

$$\mathbb{E}\left[f(\bar{X}_T) - f(x^*)\right] \leq (D_1 + H)\sqrt{\frac{\delta_0^2 + 2G_\sigma^2}{T}} \tag{13}$$

where $H = 8G_\sigma^2/\delta_0^2 + 2\log(1 + 2G_\sigma^2 T/\delta_0^2)$.

Finally, if (RS) kicks in, we have the sharper guarantee:

Theorem 4. *With notation as above, if f satisfies (RS),* AdaMir *enjoys the bound*

$$\mathbb{E}[f(\bar{X}_T) - f(x^*)] \leq (2 + D_1 + H)\left[\frac{A}{T} + \frac{B\sigma}{\sqrt{T}}\right] \tag{14}$$

where:

a) $A = \delta_0 + 2[f(X_1) - \min f] + L\big(2 + 8G_\sigma^2/\delta_0^2 + 2\log(4L^2/\delta_0^2)\big)$. (15a)

b) $B = \sqrt{(4 + 2H)/K}$. (15b)

The full proof of Theorems 3 and 4 is relegated to the supplement, but the key steps are as follows:

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 3 and 4 branches out. First, in the case of Theorem 3, 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 (13) then follows by the boxing bound. Instead, in the case of Theorem 4, the analysis is more involved and relies crucially on the lower bound $\gamma_t \geq 1/(A + B\sigma\sqrt{t})$. The bound (14) then follows by combining this lower bound for γ_t with the regret boxing bound above.

In the supplement, we also conclude a series of numerical experiments in random Fisher markets that illustrate the method's adaptation properties in an archetypal non-Lipschitz problem.

7 Concluding remarks

Our theoretical analysis confirms that AdaMir concurrently achieves 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 several future research directions, especially regarding the method's convergence properties in non-convex problems. The "best-iterate" discussion of Section 5 is a first step along the way, but many questions and problems remain open in this direction, especially regarding the convergence of the method's "last iterate" in stochastic, non-convex settings. We defer these questions to future work.

Acknowledgments and Disclosure of Funding

This research was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET) and 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 grant ALIAS (ANR-19-CE48-0018-01).

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] 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.

[8] 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.

[9] 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.

[10] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9): 142, 1998.

[11] 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.

[12] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

[13] Yair Censor and Arnold Lent. An iterative row action method for internal convex programming. *Journal of Optimization Theory and Applications*, 34:321–353, 1981.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. https://arxiv.org/abs/1808.03045, 2018.

[20] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Arkadi Semen Nemirovski and David Berkovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983.

[29] 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.

[30] 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.

[31] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

[32] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120 (1):221–259, 2009.

[33] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[34] Noam Nisan, Tim Roughgarden, Éva Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[35] Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987.

[36] Ralph Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[37] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[38] 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.

[39] 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.

[40] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170:67–96, 2018.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

A Bregman regularizers and mirror maps

Our goal in this appendix is to derive some basic properties for the class of Bregman proximal maps and mirror descent methods considered in the main body of our paper. Versions of the properties that we derive are known in the literature [see e.g., 6, 14, 32, 37, and references therein].

To begin, we introduce two notions that will be particularly useful in the sequel. The first is the convex conjugate of a Bregman function h, i.e.,

$$h^*(y) = \max_{x \in \operatorname{dom} h} \{\langle y, x\rangle - h(x)\} \tag{A.1}$$

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

$$Q(y) = \arg\max_{x \in \operatorname{dom} h} \{\langle y, x\rangle - h(x)\} \tag{A.2}$$

That the above is well-defined is a consequence of the fact that h is proper, l.s.c., convex and coercive;[5] in addition, the fact that Q takes values in $\operatorname{dom} \partial h$ follows from the fact that any solution of (A.2) must necessarily have nonempty subdifferential (see below). For completeness, we also recall here the definition of the Bregman proximal mapping

$$P_x(v) = \arg\min_{x' \in \operatorname{dom} h} \{\langle v, x - x'\rangle + D(x', x)\} \tag{prox}$$

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

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

Lemma A.1. *Let h be a regularizer in the sense of Definition 1 with K-strong convexity modulus. Then, for all $x \in \operatorname{dom} \partial h$ and all $v, y \in \mathcal{V}^*$ we have:*

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

$$\langle \nabla h(x), x - p\rangle \leq \langle y, x - p\rangle. \tag{A.3}$$

Proof. For the first equivalence, note that x solves (A.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 (prox) if and only if $\nabla h(x) + v \in \partial h(x^+)$ and therefore if and only if $x^+ = Q(\nabla h(x) + v)$.

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

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, t(p - x)\rangle] \quad \text{for all } t \in [0, 1]. \tag{A.4}$$

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. □

As a corollary, we have:

Proof of Proposition 2. Our claim follows directly from a tandem application of items (1) and (2) in Lemma A.1. □

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" [14]. This will be our main tool for deriving the main estimates for our analysis. Being more precise, we have the following lemma:

[5] 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 $\|\cdot\|$.

Lemma A.2. *Let h be a regularizer in the sense of Definition 1. Then, for all $p \in \operatorname{dom} h$ and all $x, x' \in \operatorname{dom} \partial h$, we have:*

$$D(p, x') = D(p, x) + D(x, x') + \langle \nabla h(x') - \nabla h(x), x - p \rangle. \tag{A.5}$$

Proof. By definition:

$$\begin{aligned} D(p, x') &= h(p) - h(x') - \langle \nabla h(x'), p - x' \rangle \\ D(p, x) &= h(p) - h(x) - \langle \nabla h(x), p - x \rangle \\ D(x, x') &= h(x) - h(x') - \langle \nabla h(x'), x - x' \rangle. \end{aligned} \tag{A.6}$$

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

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

Proposition A.1. *Let h be a regularizer in the sense of Definition 1 with strong convexity modulus $K > 0$. Fix some $p \in \operatorname{dom} h$ and let $x^+ = P_x(v)$ for some $x \in \operatorname{dom} \partial h$ and $v \in \mathcal{V}^*$. We then have:*

$$D(p, x^+) \leq D(p, x) - D(x^+, x) + \langle v, x^+ - p \rangle \tag{A.7}$$

and

$$D(p, x^+) \leq D(p, x) + D(x, x^+) - \langle v, x - p \rangle. \tag{A.8}$$

Proof. By the three-point identity established in Lemma A.2, we have:

$$D(p, x) = D(p, x^+) + D(x^+, x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p \rangle \tag{A.9}$$

Rearranging terms then yields:

$$D(p, x^+) = D(p, x) - D(x^+, x) + \langle \nabla h(x^+) - \nabla h(x), x^+ - p \rangle \tag{A.10}$$

By (A.3) and the fact that $x^+ = P_x(v)$ so $\nabla h(x) + v \in \partial h(x^+)$, the first inequality follows; the second one is obtained similarly. □

B Convergence analysis of AdaMir

In this appendix, we will illustrate in detail the convergence analysis of AdaMir, which we present in pseudocode form as Algorithm 1 below. For ease of presentation we shall divide our analysis, as in the main body of our paper, into two sections: the deterministic and the stochastic one.

Algorithm 1: Adaptive mirror descent (AdaMir)

```
Initialize X_0 ≠ X_1 ∈ dom ∂h; set δ_0 = [D(X_0, X_1) + D(X_1, X_0)]^{1/2}
for t = 1, 2, ..., T − 1 do
    set γ_t = (Σ_{s=0}^{t−1} δ_s^2)^{−1/2}                    # step
    get g_t ← g(X_t; ω_t)                                     # feedback
    set X_{t+1} = P_{X_t}(−γ_t g_t)                           # Bregman step
    set δ_t = [D(X_t, X_{t+1}) + D(X_{t+1}, X_t)]^{1/2} / γ_t  # Bregman residual
end for
return X̄_T ← (1/T) Σ_{t=1}^{T} X_t                            # candidate solution
```

B.1. The deterministic case. We begin with the proof of Lemma 1 which provides an upper bound to the Bregman residuals generated by AdaMir:

Proof of Lemma 1. By the definition of the Bregman proximal step in (MD) and Proposition A.1, we have:

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

Hence, by applying the (RC) condition of the objective we get:

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

We thus get:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq 2\gamma_t^2 G^2. \tag{B.3}$$

Hence, by the definition (7) of δ_t^2, we conclude that

$$\delta_t^2 = \frac{D(X_t, X_{t+1}) + D(X_{t+1}, X_t)}{\gamma_t^2} \leq 2G^2. \tag{B.4}$$

□

Proof of Lemma 2. Since the adaptive step-size policy γ_t is decreasing and bounded from below $(\gamma_t) \geq 0$ we get that its limit exist,i.e.,

$$\lim_{t \to +\infty} \gamma_t = \gamma_\infty \quad \text{for some} \quad \gamma_\infty \geq 0 \tag{B.5}$$

Assume that $\gamma_\infty = 0$. By Proposition 1, we obtain:

$$\begin{aligned} f(X_{t+1}) &\leq f(X_t) + \langle \nabla f(X_t), X_{t+1} - X_t \rangle + LD(X_{t+1}, X_t) \\ &\leq f(X_t) - \frac{1}{\gamma_t} D(X_t, X_{t+1}) \\ &\qquad - \frac{1}{\gamma_t} D(X_{t+1}, X_t) + L\left[D(X_t, X_{t+1}) + D(X_{t+1}, X_t)\right] \end{aligned} \tag{B.6}$$

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

$$f(X_{t+1}) \leq f(X_t) - \gamma_t \delta_t^2 + L\gamma_t^2 \delta_t^2 = f(X_t) - \frac{1}{2}\gamma_t \delta_t^2 - \frac{1}{2}\gamma_t \delta_t^2 + L\gamma_t^2 \delta_t^2 \tag{B.7}$$

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

$$\frac{1}{2}\gamma_t \delta_t^2 \leq f(X_t) - f(X_{t+1}) + \gamma_t \delta_t^2 \left[L\gamma_t - \frac{1}{2}\right] \tag{B.8}$$

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

$$\left[L\gamma_t - \frac{1}{2}\right] \leq 0 \quad \text{for all } t > t_0 \tag{B.9}$$

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

$$\begin{aligned} \frac{1}{2}\sum_{t=1}^{T} \gamma_t \delta_t^2 &\leq f(X_1) - f(X_{T+1}) + \sum_{t=1}^{t_0} \left[L\gamma_t - \frac{1}{2}\right] \gamma_t \delta_t^2 \\ &\leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \left[L\gamma_t - \frac{1}{2}\right] \gamma_t \delta_t^2 \end{aligned} \tag{B.10}$$

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

$$\frac{1}{2}\left[\frac{1}{\gamma_T} - \delta_0\right] \leq \frac{1}{2}\sqrt{\delta_0^2 + \sum_{t=1}^{T-1} \gamma_t \delta_t^2} \leq \sum_{t=1}^{T} \gamma_t \delta_t^2 \leq f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{t_0} \left[L\gamma_t - \frac{1}{2}\right] \gamma_t \delta_t^2 \tag{B.11}$$

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

$$\lim_{t \to +\infty} \gamma_t = \gamma_\infty > 0 \tag{B.12}$$

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

$$\gamma_t = \frac{1}{\sqrt{\delta_0^2 + \sum_{s=1}^{t-1} \delta_s^2}} \tag{B.13}$$

whereas after rearranging we obtain:

$$\sum_{s=1}^{t-1} \delta_s^2 = \frac{1}{\gamma_t^2} - \delta_0^2 \tag{B.14}$$

and therefore by taking limit on both sides we obtain:

$$\sum_{t=1}^{+\infty} \delta_t^2 = \lim_{t\to+\infty} \sum_{s=1}^{t-1} \delta_s^2 = \lim_{t\to+\infty} \frac{1}{\gamma_t^2} - \delta_0^2 = \frac{1}{\gamma_\infty^2} - \delta_0^2 < +\infty \tag{B.15}$$

and hence the result follows. □

We proceed by providing an upper bound in terms of the Bregman divergence for the distance of the algorithm's iterates from a solution of (Opt):

Lemma B.1. *For all $x^* \in \mathcal{X}^*$, the iterates of Algorithm 1 satisfy the bound*

$$D(x^*, X_t) \leq D(x^*, X_1) + \sum_{s=1}^{T} \gamma_s^2 \delta_s^2. \tag{B.16}$$

Proof. By the second part of Proposition A.1, we have:

$$\begin{aligned} D(x^*, X_{s+1}) &\leq D(x^*, X_s) - \gamma_t \langle g_t, X_t - x^* \rangle + D(X_s, X_{s+1}) \\ &\leq D(x^*, X_s) + D(X_s, X_{s+1}) \\ &\leq D(x^*, X_s) + D(X_{s+1}, X_s) + D(X_s, X_{s+1}) \end{aligned} \tag{B.17}$$

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

$$\begin{aligned} D(x^*, X_t) &\leq D(x^*, X_1) + \sum_{s=1}^{t} [D(X_s, X_{s+1}) + D(X_{s+1}, X_s)] \\ &\leq D(x^*, X_1) + \sum_{s=1}^{T} [D(X_s, X_{s+1}) + D(X_{s+1}, X_s)] \\ &= D(x^*, X_1) + \sum_{s=1}^{T} \gamma_s^2 \delta_s^2 \end{aligned} \tag{B.18}$$

where the last equality follows from the definition (7) of δ_t. □

With these intermediate results at our disposal, we are finally in a position to prove the core estimate (11) for AdaMir:

Proof of Proposition 3. By the convexity of f, the definition of the Bregman proximal step in Algorithm 1 and Proposition A.1, we have:

$$f(X_t) - f(x^*) \leq \langle g_t, X_t - x^* \rangle \leq \frac{1}{\gamma_t} \langle \nabla h(X_t) - \nabla h(X_{t+1}), X_t - x^* \rangle. \tag{B.19}$$

Hence, by applying again the three-point identity (Lemma A.2), we obtain:

$$\begin{aligned} f(X_t) - f(x^*) &\leq \frac{D(x^*, X_t) - D(x^*, X_{t+1})}{\gamma_t} + \frac{D(X_t, X_{t+1})}{\gamma_t} \\ &\leq \frac{D(x^*, X_t) - D(x^*, X_{t+1})}{\gamma_t} + \frac{D(X_t, X_{t+1}) + D(X_{t+1}, X_t)}{\gamma_t} \\ &= \frac{D(x^*, X_t) - D(x^*, X_{t+1})}{\gamma_t} + \gamma_t \delta_t^2 \end{aligned} \tag{B.20}$$

where the last equality follows readily from the definition (7) of δ_t. Therefore, by summing through $t = 1, 2, \dots, T$, we obtain:

$$\sum_{t=1}^{T}[f(X_t) - f(x^*)] \leq \frac{D(x^*, X_1)}{\gamma_1} + \sum_{t=2}^{T}\left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right]D(x^*, X_t) + \sum_{t=1}^{T}\gamma_t\delta_t^2. \tag{B.21}$$

Now, by Lemma B.1, the second term on the right-hand side (RHS) of (B.21) becomes:

$$\begin{aligned}\sum_{t=2}^{T}\left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right]D(x^*, X_t) &\leq \sum_{t=2}^{T}\left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right]\left(D(x^*, X_1) + \sum_{s=1}^{T}\gamma_s^2\delta_s^2\right) \\ &\leq \frac{D(x^*, X_1)}{\gamma_T} - \frac{D(x^*, X_1)}{\gamma_1} + \sum_{s=1}^{T}\gamma_s^2\delta_s^2 \cdot \sum_{t=1}^{T}\left[\frac{1}{\gamma_t} - \frac{1}{\gamma_{t-1}}\right] \\ &\leq \frac{D(x^*, X_1)}{\gamma_T} - \frac{D(x^*, X_1)}{\gamma_1} + \frac{\sum_{t=1}^{T}\gamma_t^2\delta_t^2}{\gamma_T}. \end{aligned} \tag{B.22}$$

Hence, by combining the above with (B.21), our claim follows. □

With the regret bound (11) at our disposal, we may finally proceed with the proof of our main result concerning the universality of AdaMir, i.e., Theorem 1 :

Proof of Theorem 1. Repeating the statement of Proposition 3, the iterate sequence X_t generated by AdaMir enjoys the bound:

$$\sum_{t=1}^{T}[f(X_t) - f(x^*)] \leq \frac{D(x^*, X_1)}{\gamma_T} + \frac{\sum_{t=1}^{T}\gamma_t^2\delta_t^2}{\gamma_T} + \sum_{t=1}^{T}\gamma_t\delta_t^2 \tag{11}$$

We now proceed to bound each term on the RHS of (11) from above. We consider three separate cases, first only under (RC),then under (RS) and finally when (RC) and (RS) holds.

Case 1. We begin with problems satisfying (RC).

- For the first term, Lemma 1 gives:

$$\frac{D(x^*, X_1)}{\gamma_T} = D(x^*, X_1)\sqrt{\sum_{t=0}^{T-1}\delta_t^2} \leq D(x^*, X_1)\sqrt{2G^2T}. \tag{B.23}$$

- For the second term, we have:

$$\sum_{t=1}^{T}\gamma_t^2\delta_t^2 \leq \sum_{t=1}^{T}\frac{\delta_t^2}{\sum_{s=0}^{t-1}\delta_s^2} = \sum_{t=1}^{T}\frac{\delta_t^2}{\delta_0^2 + \sum_{s=1}^{t-1}\delta_s^2}. \tag{B.24}$$

  Hence, by Lemmas 1 and D.5, we get:

$$\begin{aligned}\sum_{t=1}^{T}\gamma_t^2\delta_t^2 &\leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\left(1 + \sum_{t=1}^{T-1}\frac{\delta_t^2}{\delta_0^2}\right) \\ &= 2 + \frac{8G^2}{\delta_0^2} + 2\log\left(\sum_{t=0}^{T-1}\frac{\delta_t^2}{\delta_0^2}\right) \\ &\leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{2G^2T}{\delta_0^2}. \end{aligned} \tag{B.25}$$

- Finally, for the third term, we get:

$$\sum_{t=1}^{T}\gamma_t\delta_t^2 = \sum_{t=1}^{T}\frac{\delta_t^2}{\sqrt{\sum_{s=0}^{t-1}\delta_t^2}} = \sum_{t=1}^{T}\frac{\delta_t^2}{\sqrt{\delta_0^2 + \sum_{s=1}^{t-1}\delta_t^2}}. \tag{B.26}$$

Hence, Lemmas 1 and D.4 again yield:

$$\begin{aligned}\sum_{t=1}^{T}\gamma_t\delta_t^2 &\leq \frac{4G^2}{\delta_0} + 3\sqrt{2}G + 3\sqrt{\delta_0^2 + \sum_{t=1}^{T-1}\delta_t^2} \\ &\leq \frac{4G^2}{\delta_0} + 3\sqrt{2}G + 3\sqrt{\sum_{t=0}^{T-1}\delta_t^2} \\ &\leq \frac{4G^2}{\delta_0} + 3\sqrt{2}G + 3\sqrt{2G^2T}. \end{aligned} \tag{B.27}$$

The claim of Theorem 1 then follows by combining the above within the regret bound (11).

Case 2. We now turn to problems satisfying (RS). Recalling Lemma 2, we shall revisit the terms of (11). In particular, we have:

- For the first term, we have:

$$\frac{D(x^*, X_1)}{\gamma_T} = D(x^*, X_1)\sqrt{\sum_{t=0}^{T-1}\delta_t^2} \leq \frac{D(x^*, X_1)}{\gamma_\infty} \tag{B.28}$$

- For the second term, we have:

$$\sum_{t=1}^{T}\gamma_t^2\delta_t^2 \leq \frac{1}{\delta_0^2}\sum_{t=1}^{T}\delta_t^2 \leq \frac{1}{\delta_0^2\gamma_\infty^2} - 1 \tag{B.29}$$

- Finally, for the third term, we get:

$$\sum_{t=1}^{T}\gamma_t\delta_t^2 \leq \frac{1}{\delta_0}\sum_{t=1}^{T}\delta_t^2 \leq \frac{1}{\delta_0\gamma_\infty^2} - \delta_0 \tag{B.30}$$

Combining all the above, the result follows.

Case 3. Finally, we consider objectives where (RC) and (RS) hold simultaneously. Now, by working in the same spirit as in the proof of Lemma 2 we get:

$$\frac{1}{2}\gamma_t\delta_t^2 \leq f(X_t) - f(X_{t+1}) + \gamma_t\delta_t^2\left[L\gamma_t - \frac{1}{2}\right] \tag{B.31}$$

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

$$\frac{1}{2}\sum_{t=1}^{T}\gamma_t\delta_t^2 \leq f(X_1) - \min_{x\in\mathcal{X}} f(x) + \sum_{t=1}^{T}\gamma_t\delta_t^2\left[L\gamma_t - \frac{1}{2}\right] \tag{B.32}$$

Now, after denoting:

$$t_0 = \max\{t \in \mathbb{N} : 1 \leq t \leq T \ \text{ such that } \ \gamma_t \geq \frac{1}{2L}\} \tag{B.33}$$

and decomposing the sum we get:

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\gamma_t\delta_t^2 &\leq f(X_1) - \min_{x\in\mathcal{X}} f(x) + \sum_{t=1}^{t_0}\gamma_t\delta_t^2\left[L\gamma_t - \frac{1}{2}\right] + \sum_{t=t_0+1}^{T}\gamma_t\delta_t^2\left[L\gamma_t - \frac{1}{2}\right] \\ &\leq f(X_1) - \min_{x\in\mathcal{X}} f(x) + \sum_{t=1}^{t_0}\gamma_t\delta_t^2\left[L\gamma_t - \frac{1}{2}\right] \\ &\leq f(X_1) - \min_{x\in\mathcal{X}} f(x) + L\sum_{t=1}^{t_0}\gamma_t^2\delta_t^2 \end{aligned} \tag{B.34}$$

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

$$\begin{aligned}\sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 &\leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\left(1 + \sum_{t=1}^{t_0-1} \frac{\delta_t^2}{\delta_0^2}\right) \\ &= 2 + \frac{8G^2}{\delta_0^2} + 2\log\left(\frac{1}{\delta_0^2}\left[\delta_0^2 + \sum_{t=1}^{t_0-1} \delta_t^2\right]\right) \\ &= 2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{1}{\delta_0^2\gamma_{t_0}^2} \end{aligned} \tag{B.35}$$

and by definition of t_0 we get:

$$\sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 \leq 2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{4L^2}{\delta_0^2}. \tag{B.36}$$

which yields:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq f(X_1) - \min_{x\in\mathcal{X}} f(x) + L\left[2 + \frac{8G^2}{\delta_0^2} + 2\log\frac{4L^2}{\delta_0^2}\right] \tag{B.37}$$

The result then follows by plugging in the above bounds in (11). □

B.2. The stochastic case. In this appendix, we shall provide the stochastic part of our analysis. We start by providing an intermediate lemma concerning the class of (RC) objectives.

Lemma B.2. *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:*

$$\delta_t^2 \leq \tilde{G}^2 = \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2 \quad \textit{for all } t = 1, 2, \dots \quad \textit{almost surely} \tag{B.38}$$

Proof. By working in the same spirit, we get that:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t \langle g_t, X_t - X_{t+1}\rangle \tag{B.39}$$

and by recalling that:

$$g_t = \nabla f(X_t) + U_t \tag{B.40}$$

we get with probability 1:

$$\begin{aligned} D(X_t, X_{t+1}) + D(X_{t+1}, X_t) &\leq \gamma_t \left[\langle \nabla f(X_t), X_t - X_{t+1}\rangle + \langle U_t, X_t - X_{t+1}\rangle\right] \\ &\leq \gamma_t \left[G\sqrt{2D(X_{t+1}, X_t)} + \|U_t\|_* \|X_t - X_{t+1}\|\right] \end{aligned} \tag{B.41}$$

with the second inequality being obtained by (RC). Now, by invoking the strong convexity assumption of K, the (LHS) of the above becomes:

$$\begin{aligned} \gamma_t \left[G\sqrt{2D(X_{t+1}, X_t)} + \|U_t\|_* \|X_t - X_{t+1}\|\right] &\leq \gamma_t [G\sqrt{2(D(X_{t+1}, X_t) + D(X_t, X_{t+1}))} \\ &\quad + \|U_t\|_* \sqrt{\frac{2}{K}(D(X_{t+1}, X_t) + D(X_t, X_{t+1}))}] \end{aligned} \tag{B.42}$$

which in turn yields:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t \sqrt{D(X_{t+1}, X_t) + D(X_t, X_{t+1})} \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\|U_t\|_*\right] \tag{B.43}$$

Therefore, we get:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t^2 \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\|U_t\|_*\right]^2 \tag{B.44}$$

and by *stochastic first-order oracle* (SFO) we get with probability 1:

$$D(X_t, X_{t+1}) + D(X_{t+1}, X_t) \leq \gamma_t^2 \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2 \tag{B.45}$$

or equivalently,

$$\delta_t^2 = \frac{D(X_t, X_{t+1}) + D(X_{t+1}, X_t)}{\gamma_t^2} \leq \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2 \tag{B.46}$$

and the result follows. □

Finally, we provide the proof of the first theorem for the stochastic setting.

Proof of Theorem 3. By the second part of Proposition A.1, we have:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t\langle g_t, X_t - x^*\rangle + D(X_t, X_{t+1}) \\ &\leq D(x^*, X_t) - \gamma_t\langle g_t, X_t - x^*\rangle + D(X_{t+1}, X_t) + D(X_t, X_{t+1}) \\ &\leq D(x^*, X_t) - \gamma_t\langle g_t, X_t - x^*\rangle + \gamma_t^2\delta_t^2 \end{aligned} \tag{B.47}$$

which yields after rearranging and summing $t = 1, \dots, T$:

$$\sum_{t=1}^{T} \gamma_t\langle g_t, X_t - x^*\rangle \leq D(x^*, X_1) + \sum_{t=1}^{T} \gamma_t^2\delta_t^2 \tag{B.48}$$

and by recalling that $g_t = \nabla f(X_t) + U_t$ and taking expectations on both sides we get:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t\langle \nabla f(X_t), X_t - x^*\rangle\right] \leq D(x^*, X_1) + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t\langle U_t, X_t - x^*\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t^2\delta_t^2\right] \tag{B.49}$$

First, we shall the (LHS) from below. In particular, we have by convexity:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t\langle \nabla f(X_t), X_t - x^*\rangle\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t(f(X_t) - f(x^*))\right] \tag{B.50}$$

Moreover, by denoting $\tilde{G}^2 = \left[\sqrt{2}G + \sqrt{\frac{2}{K}}\sigma\right]^2$ we have with probability 1:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t(f(X_t) - f(x^*) &= \sum_{t=1}^{T} \frac{1}{\sqrt{\delta_0^2 + \sum_{s=1}^{t-1}\delta_s^2}}(f(X_t) - f(x^*) \\ &\geq \sum_{t=1}^{T} \frac{1}{\sqrt{\delta_0^2 + \tilde{G}^2 t}}(f(X_t) - f(x^*)) \\ &\geq \sum_{t=1}^{T} \frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)t}}(f(X_t) - f(x^*) \\ &\geq \frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)T}} \sum_{t=1}^{T}(f(X_t) - f(x^*) \end{aligned} \tag{B.51}$$

with the second inequality being obtained by Lemma B.2. Hence, we get:

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t\langle \nabla f(X_t), X_t - x^*\rangle\right] \geq \frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)T}} \mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\right] \tag{B.52}$$

We now turn our attention towards to the (LHS). In particular, we shall bound each term individually from above.

- For the term $\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\langle U_t, X_t - x^*\rangle\right]$:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\langle U_t, X_t - x^*\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\langle U_t, X_t - x^*\rangle\right] \\ &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\gamma_t\langle U_t, X_t - x^*\rangle|\mathcal{F}_t\right]\right] \\ &= \sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\,\mathbb{E}\left[\langle U_t, X_t - x^*\rangle|\mathcal{F}_t\right]\right] \\ &= \sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\langle\mathbb{E}[U_t|\mathcal{F}_t], X_t - x^*\rangle\right] = 0 \end{aligned} \tag{B.53}$$

  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 D.5 and Lemma B.2, we have with probability 1:

$$\sum_{t=1}^{T}\gamma_t^2\delta_t^2 \leq 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \sum_{t=1}^{T}\frac{\delta_t^2}{\delta_0^2}) \leq 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{K\delta_0^2}T) \tag{B.54}$$

  Therefore we get:

$$\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right] \leq 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2}T) \tag{B.55}$$

Thus, combining all the above we obtain:

$$\frac{1}{\sqrt{(\delta_0^2 + \tilde{G}^2)T}}\,\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\right] \leq D(x^*, X_1) + 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2}T) \tag{B.56}$$

and hence,

$$\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t) - f(x^*))\right] \leq \sqrt{(\delta_0^2 + \tilde{G}^2)T}\left[D(x^*, X_1) + 2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2})T)\right] \tag{B.57}$$

The result follows by dividing both sides by T. □

Proof of Theorem 4. By Proposition 1, we have:

$$\begin{aligned} f(X_{t+1}) &\leq f(X_t) + \langle\nabla f(X_t), X_{t+1} - X_t\rangle + LD(X_{t+1}, X_t) \\ &\leq f(X_t) + \langle\nabla f(X_t), X_{t+1} - X_t\rangle + L\left[D(X_{t+1}, X_t) + D(X_t, X_{t+1})\right] \\ &= f(X_t) + \langle g_t, X_{t+1} - X_t\rangle + \langle U_t, X_t - X_{t+1}\rangle + L\gamma_t^2\delta_t^2 \\ &\leq f(X_t) - \frac{1}{\gamma_t}\left[D(X_{t+1}, X_t) + D(X_t, X_{t+1})\right] + \|U_t\|_*\|X_t - X_{t+1}\| + L\gamma_t^2\delta_t^2 \\ &= f(X_t) - \gamma_t\delta_t^2 + \|U_t\|_*\|X_t - X_{t+1}\| + L\gamma_t^2\delta_t^2 \end{aligned} \tag{B.58}$$

Now, since h is $K-$ strongly convex we have that:

$$\|X_t - X_{t+1}\| \leq \sqrt{\frac{2}{K}\left[D(X_{t+1}, X_t) + D(X_t, X_{t+1})\right]} = \sqrt{\frac{2}{K}}\gamma_t\delta_t \tag{B.59}$$

and using the fact that the noise $\|U_t\|_* \leq \sigma$ almost surely, we have:

$$f(X_{t+1}) \leq f(X_t) - \gamma_t\delta_t^2 + \sqrt{\frac{2}{K}}\gamma_t\delta_t^2 + L\gamma_t^2\delta_t^2 \tag{B.60}$$

Therefore, after rearranging and telescoping we get:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq 2 \left[f(X_1) - \min_{x \in \mathcal{X}} f(x) + \sum_{t=1}^{T} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2}) + \sigma \sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t \right] \tag{B.61}$$

Now, let us bound each term of the (RHS) of the above individually:

- For the term $\sum_{t=1}^{T} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2})$ we first set:

$$t_0 = \max\{1 \leq t \leq T : \gamma_t \geq \frac{1}{2L}\} \tag{B.62}$$

  Then, by decomposing the said sum we get:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2}) &= \sum_{t=1}^{t_0} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2}) + \sum_{t=t_0+1}^{T} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2}) \\ &\leq \sum_{t=1}^{t_0} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2}) \\ &\leq L \sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 \end{aligned} \tag{B.63}$$

  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 B.2) we have:

$$\begin{aligned} L \sum_{t=1}^{t_0} \gamma_t^2 \delta_t^2 &= L \sum_{t=1}^{t_0} \frac{\delta_t^2}{\delta_0^2 + \sum_{s=1}^{t-1} \delta_s^2} \\ &\leq L \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{1}{\delta_0^2} \sum_{t=1}^{t_0-1} \delta_t^2) \right] \\ &\leq L \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{1}{\delta_0^2} (\delta_0^2 + \sum_{t=1}^{t_0-1} \delta_t^2) \right] \\ &\leq L \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{1}{\delta_0^2 \gamma_{t_0}^2} \right] \end{aligned} \tag{B.64}$$

  Therefore, by the definition of t_0 we finally get with probability 1:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 (L\gamma_t - \frac{1}{2}) \leq L \left[2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log \frac{4L^2}{\delta_0^2} \right] \tag{B.65}$$

- For the term $\sigma \sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t$ we have:

$$\sigma \sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t = \sigma \sqrt{\frac{2}{K}} \sum_{t=1}^{T} \sqrt{\gamma_t^2 \delta_t^2} \leq \sigma \sqrt{\frac{2}{K}} \sqrt{T} \sqrt{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2} \tag{B.66}$$

  Therefore, by working in the same spirit as above we get:

$$\begin{aligned} \sigma \sqrt{\frac{2}{K}} \sum_{t=1}^{T} \gamma_t \delta_t &\leq \sigma \sqrt{\frac{2}{K}} \sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{1}{\delta_0^2} \sum_{t=1}^{T} \delta_t^2)} \\ &\leq \sigma \sqrt{\frac{2}{K}} \sqrt{T} \sqrt{2 + \frac{4\tilde{G}^2}{\delta_0^2} + 2\log(1 + \frac{\tilde{G}^2}{\delta_0^2} T)} \end{aligned} \tag{B.67}$$

On the other hand, we may the (LHS) from below as follows:

$$\sum_{t=1}^{T}\gamma_t\delta_t^2 \geq \gamma_T\sum_{t=1}^{T}\delta_t^2 \geq \gamma_T\left[\delta_0^2-\delta_0^2+\sum_{t=1}^{T}\delta_t^2\right] = \frac{\gamma_T}{\gamma_{T+1}^2}-\delta_0^2\gamma_T = \frac{1}{\gamma_T}-\delta_0^2\gamma_T \tag{B.68}$$

So, combining the above:

$$\frac{1}{\gamma_T}-\delta_0^2\gamma_T \leq 2(f(X_1)-\min_{x\in\mathcal{X}}f(x)+L\left[2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log\frac{4L^2}{\delta_0^2}\right] + \sigma\sqrt{\frac{2}{K}}\sqrt{T}\sqrt{2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log(1+\frac{\tilde{G}^2}{\delta_0^2}T))} \tag{B.69}$$

which finally yields with probability 1:

$$\frac{1}{\gamma_T} \leq \delta_0+2(f(X_1)-\min_{x\in\mathcal{X}}f(x)+L\left[2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log\frac{4L^2}{\delta_0^2}\right]+\sigma\sqrt{\frac{2}{K}}\sqrt{T}\sqrt{2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log(1+\frac{\tilde{G}^2}{\delta_0^2}T))} \tag{B.70}$$

and hence with probability 1:

$$\gamma_T \geq \left[\delta_0+2(f(X_1)-\min_{x\in\mathcal{X}}f(x)+L\left[2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log\frac{4L^2}{\delta_0^2}\right]+\sigma\sqrt{\frac{2}{K}}\sqrt{T}\sqrt{2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log(1+\frac{\tilde{G}^2}{\delta_0^2}T))}\right]^{-1}$$

□

Therefore, by setting:

$$A = \delta_0+2(f(X_1)-\min_{x\in\mathcal{X}}f(x)+L\left[2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log\frac{4L^2}{\delta_0^2}\right] \tag{B.71}$$

and

$$B = \sigma\sqrt{\frac{2}{K}}\sqrt{2+\frac{4\tilde{G}^2}{\delta_0^2}+2\log(1+\frac{\tilde{G}^2}{\delta_0^2}T))} \tag{B.72}$$

we get that:

$$\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t)-f(x^*))\gamma_T\right] \geq \left(A+B\sqrt{T}\right)^{-1}\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t)-f(x^*))\right] \tag{B.73}$$

Moreover, working in the same spirit as in Theorem 3 we have:

$$\left(A+B\sqrt{T}\right)^{-1}\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t)-f(x^*))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}(f(X_t)-f(x^*))\gamma_T\right] \leq \left(D_1+\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right]\right) \tag{B.74}$$

which in turn yields:

$$\mathbb{E}\left[\sum_{t=1}^{T}(f(X_t)-f(x^*))\right] \leq \left(D_1+\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right]\right)\left(A+B\sqrt{T}\right) \tag{B.75}$$

The result then follows by dividing both sides by T and by the fact that $\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t^2\delta_t^2\right] = \mathcal{O}(\log T)$.

C Last iterate Convergence

Throughout this section we assume that f satisfies the following weak-secant inequality of the form:

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

for every closed subset $\mathcal{K}$ of $\mathcal{X}$ that is separated by neighborhoods from $\arg\min f$. More precisely, our proof is divided in two parts. To begin with, we first show that under (RC) or (RS) the iterates ofADAMIR possess a convergent subsequence towards the solution set $\mathcal{X}^*$. Formally stated, we have the following proposition:

Proposition C.1. *Assume that f is (RC) or (RS) 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 ADAMIR 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 neighbourhoods 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, following the spirit of Lemma B.1, we have:

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

with the last equality being obtained by the definition of (7). Now, applying (SI) we get:

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

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 $t = 1, \dots, T$, factorizing and setting $\beta_t = \sum_{t=1}^T \gamma_t$ we have:

$$D(x^*, X_{T+1}) \leq D(x^*, X_1) - \beta_t \left[\delta(\mathcal{K}) - \frac{\sum_{t=1}^T \gamma_t^2 Z_t^2}{\beta_t} \right] \tag{C.3}$$

Now, (C.3) 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.

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

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

$$\beta_T = \sum_{t=1}^T \frac{1}{\sqrt{\delta_0^2 + \sum_{j=1}^{t-1} \delta_t^2}} \geq \sum_{t=1}^T \frac{1}{\sqrt{\delta_0^2 + 2G^2 t}} \tag{C.4}$$

  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:

$$\begin{aligned} \sum_{t=1}^T \gamma_t^2 \delta_t^2 &= \sum_{t=1}^T \frac{\delta_t^2}{\delta_0^2 + \sum_{j=1}^{t-1} \delta_j^2 / \delta_0^2} \\ &\leq 2 + 8G^2/\delta_0^2 + 2\log(1 + \sum_{t=1}^{T-1} \delta_t^2/\delta_0^2) \\ &\leq 2 + 8G^2/\delta_0^2 + 2\log(1 + 2G^2 T/\delta_0^2) \end{aligned} \tag{C.5}$$

  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:

$$\frac{\sum_{t=1}^T \gamma_t^2 \delta_t^2}{\beta_T} \to 0 \tag{C.6}$$

So, combining all the above and letting $T \to +\infty$ in (C.3), we get that $D(x^*, X_{T+1}) \to -\infty$, a contradiction. Therefore, the result under (RC) follows.

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

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq \sum_{t=1}^{+\infty} \delta_t^2 < +\infty \tag{C.7}$$

which by working as in Lemma 2 also yields:

$$\lim_{t \to +\infty} \gamma_t = \gamma_\infty > 0 \tag{C.8}$$

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 (C.3). More precisely, we have:

- For β_T we have:

$$\beta_T = \sum_{t=1}^{T} \gamma_t \geq \gamma_\infty \sum_{t=1}^{T} 1 = \gamma_\infty T \tag{C.9}$$

  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 2:

$$\sum_{t=1}^{T} \gamma_t \delta_t^2 \leq \frac{1}{\delta_0} \sum_{t=1}^{T} \delta_t^2 < +\infty \tag{C.10}$$

  which yields that $\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 = \mathcal{O}(1)$, which combined with (C.9) gives that:

$$\frac{\sum_{t=1}^{T} \gamma_t^2 \delta_t^2}{\beta_T} \to 0 \tag{C.11}$$

so, again combing the above and letting $T \to +\infty$ in (C.3), we get that $D(x^*, X_{T+1}) \to -\infty$, a contradiction. Therefore, the result follows also under (RS). □

Having all this at hand, we are finally in the position to prove 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 [15], [35]).

Lemma C.1. *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 $t = 1, 2, \ldots$:*

$$\alpha_{t+1} \leq \chi \alpha_t - \beta_t + \varepsilon_t \tag{C.12}$$

Then, α_t converges.

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

$$\alpha_{t+1} \leq \chi^{t+1} \alpha_0 + \sum_{k=0}^{t} \chi^{t-k} \varepsilon_k \tag{C.13}$$

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:

$$0 \leq \alpha_t \leq \alpha_{k_{t_0}} + \sum_{m > t_{k_{t_0}}} \varepsilon_m < \frac{\delta}{2} + \alpha + \frac{\delta}{2} = \alpha + \delta \tag{C.14}$$

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

Proof of Theorem 2. We will divide our proof in two parts by distinguishing the two different regularity cases.

Case 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:

1. $\gamma_\infty > 0$: Following the same reasoning with Lemma 2 we get that:

$$\sum_{t=1}^{T} \gamma_t^2 \delta_t^2 \leq \sum_{t=1}^{+\infty} \delta_t^2 < +\infty \tag{C.15}$$

Hence, by recalling the inequality:

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) + \gamma_t^2 \delta_t^2 \text{ for all } x^* \in \mathcal{X}^* \tag{C.16}$$

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

$$\inf_{x^* \in \mathcal{X}^*} D(x^*, X_{t+1}) \leq \inf_{x^* \in \mathcal{X}^*} D(x^*, X_t) + \gamma_t^2 \delta_t^2 \tag{C.17}$$

and since the sequence $\gamma_t^2 \delta_t^2$ is summable we can directly apply Lemma C.1 which yields that the sequence $\inf_{x^* \in \mathcal{X}^*} D(x^*, X_t)$ is convergent. Now, since by Proposition C.1,ADAMIR possesses a convergent subsequence towards the solution set $\mathcal{X}^*$ the result follows.

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

$$D_\varepsilon = \{x \in \mathcal{X} : D(\mathcal{X}^*, x) < \varepsilon\}. \tag{C.18}$$

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

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \delta_t^2 \\ &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \frac{2G^2}{K} \end{aligned} \tag{C.19}$$

with the second inequality being obtained by Lemma 1. 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:

$$\langle f(x), x - x^* \rangle \geq c > 0 \text{ for some } c \equiv c(\varepsilon) > 0 \text{ and for all } x \in D_\varepsilon \setminus D_{\varepsilon/2} \text{ and } x^* \in \mathcal{X}^* \tag{C.20}$$

We consider two cases:

- $X_t \in D_\varepsilon \setminus D_{\varepsilon/2}$: In. this case, we have:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \frac{2G^2}{K} \\ &\leq D(x^*, X_t) - \gamma_t c + \gamma_t^2 \frac{2G^2}{K} \end{aligned} \tag{C.21}$$

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:

$$\begin{aligned} D(x^*, X_{t+1}) &\leq D(x^*, X_t) - \gamma_t \langle \nabla f(X_t), X_t - x^* \rangle + \gamma_t^2 \frac{2G^2}{K} \\ &\leq D(x^*, X_t) + \gamma_t^2 \frac{2G^2}{K} \end{aligned} \tag{C.22}$$

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:

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) + \frac{\varepsilon}{2} \tag{C.23}$$

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 C.1 and the fact that $\gamma_t \to 0$.

Case 2: The (RS) case. Recall that we have the following inequality,

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) + \gamma_t^2 \delta_t^2 \quad \text{for all} \quad x^* \in \mathcal{X}^* \tag{C.24}$$

whereas taking infima on both sides relative to $\mathcal{X}^*$ we readily get:

$$\inf_{x^* \in \mathcal{X}^*} D(x^*, X_{t+1}) \leq \inf_{x^* \in \mathcal{X}^*} D(x^*, X_t) + \gamma_t^2 \delta_t^2 \tag{C.25}$$

Now, by recalling that by Lemma 2, we have $\gamma_t^2 \delta_t^2$ is summable. we can apply directly Lemma C.1. Thus, we have the sequence $\inf_{x^* \in \mathcal{X}^*} D(x^*, X_t)$ is convergent. Moreover, Proposition C.1 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:

$$\inf_{x^* \in \mathcal{X}^*} \|x^* - X_t\|^2 \leq \inf_{x^* \in \mathcal{X}^*} D(x^*, X_t) \to 0 \tag{C.26}$$

and the proof is complete. □

D Lemmas on numerical sequences

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. [3] and McMahan and Streeter [27]; 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. [18, Lem. 14], so we only provide a proof for two ancillary results that would otherwise require some more menial bookkeeping.

Lemma D.1 (27, 23)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sqrt{\sum_{t=1}^{T} \alpha_t} \leq \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T} \alpha_t} \tag{D.1}$$

Lemma D.2 (23)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sum_{t=1}^{T} \frac{\alpha_t}{1 + \sum_{i=1}^{t} \alpha_i} \leq 1 + \log(1 + \sum_{t=1}^{T} \alpha_t) \tag{D.2}$$

Lemma D.3. *Let $b_1, \dots, b_t$ a sequence of non-negative numbers with $b_1 > 0$. Then, the following inequality holds:*

$$\sum_{t=1}^{T} \frac{b_t}{\sum_{i=1}^{t} b_i} \leq 2 + \log\left(\frac{\sum_{t=1}^{T} b_t}{b_1}\right) \tag{D.3}$$

Proof. It is directly obtained by applying Lemma D.2 for the sequence $\alpha_t = b_t / b_1$. □

The following set of inequalities are due to [4]. For completeness, we provide a sketch of their proof.

Lemma D.4 (4)**.** *For all non-negative numbers: $\alpha_1, \dots \alpha_t \in [0, \alpha]$, $\alpha_0 \geq 0$, the following inequality holds:*

$$\sqrt{\alpha_0 + \sum_{t=1}^{T-1} \alpha_i} - \sqrt{\alpha_0} \leq \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\alpha_0 + \sum_{i=1}^{t-1} \alpha_j}} \leq \frac{2\alpha}{\sqrt{\alpha_0}} + 3\sqrt{\alpha} + 3\sqrt{\alpha_0 + \sum_{t=1}^{T-1} \alpha_t} \tag{D.4}$$

Lemma D.5. *For all non-negative numbers: $\alpha_1, \dots \alpha_t \in [0, \alpha]$, $\alpha_0 \geq 0$, we have:*

$$\sum_{t=1}^{T} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} \leq 2 + \frac{4\alpha}{\alpha_0} + 2\log\left(1 + \sum_{t=1}^{T-1} \frac{\alpha_t}{\alpha_0}\right) \tag{D.5}$$

Proof. Let us denote

$$T_0 = \min\{t \in [T] : \textstyle\sum_{j=1}^{t-1} \alpha_j \geq \alpha\} \tag{D.6}$$

Then, dividing the sum by T_0, we get:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} &\leq \sum_{t=1}^{T_0-1} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} + \sum_{t=T_0}^{T} \frac{\alpha_t}{\alpha_0 + \sum_{i=1}^{t-1} \alpha_i} \\
&\leq \frac{1}{\alpha_0} \sum_{t=1}^{T_0-1} \alpha_t + \sum_{t=T_0}^{T} \frac{\alpha_t}{1/2\alpha_0 + 1/2\alpha + 1/2 \sum_{j=1}^{t-1} \alpha_j} \\
&\leq \frac{\alpha}{\alpha_0} + 2 \sum_{t=T_0}^{T} \frac{\alpha_i/\alpha_0}{1 + \sum_{j=T_0}^{t} \alpha_j/\alpha_0} \\
&\leq \frac{2\alpha}{\alpha_0} + 2 + 2\log\left(1 + \sum_{t=T_0}^{T} \alpha_i/\alpha_0\right) \\
&\leq \frac{2\alpha}{\alpha_0} + 2 + 2\log\left(1 + \sum_{t=1}^{T} \alpha_i/\alpha_0\right)
\end{aligned} \tag{D.7}$$

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 D.2. □

E Fisher markets: A case study

E.1. The Fisher market model. In this appendix, we illustrate the convergence properties of AdaMir in a Fisher equilibrium problem with linear utilities – both stochastic and deterministic. Following [34], a Fisher market consists of a set $\mathcal{N} = \{1, \dots, n\}$ of n *buyers* – or *players* – that seek to share a set $\mathcal{M} = \{1, \dots, m\}$ of m 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* [22]: each player $i = 1, \dots, n$ bids x_{ik} per unit of the k-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 $\sum_{k=1}^{m} x_{ik} \leq 1$ for all $i = 1, \dots, n$. The price of the k-th good is then set to be the sum of the players' bids, i.e., $p_k = \sum_{i \in \mathcal{N}} x_{ik}$; then, each player gets a prorated fraction of each good, namely $w_{ik} = x_{ik}/p_k$.

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

$$u_i(x_i; x_{-i}) = \sum_{k \in \mathcal{M}} \theta_{ik} w_{ik} = \sum_{k \in \mathcal{M}} \frac{\theta_{ik} x_{ik}}{\sum_{j \in \mathcal{N}} x_{jk}}, \tag{E.1}$$

where $x_i = (x_{ik})_{k \in \mathcal{M}}$ 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

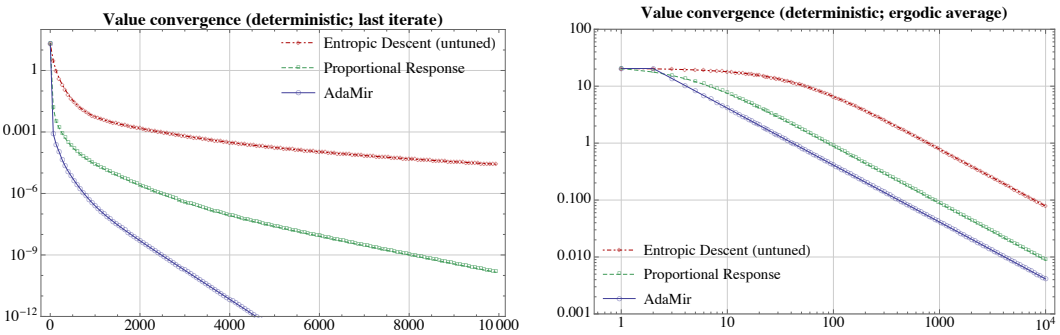
$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \tag{Eq}$$

for all $i \in \mathcal{N}$ and all $x_i = (x_{ik})_{k \in \mathcal{M}}$ such that $x_{ik} \geq 0$ and $\sum_{k \in \mathcal{M}} x_{ik} = 1$.[6]

As was observed by Shmyrev [38], the equilibrium problem (Eq) can be rewritten equivalently as

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

[6] It is trivial to see that, in this market problem, all users would saturate their budget constraints at equilibrium, i.e., $\sum_{k \in \mathcal{M}} x_{ik} = 1$ for all $i \in \mathcal{N}$.



(a) Last-iterate convergence

(b) Ergodic convergence

Figure 1: The convergence speed of (EGD), (PR) and ADAMIR in a stationary Fisher market.

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

$$f(x) = \mathbb{E}[F(x;\omega)]. \tag{E.2}$$

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. [8], the problem satisfies (RS) over $\mathcal{X} = \{x \in \mathbb{R}^{nm}_{+} : \sum_{k\in\mathcal{M}} x_{ik} = 1\}$ relative to the negative entropy function $h(x) = \sum_{ik} x_{ik} \log x_{ik}$. As a result, mirror descent methods based on this Bregman function are natural candidates for solving (E.2).

In more detail, following standard arguments [6], the general mirror descent template (MD) relative to h can be written as

$$x^{+}_{ik} = \frac{x_{ik} \exp(-\gamma g_{ik})}{\sum_{l\in\mathcal{M}} x_{il} \exp(-\gamma g_{il})} \tag{E.3}$$

where the (stochastic) gradient vector $g \equiv g(x;\theta)$ is given in components by

$$g_{ik} = 1 + \log p_k - \log \theta_{ik}. \tag{E.4}$$

Explicitly, this leads to the entropic gradient descent algorithm

$$X_{ik,t+1} = \frac{X_{ik,t}(\theta_{ik}/p_k)^{\gamma_t}}{\sum_{l\in\mathcal{M}} X_{il,t}(\theta_{il}/p_l)^{\gamma_t}} \tag{EGD}$$

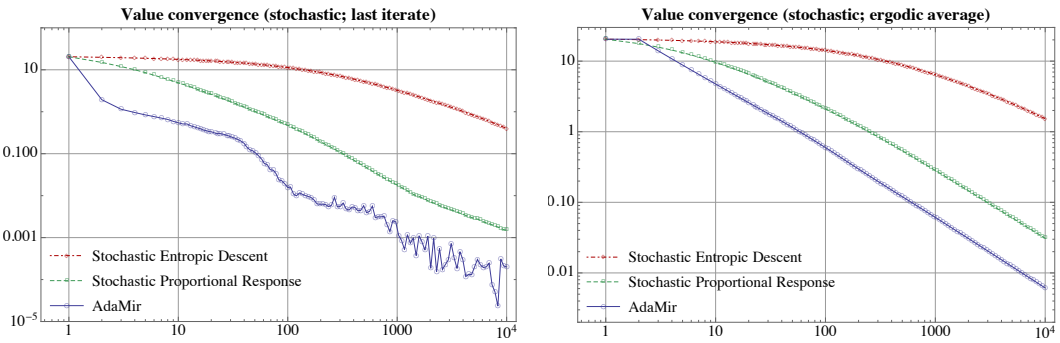
In particular, as a special case, the choice $\gamma = 1$ gives the *proportional response* (PR) algorithm of Wu and Zhang [42], namely

$$X_{ik,t+1} = \frac{\theta_{ik} w_{ik,t}}{\sum_{l\in\mathcal{M}} \theta_{il} w_{il,t}}, \tag{PR}$$

where $w_{ik,t} = X_{ik,t} / \sum_{j\in\mathcal{N}} X_{jk,t}$. As far as we aware, the PR algorithm is considered to be the most efficient method for solving *deterministic* Fisher equilibrium problems [8].

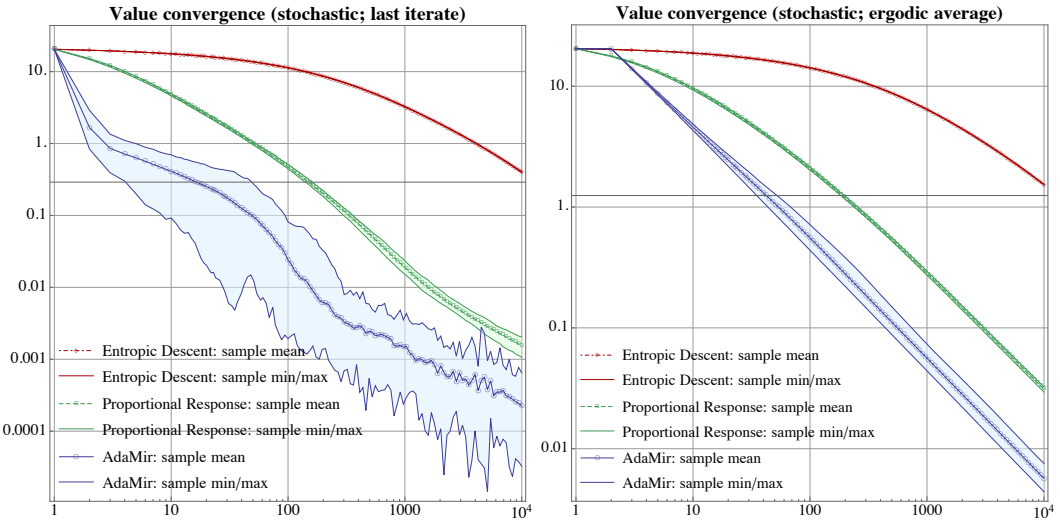
E.2. Experimental validation and methodology. For validation purposes, we ran a series of numerical experiments on a synthetic Fisher market model with $n = 50$ players sharing $m = 5$ goods, and utilities drawn uniformly at random from the interval $[2, 8]$. For stationary markets, the players' marginal 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



(a) Last-iterate convergence **(b)** Ergodic convergence

Figure 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 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.

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 [25] and Antonakopoulos et al. [1], 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. 1–3. 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. 1. For stochastic market models, we present a sample realization in Fig. 2, and a statistical study over $S = 50$ sample realizations in Fig. 3. In all cases, ADAMIR 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.