

Accelerated optimization algorithms and ordinary differential equations: the convex non Euclidean case

Paul Dobson

P.DOBSON_1@HW.AC.UK

*Maxwell Institute for Mathematical Sciences and Mathematics Department,
Heriot-Watt University,
Edinburgh, EH14 4AS, UK*

Jesus María Sanz-Serna

JMSANZSERNA@GMAIL.COM

*Departamento de Matemáticas,
Universidad Carlos III de Madrid,
Avenida Universidad 30, 28911, Leganés, Madrid*

Konstantinos C. Zygalakis

K.ZYGALAKIS@ED.AC.UK

*Maxwell Institute for Mathematical Sciences and School of Mathematics
University of Edinburgh
Peter Guthrie Tait Rd, EH9 3FD, Edinburgh*

Abstract

We study the connections between ordinary differential equations and optimization algorithms in a non-Euclidean setting. We propose a novel accelerated algorithm for minimising convex functions over a convex constrained set. This algorithm is a natural generalization of Nesterov’s accelerated gradient descent method to the non-Euclidean setting and can be interpreted as an additive Runge-Kutta algorithm. The algorithm can also be derived as a numerical discretization of the ODE appearing in Krichene et al. (2015a). We use Lyapunov functions to establish convergence rates for the ODE and show that the discretizations considered achieve acceleration beyond the setting studied in Krichene et al. (2015a). Finally, we discuss how the proposed algorithm connects to various equations and algorithms in the literature.

Keywords: Lyapunov function, probability simplex, convex optimization, mirror map, gradient descent, accelerated methods

1 Introduction

Optimization lies at the heart of many problems in statistics and machine learning. We are interested in solving the following problem

$$\min_{x \in \mathcal{X}} f(x) \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is closed and convex and the objective function f is convex and continuously differentiable in an open set that contains \mathcal{X} . Numerous different algorithms (Polyak, 1987; Boyd and Vandenberghe, 2004; Nocedal and Wright, 2006; Beck, 2017) have been proposed for this problem both in the case where $\mathcal{X} = \mathbb{R}^d$ as well as when \mathcal{X} is a proper convex subset of \mathbb{R}^d . These different algorithms can be classified depending on the type of information they use. For example, first-order methods make use only of the gradient of f , ∇f . Second-order methods use additionally second derivatives utilising in some shape or form the Hessian of f , $\nabla^2 f$ and this enables faster convergence.

In the last few years, first-order methods have gained in popularity despite their slower convergence rate since data sets and problems have become larger (Wright and Recht, 2022). In the unconstrained case, the simplest first-order method is gradient descent which however does not give optimal convergence rates within the class of (strongly) convex and gradient Lipschitz functions (Nesterov, 2014). On the other hand, Nesterov proposed a family of accelerated first-order methods with optimal convergence rates. In the constrained case, one popular algorithm for its simplicity is projected gradient descent (Bubeck, 2015), which is the composition of a gradient descent step and a projection to \mathcal{X} . On the other hand mirror gradient descent (Beck, 2017), which is an adaptation of gradient descent with the Euclidean distance replaced by a Bregman divergence, avoids calculating such projection. Nevertheless, both methods fail to yield optimal convergence rates for convex and gradient Lipschitz functions. Two examples of accelerated methods in this constrained setting are the methods proposed in Krichene et al. (2015a) and Tseng (2008).

In a different direction in the last few years, there has been a renewed interest in connecting optimization algorithms with ordinary differential equations (ODEs). This has been partially driven by the desire to understand better the acceleration phenomenon starting with the seminal paper of Su et al. (2016) that showed in the convex case Nesterov’s accelerated method corresponds to a discretization of a particular second-order ODE. This result sparked a lot of subsequent research on the links between optimization and numerical solutions of ODEs (Scieur et al., 2016; Wilson et al., 2021) as well as borrowing ideas from dynamical systems and control theory to prove convergence of optimization algorithms (Wilson et al., 2021; Lessard et al., 2016; Fazlyab et al., 2018).

An important question is why certain discretizations of second-order ODEs like the one that appeared in Su et al. (2016) may or may not accelerate. The papers (Shi et al., 2022, 2019) explain the behaviour of different algorithms using the high-resolution ODEs framework. Following a different direction Sanz Serna and Zygalkakis (2021) uses Lyapunov functions combined with integral quadratic constraints (Lessard et al., 2016; Fazlyab et al., 2018) to provide sufficient conditions that a numerical discretization should satisfy to yield acceleration. Furthermore, these ideas are extended in Dobson et al. (2023) to obtain sharper convergence rates for the Nesterov’s accelerated method in the strongly convex case as well as giving an interpretation of it as an additive Runge-Kutta method (Cooper and Sayfy, 1980, 1983).

The majority of the literature proposes and studies these second-order ODEs in the Euclidean setting. One exception is Krichene et al. (2015a) which generalises the ODE from Su et al. (2016) in a non-Euclidean setting using appropriate Bregman divergences. In addition, it proposes a discretization of this ODE that uses a suitable regularising function and inherits the favourable properties of the ODE leading to an accelerated optimization algorithm. An alternative way to generalise the ODE from Su et al. (2016) is given in Wilson et al. (2021), which is analysed using Langrangians that contain appropriate Bregman divergences.

In this work, we provide a novel discretization of the ODE appearing in Krichene et al. (2015a) leading to an accelerated optimization method for the problem (1) without requiring the use of a regularisation function. The proposed method is the natural extension of Nesterov’s method in the non-Euclidean setting and corresponds to an additive Runge Kutta discretization of the underlying non-Euclidean ODE. Furthermore, we extend the analysis from Krichene et al. (2015a) both for the ODE dynamics as well as their discretizations to be able to deal with scenarios not previously covered, for example in the case of simplex when the minimizer is on the boundary. Finally, we make explicit connections between the various non-Euclidean ODEs (Krichene et al., 2015a; Wilson et al., 2021) and optimization algorithms (Tseng, 2008).

The rest of the paper is organised as follows. In Section 2 we revisit the problem when $\mathcal{X} = \mathbb{R}^d$ and discuss the connection between ODEs and optimization algorithms. We then in Section 3 discuss in detail a natural generalization of gradient flow and gradient descent in the non-Euclidean setting. Section 4 contains our main results. In particular, after summarising the results in Krichene et al. (2015a), we propose our novel algorithm and explain why it generalises Nesterov’s method in the non-Euclidean setting as well as its connection with additive Runge-Kutta methods. Furthermore, we establish convergence for our algorithm in the setting proposed in Krichene et al. (2015a) and extend these results in a more general setting. This section concludes by discussing transformations of our algorithm for the different ODEs discussed previously in the literature. Several numerical experiments are presented in 5 that illustrate the behaviour of our proposed method as well as compare it with the method in Krichene et al. (2015a). Finally, Section 6 gathers the proofs of theorems from earlier sections.

2 ODEs and optimization in the Euclidean setting

In this preparatory section we consider the particular case where in (1) $\mathcal{X} = \mathbb{R}^d$ and \mathbb{R}^d is endowed with the standard Euclidean norm.

2.1 Gradient flow and gradient descent

The simplest ODE associated with the problem (1) is given by the gradient flow

$$\dot{x}(t) = -\nabla f(x(t)). \quad (2)$$

When (1) has a minimizer x^* , $f(x(t)) - f(x^*)$ approaches 0 at a rate $\mathcal{O}(1/t)$, as it may be proved e.g. by means of the Lyapunov function

$$V(x, t) = t(f(x) - f(x^*)) + \frac{1}{2} \|x - x^*\|^2. \quad (3)$$

In fact, since V is nonincreasing along solutions of (2), one has

$$f(x(t)) - f(x^*) \leq \frac{1}{t} V(x(0), 0) = \frac{1}{2t} \|x(0) - x^*\|^2, \quad t > 0.$$

The simplest method to discretize (2) is the explicit Euler rule, that leads to the standard gradient descent algorithm

$$x_{k+1} = x_k - h \nabla f(x_k), \quad (4)$$

where h is the timestep/learning rate. Since $x_k \approx x(kh)$ one would expect that $f(x_k) - f(x^*)$ would decay like $1/k$. In fact, when f is L_f -smooth, i.e.

$$\forall x, y \in \mathbb{R}^d, \quad \|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|,$$

and $h \leq 1/L_f$, the $\mathcal{O}(1/k)$ decay may be proved by using the discrete Lyapunov function

$$V_k(x) = kh(f(x) - f(x^*)) + \frac{1}{2} \|x - x^*\|^2,$$

which is the obvious counterpart of (3). A more sophisticated Lyapunov function valid for $1/L_f \leq h < 2/L_f$ may be seen in Fazlyab et al. (2018).

There is no need to derive (4) by seeing it as an integrator for an ODE. The algorithm may be written as

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2h} \|x - x_k\|^2 \right\}, \quad (5)$$

with a clear optimization interpretation: $f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$ is the linear approximation to the objective function and $(1/(2h)) \|x - x_k\|^2$ represents a penalty term.

2.2 Acceleration

Algorithms with acceleration for convex objective functions, including the celebrated Nesterov method (Nesterov, 2014), raise to $\mathcal{O}(1/k^2)$ the $\mathcal{O}(1/k)$ rate of convergence of gradient descent. Even though no system of ODEs was originally used to derive Nesterov's algorithm, the well-known reference Su et al. (2016) showed the relation between the algorithm and the second-order ODE

$$\ddot{x}(t) + \frac{r+1}{t} \dot{x}(t) + \nabla f(x(t)) = 0,$$

where $r \geq 2$. For our purposes, it is useful to rewrite the equation as a first order system

$$\dot{z}(t) = -\frac{t}{r} \nabla f(x(t)), \quad (6a)$$

$$\dot{x}(t) = \frac{r}{t} (z(t) - x(t)), \quad (6b)$$

with Lyapunov function

$$V(x, z, t) = \frac{t^2}{r^2} (f(x) - f(x^*)) + \frac{1}{2} \|z - x^*\|^2,$$

which implies the following $\mathcal{O}(1/t^2)$ decay estimate of $f(x(t))$ towards the optimal value $f(x^*)$

$$f(x(t)) - f(x^*) \leq \frac{r^2}{t^2} V(x(0), z(0), 0) = \frac{r^2}{2t^2} \|z(0) - x^*\|^2, \quad t > 0.$$

Following the line of thought in the preceding subsection, one would expect that integrators $(z_k, x_k) \mapsto (z_{k+1}, x_{k+1})$ for (6), where (z_k, x_k) approximate $(z(k\delta), x(k\delta))$ (δ is the time-step), may offer the potential of yielding optimization algorithms for which $f(x_k) - f(x^*)$ decays like $1/k^2$, i.e. algorithms that show acceleration. Unlike the case of the gradient flow, this is not a simple task since standard discretizations such a Runge-Kutta algorithms do not lead to acceleration, see the discussion in Sanz Serna and Zygalkis (2021) and Dobson et al. (2023). In addition, even for discretizations with acceleration, a Lyapunov function of the ODE, may not work for the discrete algorithm.

As proved in a more general setting in Section 4.2, if $\{\gamma_k\}_{k=0}^\infty$ is a sequence with $\gamma_0 = 1$, $\gamma_k \geq 1$, $k = 1, 2, \dots$, the algorithm

$$y_k = x_k + \frac{1}{\gamma_k} (z_k - x_k), \quad (7a)$$

$$z_{k+1} = z_k - \gamma_k h \nabla f(y_k), \quad (7b)$$

$$x_{k+1} = y_k + \frac{1}{\gamma_k} (z_{k+1} - z_k), \quad (7c)$$

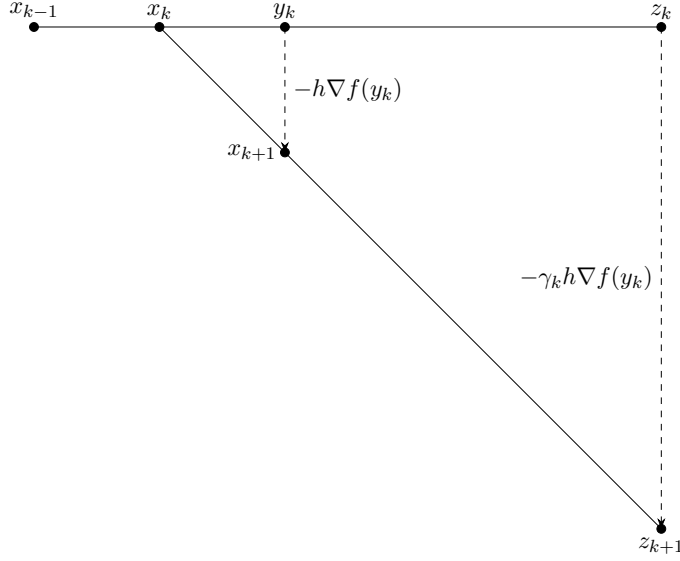


Figure 1: An illustration of one step of the Nesterov algorithm in a Euclidean setting

provides, for suitable choices of the learning rate h and the coefficients γ_k , a consistent, albeit nonstandard, discretization of (6) that leads to acceleration. Figure 1 illustrates one step of (7). The point y_k is determined as a convex combination of x_k and z_k , then z_{k+1} is obtained by moving from z_k in the direction of the gradient and finally x_{k+1} is obtained as a convex combination of x_k and z_{k+1} .

The algorithm may be reformulated by eliminating the variable z . Using (7b) and (7c), we obtain

$$x_{k+1} = y_k - h\nabla f(y_k) \quad (8)$$

and, after setting $z_k = x_k + (\gamma_{k-1} - 1)(x_k - x_{k-1})$, (7a) becomes, for $k \geq 1$,

$$y_k = x_k + \beta_{k-1}(x_k - x_{k-1}), \quad \beta_{k-1} = (\gamma_{k-1} - 1)/\gamma_k, \quad (9)$$

($y_0 = x_0$). In the formulation (8)–(9), one first computes y_k by extrapolation from x_{k-1} and x_k and then moves from y_k to x_{k+1} by a gradient descent substep. The relations (8)–(9) are the well-known formulas for the accelerated Nesterov method (Nesterov, 2014) when written as a three-term recursion linking x_{k-1} , x_k and x_{k+1} as in e.g. Fazlyab et al. (2018), with

$$\gamma_k = \frac{1}{2} \left(1 + \sqrt{1 + 4\gamma_{k-1}^2} \right), \quad k = 1, 2, \dots, \quad \gamma_0 = 1. \quad (10)$$

3 Non-Euclidean optimization

It is well known that, in many instances (see e.g. the discussion in (Beck, 2017, Example 9.19)), it is useful to consider the problem (1) when the norm in \mathbb{R}^d is not Euclidean. In what follows, \mathbb{E} denotes \mathbb{R}^d endowed with an arbitrary norm $\|\cdot\|$ and \mathbb{E}^* is the dual space with dual norm $\|\cdot\|_*$. The map $\langle \cdot, \cdot \rangle : \mathbb{E}^* \times \mathbb{E} \rightarrow \mathbb{R}$ denotes the standard pairing between \mathbb{E}^* and \mathbb{E} , i.e. the real number $\langle \zeta, x \rangle = \sum_j \zeta_j x_j$ is the value of the linear form $\zeta \in \mathbb{E}^*$ acting on the vector $x \in \mathbb{E}$.

3.1 Mirror gradient ODE and mirror descent

In the non Euclidean setting, the gradient flow equation (2) is meaningless because $\dot{x}(t) \in \mathbb{E}$ is a primal vector and $\nabla f(x(t)) \in \mathbb{E}^*$ is a dual vector. Nemirovsky and Yudin (Nemirovsky and Yudin, 1983) suggested alternative ODEs of the form

$$\dot{\zeta}(t) = -\nabla f(\chi(\zeta(t))), \quad x(t) = \chi(\zeta(t)), \quad (11)$$

where ζ takes values in \mathbb{E}^* and χ maps \mathbb{E}^* into \mathbb{E} . The dynamics take place in the dual space and the primal variable x just *mirrors* the behaviour of ζ ; for this reason, χ is referred to as the *mirror map*. In Nemirovsky and Yudin (1983), it is assumed that $\chi = \nabla\psi^*$, where $\psi^* : \mathbb{E}^* \rightarrow \mathbb{R}$ is a differentiable function which is used to construct Lyapunov functions for (11). There is much freedom in the choice of ψ^* (Nemirovsky and Yudin, 1983); throughout this paper the attention is restricted to cases where the following standing requirement is met.

Assumption 1 *The mirror map satisfies $\chi = \nabla\psi^*$ for some convex and differentiable function $\psi^* : \mathbb{E}^* \rightarrow \mathbb{R}$ and takes values in \mathcal{X} . In addition, it is L_χ -smooth i.e. for each $\xi, \zeta \in \mathbb{E}^*$:*

$$\|\chi(\xi) - \chi(\zeta)\| \leq L_\chi \|\xi - \zeta\|_*.$$

If D_{ψ^*} denotes the corresponding Bregman divergence, so that for $\xi, \zeta \in \mathbb{E}^*$,

$$D_{\psi^*}(\xi, \zeta) = \psi^*(\xi) - \psi^*(\zeta) - \langle \xi - \zeta, \nabla\psi^*(\zeta) \rangle,$$

and ζ^* is such that $x^* = \chi(\zeta^*)$ is a minimizer of (1), it is easy to prove (see e.g. Krichene et al. (2015a)) that

$$\frac{d}{dt} D_{\psi^*}(\zeta(t), \zeta^*) \leq -(f(x(t)) - f(x^*)) \leq 0, \quad (12)$$

a fact that may be used to establish convergence (Nemirovsky and Yudin, 1983).

Some examples follow.

- *Unconstrained Euclidean case.* Here $\mathcal{X} = \mathbb{E}$, the norms $\|\cdot\|$ and $\|\cdot\|_*$ are Euclidean, and $\psi^*(\cdot) = (1/2)\|\cdot\|^2$. In this case, the spaces \mathbb{E} and \mathbb{E}^* may be identified with one another, χ is the identity, $D_{\psi^*}(\xi, \zeta) = (1/2)\|\xi - \zeta\|^2$ for each ξ and ζ , and the pairing $\langle \cdot, \cdot \rangle$ may be identified with the Euclidean inner product. The ODE (11) reduces to the gradient flow equation (2)
- *The simplex.* Here \mathcal{X} is the probability simplex

$$\Delta = \left\{ x = (x_1, \dots, x_d) \in \mathbb{R}^d : \sum_{j=1}^d x_j = 1, \quad x_j \geq 0, \quad j = 1, \dots, d \right\}.$$

Choosing $\psi^*(\zeta) = \log \sum_j e^{\zeta_j}$, the i -component of $\chi(\zeta)$ equals $e^{\zeta_i} / \sum_j e^{\zeta_j}$; clearly χ maps \mathbb{E}^* onto the *relative interior* of \mathcal{X} (Beck, 2017), i.e. the subset Δ_+ of Δ consisting of points with positive components. The mirror map is 1-smooth (Beck, 2017, Example 5.15) when $\|\cdot\|_*$ is either the ℓ^∞ or the ℓ^2 norm (in which case $\|\cdot\|$ is respectively the ℓ^1 or the ℓ^2 norm).

- *The hypercube.* $\mathcal{X} = [0, 1]^d$. One may choose $\psi^*(\zeta) = \sum_j \log(e^{\zeta_j} + 1)$, so that the i -th component of $\chi(\zeta)$ is $e^{\zeta_i} / (e^{\zeta_i} + 1)$. Now χ is a diffeomorphism of \mathbb{E}^* onto the interior $(0, 1)^d$ of \mathcal{X} . If $\|\cdot\|$ is any of the ℓ^p norms, $p \in [1, \infty]$, it is easy to check that $L_\chi = 1/4$.

The Euler discretization of (11) yields the following mirror descent algorithm

$$\zeta_{k+1} = \zeta_k - h\nabla f(\chi(\zeta_k)), \quad x_{k+1} = \chi(\zeta_{k+1}). \quad (13)$$

As in the ODE, the primal variable x just mirrors the evolution of the dual variable ζ .

3.2 Writing the mirror flow ODE and mirror descent in the primal space

We now write the ODE (11) in terms of the primal variable x . If χ is differentiable with Jacobian χ' , (11) implies:

$$\dot{x}(t) = -\chi'(\zeta(t))\nabla f(x(t)). \quad (14)$$

This equation still contains ζ ; in order to eliminate ζ , we have to demand that the mirror map, that we recall it is assumed throughout to satisfy Assumption 1, has additional properties.

Assumption 2 *The mirror map $\chi = \nabla\psi^*$ is differentiable and its image is the relative interior $\text{ri}(\mathcal{X})$. In addition, ψ^* is the convex conjugate of a function $\psi = \phi + \delta_{\mathcal{X}}$, where $\phi : \mathbb{E} \rightarrow (-\infty, \infty]$ is proper, convex, differentiable over $\text{ri}(\mathcal{X})$ and satisfies $\mathcal{X} \subseteq \text{dom}(\phi)$.*

This additional assumption holds in the three examples just considered. In the Euclidean setting, $\phi(\cdot) = (1/2)\|\cdot\|^2$. For the simplex, ϕ is given by

$$\phi(x) = \sum_{j=1}^n x_j \log x_j \quad (15)$$

(negative entropy) if x is in the nonnegative orthant, with $\phi(x) = \infty$ else. For the hypercube,

$$\phi(x) = \sum_{j=1}^n (x_j \log x_j + (1 - x_j) \log(1 - x_j)),$$

(negative bit entropy) if $x \in [0, 1]^d$, with $\phi(x) = \infty$ else.

We next present a lemma that we shall use repeatedly. Some notation is required. The symbols \mathcal{A} , \mathcal{V} will denote respectively the affine hull of \mathcal{X} and the corresponding linear subspace:

$$\begin{aligned} \mathcal{V} &= \text{span}\{x - z : x, z \in \mathcal{X}\}, \\ \mathcal{A} &= \mathcal{X} + \mathcal{V}. \end{aligned}$$

For instance, for the simplex, \mathcal{A} has the equation $\sum_j x_j = 1$ and \mathcal{V} consists of all vectors with $\sum_j x_j = 0$. The notation \mathcal{N} refers to the vector subspace of \mathbb{E}^* orthogonal to $\mathcal{V} \subseteq E$; thus $\zeta \in \mathcal{N}$ if and only if $\langle \zeta, x \rangle = 0$ for each $x \in \mathcal{V}$, or, equivalently, if and only if $\langle \zeta, x_1 - x_2 \rangle = 0$ for all x_1, x_2 in \mathcal{A} . For the simplex, \mathcal{N} is spanned by the vector $\mathbf{1}$ whose entries are all equal to 1 (see the left panel in Figure 2). For the Euclidean case and the hypercube, $\mathcal{V} = E$ and therefore $\mathcal{N} = \{0\}$.

Lemma 1 *If Assumptions 1 and 2 hold, then:*

1. *As z varies in $\text{ri}(\mathcal{X})$, the inverse images by the mirror map $\chi^{-1}(z) = \{\zeta : \chi(\zeta) = z\}$ provide a partition of \mathbb{E}^* into pairwise disjoint sets.*

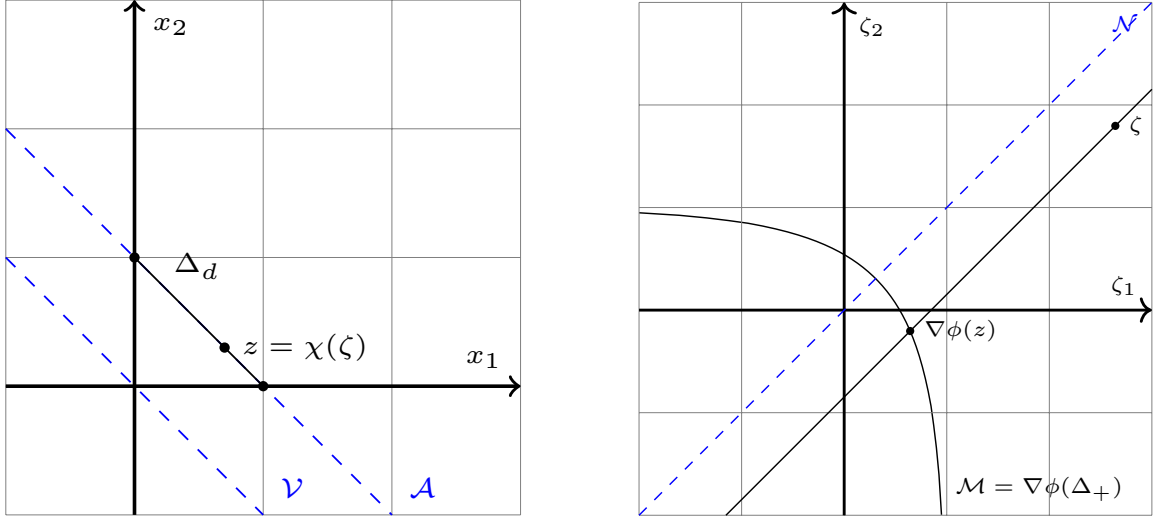


Figure 2: Lemma 1 for the case of the simplex. The left and right panels correspond to the primal and dual spaces. Each point $\zeta \in \mathbb{E}^*$ is mapped by χ into a point $z = \chi(\zeta) \in \text{ri}(\mathcal{X}) = \Delta_+$; the image $\nabla\phi(z)$ does not in general coincide with ζ , but ζ and $\nabla\phi(z)$ differ in an element in \mathcal{N} . The straight lines of slope 1 partition the dual space; each line is mapped into a single point by χ .

2. For each $z \in \text{ri}(\mathcal{X})$, its inverse image $\chi^{-1}(z)$ is the affine set $\nabla\phi(z) + \mathcal{N}$. In particular, $\nabla\phi(z) \in \chi^{-1}(z)$, i.e. $\chi(\nabla\phi(z)) = z$.
3. If $\zeta \in \mathbb{E}^*$, then $\nabla\phi(\chi(\zeta)) - \zeta \in \mathcal{N}$. Therefore for each x_1, x_2 in the affine hull \mathcal{A} of \mathcal{X} ,

$$\langle \nabla\phi(\chi(\zeta)) - \zeta, x_1 - x_2 \rangle = 0.$$

4. If $\zeta \in \mathbb{E}^*$ and $\eta \in \mathcal{N}$, then

$$\chi(\zeta) = \chi(\zeta + \eta), \quad \chi'(\zeta) = \chi'(\zeta + \eta).$$

Proof The proof can be found in Section 6. ■

The lemma shows that there is a bijection between points $z \in \text{ri}(\mathcal{X})$ and sets $\chi^{-1}(z) = \nabla\phi(z) + \mathcal{N} \subseteq \mathbb{E}^*$. There are two essentially different scenarios:

1. $\mathcal{N} = \{0\}$. Each set $\nabla\phi(z) + \mathcal{N}$ consists of the single point $\nabla\phi(z)$. In other words, the mapping $\nabla\phi$ restricted to $\text{ri}(\mathcal{X})$ is the inverse of χ and there is a one-to-one correspondence $z = \chi(\zeta), \zeta = \nabla\phi(z)$, between points $z \in \text{ri}(\mathcal{X})$ and points $\zeta \in \mathbb{E}^*$.
2. $\mathcal{N} \neq \{0\}$. In this case, for each $z \in \text{ri}(\mathcal{X})$, the affine set $\nabla\phi(z) + \mathcal{N}$ has dimension $\dim(\mathcal{N}) > 0$. The point $\nabla\phi(z)$ is in the set $\nabla\phi(z) + \mathcal{N}$ but it does not belong to any set $\nabla\phi(z') + \mathcal{N}$ if $z' \neq z$. The mapping $z \in \text{ri}(\mathcal{X}) \rightarrow \nabla\phi(z)$ provides a one-to-one parameterization of a manifold \mathcal{M} of dimension $\dim(\mathbb{E}^*) - \dim(\mathcal{N})$. See Figure 2 for the simplex with $d = 2$, where \mathcal{M} is a curve.

We may now rewrite (11) in the primal space when the additional Assumption 2 holds. We consider two situations.

- *The case $\mathcal{N} = \{0\}$.* The one-to-one correspondence between $x = \chi(\zeta)$ and $\zeta = \nabla\phi(x)$ may be used to express (11) as

$$\frac{d}{dt}\nabla\phi(x) = -\nabla f(x). \quad (16)$$

When \mathcal{N} is not reduced to $\{0\}$, this differential equation is meaningless because the left hand-side is constrained to be a vector tangent to the manifold \mathcal{M} and $\nabla f(x)$ is not constrained in that way (refer to the right panel in Figure 2).

- *General \mathcal{N} .* By Part 3 in the Lemma, $\eta := \nabla\phi(z) - \zeta \in \mathcal{N}$. Then by Part 4, $\chi'(\zeta) = \chi'(\nabla\phi(z))$ and from (14)

$$\dot{x} = -\chi'(\nabla\phi(x))\nabla f(x). \quad (17)$$

In the particular case with $\mathcal{N} = \{0\}$, this reduces to (16), because $\chi'(\nabla\phi(x))$ is the inverse of $(\nabla\phi(x))'$, as it is seen by differentiating $\chi(\nabla\phi(x)) = x$.

Example 1 *In the case of the simplex, straightforward differentiation of the expression for χ shows that (17) reads:*

$$\dot{x} = D(x)\left(-\nabla f(x) + \langle -\nabla f(x), x \rangle \mathbf{1}\right), \quad (18)$$

where $D(x)$ is the diagonal matrix with entries x_i . For x in the relative interior, this matrix is the inverse Jacobian of the map $\nabla\phi$, whose components are $1 + \log x_i$ (i.e. the inverse of the Hessian of ϕ). The right-hand side of (18) is of the form $D(x)v$, where the vector $v = \nabla f(x) + \langle \nabla f(x), x \rangle \mathbf{1}$ is a linear combination of $\nabla f(x)$ and $\mathbf{1}$, with a coefficient $\langle \nabla f(x), x \rangle$ that ensures that v is tangent at $\nabla\phi(x)$ to the manifold \mathcal{M} . Multiplying v by the inverse Jacobian results in a vector $D(x)v$ that is tangent to the relative interior of the simplex, the inverse image of \mathcal{M} by $\nabla\phi$. Analytically this corresponds to the fact that, in (18), $(d/dt)\sum_j x_j = 0$ so that this differential equation preserves the constraint $\sum_j x_j = 1$. Furthermore, due to the presence of $D(x)$, the i -th component of the right hand-side of (18) vanishes if $x_i = 0$. Therefore the points on the hyperplanes $x_i = 0$ in \mathbb{E} are equilibria of the differential equation, so that, as t varies, the $x_i(t)$ will remain positive if $x(0)$ is in the relative interior.

Similarly, when Assumption 2 holds, it is possible to rewrite the discretization (13) purely in terms of primal variables. For the case $\mathcal{N} = \{0\}$, we have

$$\nabla\phi(x_{k+1}) = \nabla\phi(x_k) - h\nabla f(x_k), \quad (19)$$

that coincides with the Euler discretization of (16). Once $\nabla\phi(x_{k+1})$ has been computed by this formula, x_{k+1} is retrieved as $\chi(\nabla\phi(x_{k+1}))$.

For general, \mathcal{N} we note that

$$x_{k+1} = \chi(\zeta_k - h\nabla f(x_k));$$

Part 3 in the Lemma implies $\eta = \nabla\phi(x_k) - \zeta_k \in \mathcal{N}$, and then, by Part 4, we have the following well-known formulation (see (Beck, 2017, Remark 9.6)) of the mirror descent algorithm:

$$x_{k+1} = \chi(\nabla\phi(x_k) - h\nabla f(x_k)). \quad (20)$$

As it may have been expected, this is a consistent discretization of the differential equation (17), as it may be seen by Taylor expansion of the right hand-side. When $\mathcal{N} = \{0\}$, the application of $\nabla\phi$ to (20) yields (19).

The algorithm (20) may be formulated from an optimization point of view, without using ODEs as stepping stones. Such a formulation is based on the the (primal) Bregman divergence D_ϕ associated with ϕ :

$$D_\phi(x, z) = \phi(x) - \phi(z) - \langle \nabla\phi(z), x - z \rangle.$$

For example in the case of the simplex, D_ϕ is the *Kullback-Leiber divergence*. Since ϕ is finite in \mathcal{X} and differentiable over $\text{ri}(\mathcal{X})$, $D_\phi(x, z)$ is defined at least for $x \in \mathcal{X}$ and $z \in \text{ri}(\mathcal{X})$. It is well known (see (Beck, 2017, Remark 9.6)) that (20) is equivalent to

$$x_{k+1} = \arg \min_{x \in \mathcal{X}} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{h} D_\phi(x, x_k) \right\}.$$

which is the direct non Euclidean counterpart of the gradient descent formula (5).

4 Non-Euclidean accelerated ODEs and numerical schemes

Mirror descent may only achieve a $\mathcal{O}(1/k)$ rate of convergence (Beck, 2017), something to be expected from the fact that in the Euclidean setting reduces to gradient descent. We now consider ODEs and algorithms that may provide rates $\mathcal{O}(1/t^2)$ or $\mathcal{O}(1/k^2)$ in non Euclidean scenarios.

4.1 A Primal/Dual ODE and a discretization

In order to construct optimization algorithms, Krichene et al. (2015a) considered the system

$$\dot{\zeta}(t) = -\frac{t}{r} \nabla f(x(t)), \tag{21a}$$

$$\dot{x}(t) = \frac{r}{t} (\chi(\zeta(t)) - x(t)), \tag{21b}$$

where $r > 0$ is a parameter. The variable x takes values in \mathbb{E} and ζ takes values in \mathbb{E}^* ; the paper (Krichene et al., 2015a) proves that if the initial data $(\zeta_0, x_0) \in \mathbb{E}^* \times \mathcal{X}$ satisfy¹ $\chi(\zeta_0) = x_0$, then the system has a unique continuously differentiable solution $(\zeta(t), x(t))$ for $0 \leq t < \infty$. Note that (21) is a natural generalization of (6) as in fact coincides with it in the Euclidean case $\chi(\zeta) = \zeta$.

If, $r \geq 2$, x^* is a minimizer, $\chi(\zeta^*) = x^*$, and Assumption 1 is satisfied, then

$$V(x, \zeta, t) = \frac{t^2}{r^2} (f(x) - f(x^*)) + D_{\psi^*}(\zeta, \zeta^*) \geq 0 \tag{22}$$

is a Lyapunov function for the system, i.e. $(d/dt)V(x(t), \zeta(t), t) \leq 0$ along solutions of (21). This immediately implies the following decay estimate of $f(x(t))$ towards the optimal value $f(x^*)$

$$f(x(t)) - f(x^*) \leq \frac{r^2}{t^2} V(x(0), \zeta(0), 0) = \frac{r^2}{t^2} D_{\psi^*}(\zeta(0), \zeta^*), \quad t > 0.$$

Note for future reference that V includes the dual variable ζ through D_{ψ^*} as in (12).

1. The requirement $\chi(\zeta_0) = x_0$ is imposed in view of the singularity of (21a) at $t = 0$.

As in the situation we discussed in the Euclidean case, discretizations $(\zeta_k, x_k) \mapsto (\zeta_{k+1}, x_{k+1})$ of (21), where (ζ_k, x_k) approximate $(\zeta(k\delta), x(k\delta))$ ($\delta > 0$ is the time-step), may offer the potential of providing optimization algorithms for which $f(x_k) - f(x^*)$ decays like $1/k^2$, i.e. algorithms that show acceleration. An algorithm with acceleration was suggested in Krichene et al. (2015a). It uses a learning rate $h > 0$, parameters $r > 0$ and $\gamma > 0$ and a *regularization* function R such that for $x, y \in \mathcal{X}$,

$$\frac{\ell_R}{2} \|x - y\|^2 \leq R(x, y) \leq \frac{L_R}{2} \|x - y\|^2.$$

We will refer to it as accelerated mirror descent with regularization (AMDR) and it is given in Algorithm 1. If $\delta = \sqrt{h}$, then the algorithm may be seen as a numerical method to integrate the

Algorithm 1 Accelerated Mirror Descent with Regularization (AMDR) (Krichene et al., 2015a)

Require: $N \in \mathbb{N}, h > 0, r \geq 0, \gamma > 0$ and regularizer R

0. **Initialize:** $x_0 \in \mathcal{X}, \zeta_0 \in \mathbb{E}^*$ ($\chi(\zeta_0) = x_0$)

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

1. $y_k \leftarrow x_k + \frac{r}{r+k} (\chi(\zeta_k) - x_k)$

2. $\zeta_{k+1} \leftarrow \zeta_k - \frac{kh}{r} \nabla f(y_k)$

3. $x_{k+1} \leftarrow \operatorname{argmin}_{x \in \mathcal{X}} (\gamma h \langle \nabla f(y_k), x \rangle + R(x, y_k))$

end for

return x_N

system (21), with ζ_k and x_k approximations to $\zeta(k\delta)$ and $x(k\delta)$ respectively. This is easily proved after taking into consideration that x_{k+1} and y_k differ by an $\mathcal{O}(\delta^2)$ amount (Krichene et al., 2015a).

The discretization in AMDR was constructed so as to *inherit* the Lyapunov function (22). This is a nontrivial task, because, typically, numerical integrators, even if very accurate, fail to reproduce the large t properties of the system being integrated; see the discussion in Sanz Serna and Zygalakis (2021). For this algorithm, it is proved in (Krichene et al., 2015a, Lemma 2) that, if $\gamma \geq L_R L_{\mathcal{X}}$ and $h \leq \ell_R / (2L_f \gamma)$, then

$$V(x_{k+1}, \zeta_{k+1}, (k+1)\delta) - V(x_k, \zeta_k, k\delta) \leq \frac{(2k+1-kr)}{r^2} (f(x_{k+1}) - f(x^*)).$$

For $r \geq 3, k \geq 1$, the right hand-side is ≤ 0 and thus the bound establishes an $\mathcal{O}(1/k^2)$ decay of $f(x_k) - f(x^*)$ (acceleration).

4.2 An alternative primal/dual discretization

The need for the regularisation function R in AMDR could be problematic. For example, in the case of the simplex, one possible choice of R is an ϵ -smooth entropy function (Krichene et al., 2015a,b). In that case, there is an efficient algorithm to implement Step 3 of AMDR, but unfortunately the value of γ to be used depends on ϵ and the learning rate h can become prohibitively small. Furthermore, beyond the simplex setting, it might not be obvious how to set R and implement Step 3. Motivated by this, we propose Algorithm 2 (AMD) that is a natural generalization of Nesterov's method in the non-Euclidean setting and makes no use of a regularization step. A learning rate $h > 0$ and a sequence $\{\gamma_k\}_{k=0}^{\infty}$ with $\gamma_0 = 1, \gamma_k \geq 1, k = 1, 2, \dots$, are required. Note that Step 3 is

Algorithm 2 Accelerated Mirror Descent (AMD)

Require: $N \in \mathbb{N}$, $\{\gamma_k\}_{k=0}^{N-1}$, $h > 0$
 0. Initialize $x_0 \in \mathcal{X}$, $\zeta_0 \in \mathbb{E}^*$,
for $k = 0, \dots, N - 1$ **do**
 1. $y_k \leftarrow x_k + \frac{1}{\gamma_k}(\chi(\zeta_k) - x_k)$
 2. $\zeta_{k+1} \leftarrow \zeta_k - \gamma_k h \nabla f(y_k)$
 3. $x_{k+1} \leftarrow y_k + \frac{1}{\gamma_k}(\chi(\zeta_{k+1}) - \chi(\zeta_k))$
end for
return x_N

equivalent to

$$x_{k+1} = x_k + \frac{1}{\gamma_k}(\chi(\zeta_{k+1}) - x_k). \quad (23)$$

In Step 1, y_k is a convex combination of x_k and $\chi(z_k)$ and in (23) x_{k+1} is a convex combination of x_k and $\chi(z_{k+1})$. By induction, all the y_k are in \mathcal{X} (so that $\nabla f(y_k)$ makes sense) and all the x_k are also in \mathcal{X} .

In the Euclidean case this algorithm reduces to (7), i.e. to the Nesterov algorithm implemented as a one-step recursion with the help of the variable z . The standard implementation (8)–(9) of Nesterov’s algorithm cannot be directly applied to the non Euclidean scenario since (8) mixes primal and dual variables. In AMD such a mixing is avoided; in Step 2, the gradients are accumulated in a dual variable (as in mirror descent (13)) and the primal mirror images of the dual variable are used to perform the convex combinations in Steps 1 and 3.

After defining $\delta = \sqrt{h}$ and $\tilde{t}_k = r\delta\gamma_k$, Steps 1 and 2 in Algorithm 2 and (23) imply

$$\begin{aligned} \frac{1}{\delta}(y_k - x_k) &= \frac{r}{\tilde{t}_k}(\chi(z_k) - x_k), \\ \frac{1}{\delta}(\zeta_{k+1} - \zeta_k) &= -\frac{\tilde{t}_k}{r}\nabla f(y_k), \\ \frac{1}{\delta}(x_{k+1} - x_k) &= \frac{r}{\tilde{t}_k}(\chi(z_{k+1}) - x_k). \end{aligned}$$

If we now assume that

$$\gamma_k = \frac{k}{r} + o(k), \quad k \rightarrow \infty, \quad (24)$$

then $\tilde{t}_k \rightarrow k\delta$ as $k \rightarrow \infty$, $\delta \rightarrow 0$ with $k\delta$ constant. We therefore have the following result:

Theorem 2 *Suppose that Assumption 1 and (24) hold, then AMD provides a consistent one-step numerical integrator $(\zeta_k, x_k) \mapsto (\zeta_{k+1}, x_{k+1})$ for the system of ODEs (21).*

The following result (proved in Section 6) implies that, for suitable choices of the learning rate h and the constants γ_k , AMD is indeed an accelerated optimization method.

Theorem 3 *Suppose that Assumption 1 holds and that the coefficients γ_k satisfy*

$$\gamma_k^2 - \gamma_{k-1}^2 - \gamma_k \leq 0, \quad k = 1, 2, \dots \quad (25)$$

Assume that f is L_f -smooth, i.e. for each x, y in the domain of f :

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_f \|x - y\| \quad (26)$$

and that

$$h \leq \frac{1}{L_f L_\chi}. \quad (27)$$

Let x^* be a minimizer of (1) and $\zeta^* \in \mathbb{E}^*$ satisfy $\chi(\zeta^*)$, and define for $k = 0, 1, \dots$

$$V_k(x_k, \zeta_k) = (\gamma_k^2 - \gamma_k)h(f(x_k) - f(x^*)) + D_{\psi^*}(\zeta_k, \zeta^*). \quad (28)$$

Then, along trajectories generated by AMD

$$V_{k+1}(x_{k+1}, \zeta_{k+1}) \leq V_k(x_k, \zeta_k), \quad k = 0, 1, \dots$$

and therefore

$$(\gamma_k^2 - \gamma_k)h(f(x_k) - f(x^*)) \leq D_{\psi^*}(\zeta_0, \zeta^*), \quad k = 0, 1, \dots$$

Under the consistency condition (24), the right hand-side of (28) converges, in the limit $\delta \rightarrow 0$, $k\delta \rightarrow t$, to the Lyapunov function (22) of the differential equations. The choice

$$\gamma_k = \frac{k+r}{r}, \quad k = 0, 1, \dots \quad (29)$$

fulfills the consistency requirement (24) and, if $r \geq 2$, also the condition (25). Recall that the same condition on r is required for (22) to be a Lyapunov function for the differential equations unlike the case of AMDR. When the coefficients γ_k are chosen as in (29), the theorem shows a decay $f(x_k) - f(x^*)$ like $1/((\gamma_k^2 - \gamma_k)h) \sim r^2/(k^2h)$ as $k \rightarrow \infty$.

The best decay of $f(x_k) - f(x^*)$ that may be proved with the theorem occurs when the γ_k are chosen as large as possible subject to (25), i.e. when the inequality in (25) becomes an equality. In this case, we have the well-known recurrence (10). These coefficients are slightly larger than (29) with $r = 2$ and, accordingly, guarantee a slightly better convergence. One may prove that with this recurrence, as $k \rightarrow \infty$, $\gamma_k = k/2 + (1/4) \log k + o(\log k)$, to be compared with the estimate $\gamma_k = k/2 + \mathcal{O}(1)$ valid for (29) with $r = 2$.

4.2.1 CONNECTION WITH ADDITIVE RUNGE KUTTA METHODS

As pointed out in Dobson et al. (2023), when run with parameters adapted to strongly convex objective functions in Euclidean space, Nesterov algorithms may be interpreted as an *Additive Runge-Kutta* method (Cooper and Sayfy, 1980, 1983) for integrating ordinary differential equations. A similar interpretation exists here. We set $\xi = (\zeta, x)$ and write (21) by additively decomposing the right hand-side as

$$\dot{\xi} = g^{[1]}(\xi, t) + g^{[2]}(\xi, t) + g^{[3]}(\xi, t),$$

with

$$g^{[1]}(\xi, t) = \begin{bmatrix} 0 \\ -\frac{r}{t}x \end{bmatrix}, \quad g^{[2]}(\xi, t) = \begin{bmatrix} 0 \\ \frac{r}{t}\chi(\zeta) \end{bmatrix}, \quad g^{[3]}(\xi, t) = \begin{bmatrix} -\frac{t}{r}\nabla f(x) \\ 0 \end{bmatrix}.$$

Then the step $(\zeta_k, x_k) \mapsto (\zeta_{k+1}, x_{k+1})$ may be written in a Runge-Kutta fashion as

$$\xi_{k+1} = \xi_k + \delta g^{[1]}(\Xi_{k,1}, \tilde{t}_k) + \delta g^{[2]}(\Xi_{k,2}, \tilde{t}_k) + \delta g^{[3]}(\Xi_{k,3}, \tilde{t}_k)$$

with the so-called stage vectors defined by

$$\begin{aligned} \Xi_{k,1} &= \xi_k, \\ \Xi_{k,2} &= \xi_k + \delta g^{[1]}(\Xi_{k,1}, \tilde{t}_k) + \delta g^{[2]}(\Xi_{k,1}, \tilde{t}_k), \\ \Xi_{k,3} &= \xi_k + \delta g^{[1]}(\Xi_{k,1}, \tilde{t}_k) + \delta g^{[2]}(\Xi_{k,2}, \tilde{t}_k) + \delta g^{[3]}(\Xi_{k,2}, \tilde{t}_k). \end{aligned}$$

Note that $\Xi_{k,1} = (\zeta_k, x_k)$, $\Xi_{k,2} = (\zeta_k, y_k)$, $\Xi_{k,3} = (\zeta_{k+1}, y_k)$. Thus the successive computations of $\Xi_{k,2}$, $\Xi_{k,3}$, and ξ_{k+1} in the Additive Runge-Kutta scheme represent the computations of y_k , ζ_{k+1} , x_{k+1} .

4.3 Convergence when x^* is not in the image of the mirror map

The Lyapunov functions (22)–(28), used for establishing convergence for the ODE (21) and its discretizations AMDR and AMD, contain a term $D_{\psi^*}(\zeta, \zeta^*)$, where $\chi(\zeta^*) = x^*$. Therefore they cannot be used to establish convergence when the minimizer x^* is not in the image of the mirror map. In the case of the simplex, this implies that one cannot treat minimizers x^* having one or more zero components. Similarly, for the hypercube, minimizers having some of their components equal to 0 or 1 cannot be dealt with. In this subsection we remove this limitation by using Lyapunov functions that, as distinct from those considered above or in Krichene et al. (2015a), are formulated purely in terms of *primal* variables. Accordingly we will operate with *Bregman divergences* defined in \mathbb{E} rather than in \mathbb{E}^* and this will require that Assumption 2 holds.

4.3.1 THE DIFFERENTIAL SYSTEM

Using the primal Bregman divergence D_ϕ , for the system of differential equations (21), in lieu of the Lyapunov function (22), we may alternatively consider

$$\widehat{V}(x, \zeta, t) = \frac{t^2}{r^2} (f(x) - f(x^*)) + D_\phi(x^*, \chi(\zeta)), \quad (30)$$

where we note that $D_\phi(x^*, \chi(\zeta))$ is well defined because χ takes values in $\text{ri}(\mathcal{X})$. Now the existence of ζ^* with $\chi(\zeta^*) = x^*$ is not required. If such a ζ^* exists, then the numerical values of (22) and (30) coincide, according to well-known properties of the Bregman divergence. The following theorem, proved in Section 6, shows that, for $r \geq 2$, \widehat{V} is indeed a Lyapunov function and therefore $f(x(t)) - f(x^*)$ decays like $1/t^2$.

Theorem 4 *Suppose that Assumptions 1 and 2 hold. If $r \geq 2$, then along solutions of (21), $(d/dt)\widehat{V} \leq 0$.*

4.3.2 ALGORITHM 1 (AMDR)

For AMDR, Lemma 2 in Krichene et al. (2015a) may be replaced by the following new result whose proof is given in Section 6. It implies that for $r \geq 3$ we shall have acceleration even if x^* is not in the image of χ .

Theorem 5 *Suppose that Assumptions 1 and 2 hold. If f is L_f -smooth, $\gamma \geq L_R L_\chi$ and $h \leq \ell_R / (2L_f \gamma)$, then for AMDR*

$$\widehat{V}(x_{k+1}, \zeta_{k+1}, (k+1)\delta) - \widehat{V}(x_k, \zeta_k, k\delta) \leq \frac{(2k+1-kr)h}{r^2} (f(x_{k+1}) - f(x^*)).$$

4.3.3 ALGORITHM 2 (AMD)

In the context of the AMD instead of using the discrete Lyapunov function (28), we will alternatively consider for $k = 0, 1, \dots$

$$\widehat{V}_k(x_k, \zeta_k) = (\gamma_k^2 - \gamma_k)h(f(x_k) - f(x^*)) + D_\phi(x^*, \chi(\zeta_k)). \quad (31)$$

By using \widehat{V} , Theorem 3 may be strengthened as follows (see Section 6 for the proof):

Theorem 6 *Suppose that Assumptions 1 and 2 hold and that the coefficients γ_k satisfy (25). Assume that f is L_f -smooth and that*

$$h \leq \frac{1}{L_f L_\chi}.$$

Let x^ be a minimizer of (1). Then, for (x_{k+1}, ζ_{k+1}) given by AMD*

$$\widehat{V}_{k+1}(x_{k+1}, \zeta_{k+1}) \leq \widehat{V}_k(x_k, \zeta_k), \quad k = 0, 1, \dots$$

and therefore

$$(\gamma_k^2 - \gamma_k)h(f(x_k) - f(x^*)) \leq D_\phi(x^*, \chi(\zeta_0)), \quad k = 0, 1, \dots$$

Thus the decay of $f(x_k) - f(x^*)$ takes place whether x^* is in the image of χ or otherwise.

4.4 A primal accelerated ODE and its discretizations

In the literature several ODEs and algorithms have appeared that have similarities to those in Krichene et al. (2015a). We now investigate this further in the case of ODEs appearing in (Wibisono et al., 2016; Wilson et al., 2021) and one of the algorithms given in Tseng (2008). One notable difference is that, contrary to the setting in Krichene et al. (2015a), only primal variables are used in those references. Motivated by this, we will now give formulations of the ODE (21) and Algorithms 1 and 2 that only make use of primal variables. The developments parallel those in Section 3.2. In this subsection it is assumed that Assumption 2 holds.

4.4.1 PRIMAL WRITING IN THE CASE $\mathcal{N} = \{0\}$

The differential system In the case where \mathcal{N} is trivial, the one-to-one correspondence between $z = \chi(\zeta)$ and $\zeta = \nabla\phi(z)$ may be used to rewrite the system (21) in terms of the primal variable z as follows:

$$\frac{d}{dt} \nabla\phi(z(t)) = -\frac{t}{r} \nabla f(x(t)), \quad (32a)$$

$$\dot{x}(t) = \frac{r}{t} (z(t) - x(t)). \quad (32b)$$

For the value $r = 2$, this system is a particular instance of the systems derived in Wilson et al. (2021) through Lagrangian functions. Note that when \mathcal{N} is not reduced to $\{0\}$, the differential equation (32a) is meaningless because the left hand-side is constrained to be a vector tangent to the manifold \mathcal{M} and $\nabla f(x)$ is not constrained in that way (see the right panel in Figure 2).

The algorithms As is the case for the differential system, when $\mathcal{N} = \{0\}$, AMDR and AMD may be expressed avoiding dual variables. Restricting the attention to AMD (AMDR may be dealt with analogously), we have

$$\begin{aligned} y_k &= x_k + \frac{1}{\gamma_k}(z_k - x_k), \\ \nabla\phi(z_{k+1}) &= \nabla\phi(z_k) - \gamma_k h \nabla f(y_k), \\ x_{k+1} &= y_k + \frac{1}{\gamma_k}(z_{k+1} - z_k). \end{aligned} \tag{33}$$

Formulas similar to (33) appear in (Wibisono et al., 2016; Wilson et al., 2021). Once $\nabla\phi(z_{k+1})$ has been found via (33), z_{k+1} is recovered as $\chi(\nabla\phi(z_{k+1}))$. Note that when \mathcal{N} is not reduced to $\{0\}$, (33) is meaningless because nothing guarantees that its right hand-side lies in \mathcal{M} , the set where $\nabla\phi$ takes values.

4.4.2 PRIMAL WRITING IN THE GENERAL CASE

The differential system It is possible to obtain a system of differential equations satisfied by the primal variables $z(t) = \chi(\zeta(t))$ and $x(t)$ without the hypothesis $\mathcal{N} = \{0\}$, where $\zeta(t)$ and $x(t)$ are solutions of (21). In fact, by arguing as in Section 3.2 and denoting by z the mirror of the variable ζ , one may write:

$$\dot{z} = \chi'(\nabla\phi(z(t))) \left(-\frac{t}{r} \nabla f(x(t)) \right), \tag{34a}$$

$$\dot{x} = \frac{r}{t}(z(t) - x(t)); \tag{34b}$$

in the particular where $\mathcal{N} = \{0\}$, this reduces to (32). The system (34) has to be initialized with $x(0) = z(0)$ in the relative interior of \mathcal{X} , and from (30), it has the Lyapunov function:

$$\frac{t^2}{r^2}(f(x) - f(x^*)) + D_\phi(x^*, z). \tag{35}$$

Example 2 (Example 1 continued) *In the case of the simplex, (34a) reads*

$$\dot{z} = D(z) \left(-\frac{t}{r} \nabla f(x) + \langle -\frac{t}{r} \nabla f(x), z \rangle \mathbf{1} \right), \tag{36}$$

where $D(z)$ is the diagonal matrix with entries z_i ;

The algorithms It is also possible, without any assumption on the dimension of \mathcal{N} , to express AMDR and AMD in terms of the primal points $z_k = \chi(\zeta_k)$. For brevity we only give details for AMD, which, by arguing as in Section 3.2, may be reformulated as

$$y_k = x_k + \frac{1}{\gamma_k}(z_k - x_k), \tag{37a}$$

$$z_{k+1} = \chi(\nabla\phi(z_k) - \gamma_k h \nabla f(y_k)), \tag{37b}$$

$$x_{k+1} = y_k + \frac{1}{\gamma_k}(z_{k+1} - z_k). \tag{37c}$$

This is a particular case of the algorithms studied in Tseng (2008). Note that in Tseng (2008) there is no discussion of the continuous time limit; however it is not difficult to show that, when (24) is fulfilled, the algorithm (37) is a consistent discretization of the system (34). Furthermore, in terms of primal variables, the discrete-time Lyapunov function in (31) is given by

$$(\gamma_k^2 - \gamma_k)h(f(x_k) - f(x^*)) + D_\phi(x^*, z_k).$$

Under the consistency requirement (24), this is an approximation to the continuous-time Lyapunov function (35).

5 Numerical experiments

We now illustrate the performance of AMDR and AMD. The standard mirror descent algorithm (20) will be used as a benchmark. All experiments reported correspond to the simplex. Recall that in this case it is possible to run AMDR with an efficient regularizer, something that may or may not be the situation for other instances of \mathcal{X} .

In the experiments that follow, we set $r = 3$ for AMDR and, as in Krichene et al. (2015a), use $\gamma = 1$ and perform Step 3 by means of the efficient procedure in (Krichene et al., 2015a, Algorithm 4) with $\epsilon = 0.3$. It turns out that with this setting, the computational costs per step of AMDR and AMD are virtually identical and also coincide with those of mirror descent. For AMD we use γ_k given by (10).

5.1 Non strongly convex objective function

In order to check that AMDR and AMD provide acceleration, we first consider an extremely simple toy example with $d = 2$, $f(x) = (1/p)((x_1 - 1/2)^p + (x_2 - 1/2)^p)$, $p = 10$. The initial condition is chosen as $[0.999, 0.001]^T$ and the three algorithms were run with different choices of the learning rate. Results for the representative value $h = 1$ may be seen in Figure 3. While for mirror descent, the decay is slightly better than $1/k$, for AMDR and AMD the decay is slightly better than $1/k^2$. Increasing the value of the parameter p results in rates that become closer to $1/k$ for mirror descent and to $1/k^2$ for the other two algorithms.

5.2 Quadratic objective function

Now the objective function is $f(x) = (1/2)x^T B^T Bx$, with B a $d \times d$ matrix with entries given by independent standard normal random variables. The initial x_0 is chosen randomly by generating a vector with independent, uniformly random components in $[0, 1]$ and then rescaling to ensure that $\sum_j x_j = 1$. The smoothness constant L_f for the gradient $\nabla f(x) = B^T Bx$ is the norm of $B^T B$ as an operator from (\mathbb{R}^d, ℓ^1) to $(\mathbb{R}^d, \ell^\infty)$, which is given by the maximum $m(B^T B)$ of the absolute value of the entries. In Assumption 1, $L_\chi = 1$ and, in view of condition (27) in Theorem 3, we run AMD with a learning rate $h = 1/m(B^T B)$; the same value is used for mirror descent. For AMDR we follow the prescription in Krichene et al. (2015a) and set

$$h = \sqrt{\epsilon / (2(1 + d\epsilon)m(B^T B)\gamma)}.$$

The experiment in Figure 4 has $d = 1000$ and 50,000 steps. The minimizer is not in the relative interior (in fact has 323 vanishing components) and the results in Section 4.3 are necessary to

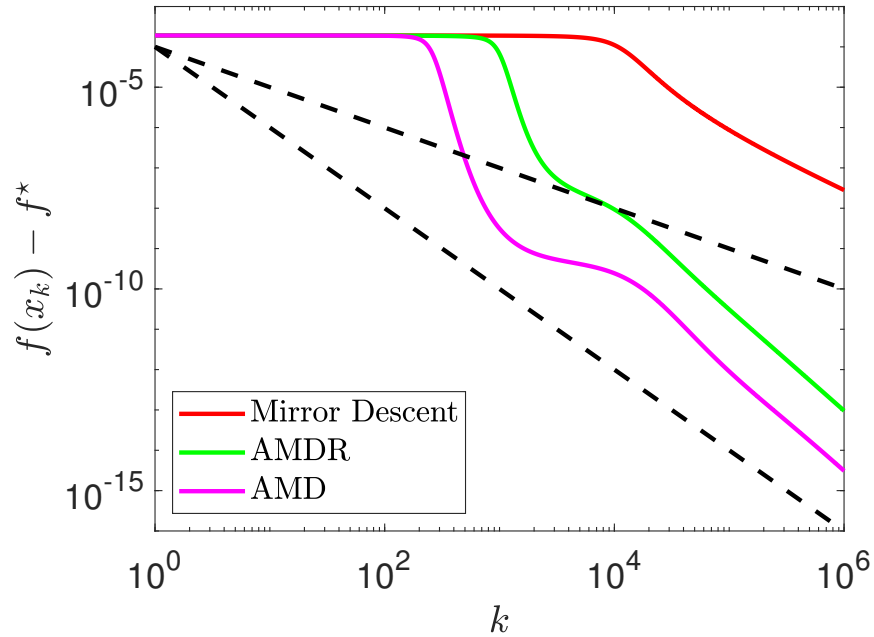


Figure 3: Non strongly convex objective function, $f(x_k) - f(x^*)$ vs. k . The dotted lines have slopes corresponding to decays $1/k$ and $1/k^2$.

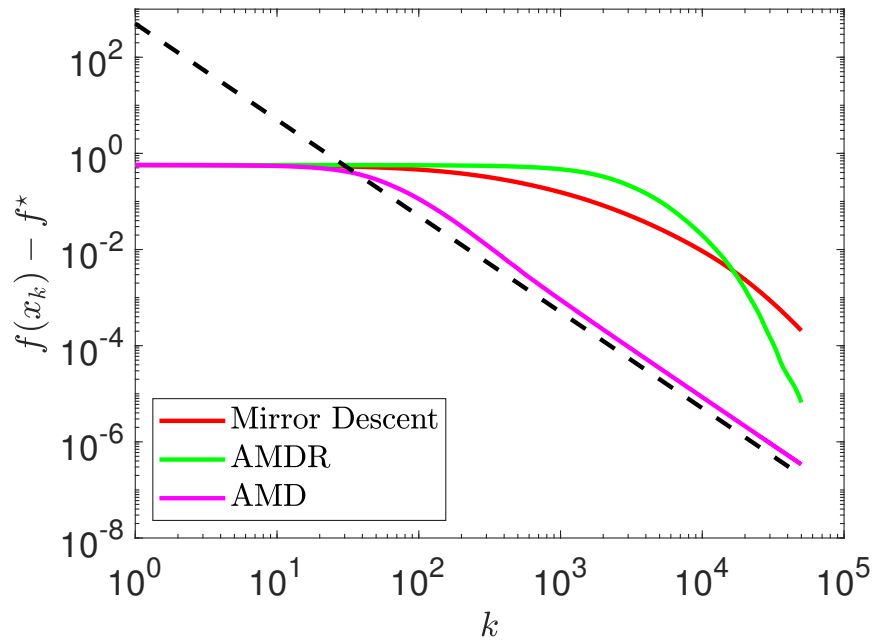


Figure 4: Quadratic objective function, $f(x_k) - f(x^*)$ vs. k . The dotted line has a slope corresponding to a decay $1/k^2$.

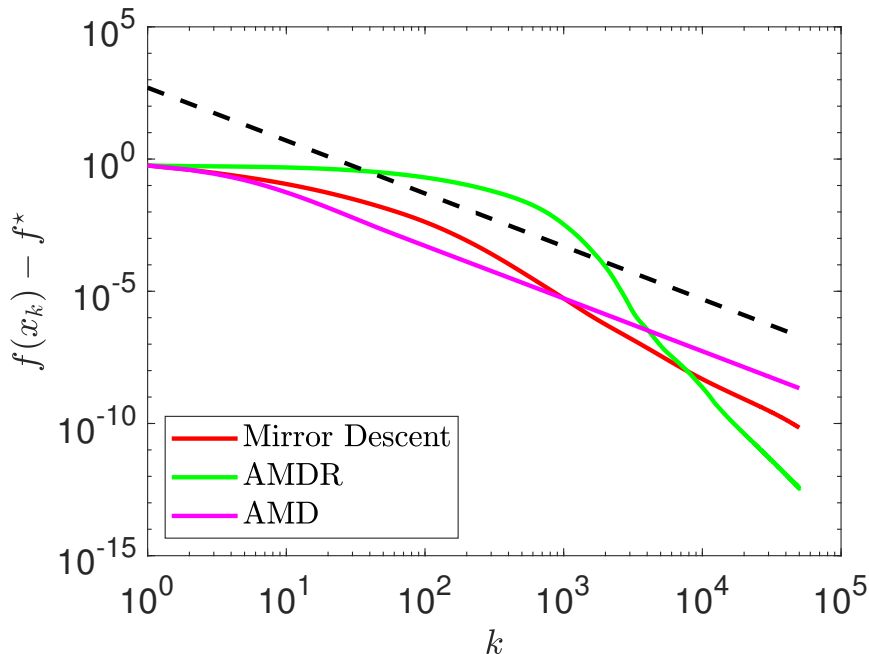


Figure 5: Quadratic objective function, larger learning rates, $f(x_k) - f(x^*)$ vs. k . The dotted line has the same equation as the reference line in Figure 4 so as to make it easy to compare both figures.

establish the convergence of AMDR and AMD. In addition, in the experiment, $m(B^T B) \approx 1.2 \times 10^3$ which leads to learning rates $h \approx 8.5 \times 10^{-4}$ for mirror descent and AMD and $h \approx 6.5 \times 10^{-4}$ for AMDR. The figure clearly bears out the $1/k^2$ acceleration proved in Theorem 6 for AMD. Mirror descent and AMDR lead initially to very little decay in f but they decay faster than $1/k^2$ once they are near the minimizer. In this particular experiment they are both outperformed by AMD.

5.3 Quadratic objective function, larger learning rates

Numerical experimentation reveals that the recipes we have just used to determine the learning rates are too pessimistic; the three algorithms tested may operate with substantially larger values of h , thus providing a larger decay in f for a given number of iterations. In fact, the values of h were based on the size of $m(B^T B)$ (an operator norm for the matrix in $\nabla f(x) = B^T Bx$). However, in (36) or (18) we see that in the differential equations being approximated by the algorithms, $\nabla f(x)$ is premultiplied by $D(z) = \text{diag}(z)$ or $D(x) = \text{diag}(x)$ respectively. Once the solution is close to the minimizer, those matrices are close to $D(x^*)$ and it is reasonable to think that the learning rates should really be determined by the size of the matrix $D(x^*)B^T B$ rather than by the size of $B^T B$. If $D(x^*)B^T B$ is much smaller than $B^T B$ learning rates based on the size of $B^T B$ may be expected to be unduly pessimistic.

These considerations may be related to the notion of *relative smoothness* introduced in Lu et al. (2018) (see also Bauschke et al. (2017)), an alternative to the notion of (absolute) smoothness in (26). For the sake of brevity, we only present the concept of relative smoothness as it applies to the simplex. A real function g , twice continuously differentiable, is said to be L_r -relatively smooth

with respect to the negative entropy ϕ in (15), if for z in the relative interior of the simplex

$$\nabla^2 g(z) \preceq L_r \nabla^2 \phi(z).$$

Recalling that $D(z)$ is the inverse of $\nabla^2 \phi(z)$, we may equivalently write

$$D(z)^{1/2} \nabla^2 g(z) D(z)^{1/2} \preceq L_r I,$$

and therefore the best possible L_r is given by the maximum of the spectral radius of the symmetric matrix $D(z)^{1/2} \nabla^2 g(z) D(z)^{1/2}$ as z ranges in the relative interior. Note that this symmetric matrix is similar to the matrix $D(z) \nabla^2 g(z)$ and therefore shares its eigenvalues. For $f(z) = (1/2) z^T B^T B z$, $D(z) \nabla^2 f(z) = D(z) B^T B$ and therefore L_r is the maximum eigenvalue of $D(z) B^T B$ (while as pointed out above L_f is the maximum of the entries of $B^T B$). It may then be conjectured that the assumption of relative smoothness of f as in (26) could be replaced by the assumption that the objective function be L_r -relatively smooth and that the algorithms may be operated with learning rates based on using the value of L_r rather than on the value of L_f .

To investigate this conjecture, we revisit the experiment in Figure 4. There, as mentioned before, x^* has 323 vanishing entries. In addition the maximum entry of x^* happens to be 5.9×10^{-3} (since the entries add up to 1 they might be expected to be small). Thus, the entries of $D(x^*) B^T B$ are more than two orders of magnitude smaller than those $B^T B$. We introduced the symmetric matrix

$$M = D(x^*)^{1/2} B^T B D(x^*)^{1/2}$$

and estimated L_r by the spectral radius $\rho(M)$.² Then we used the learning rate $h = 1/(L_\chi L_r)$ (rather than $1/(L_\chi L_f)$) for mirror descent and AMD and

$$h = \sqrt{\epsilon / (2(1 + d\epsilon)\rho(M))\gamma}$$

for AMDR. Figure 5 has the same realizations of the random elements $B^T B$ and x_0 as Figure 4, the only difference being that we now use the L_r -based values of the learning rates just described; these turn out to be $h \approx 1.3 \times 10^{-1}$ for mirror descent and AMD and $h \approx 8.1 \times 10^{-3}$ for AMDR. In Figure 5 each optimization method qualitatively behaves very much as it did in Figure 4; however for each method the size of $f(x_k) - f(x^*)$ for given k is now clearly smaller than it was.

This experiment shows the interest of future analyses of AMDR and AMD replacing the notion of absolute smoothness by the notion of relative smoothness, similarly to what it is done in (Lu et al., 2018, Theorem 3.1).

6 Proofs

This section contains the proofs of the results in the paper.

6.1 Proof of Lemma 1

Part 1. This is a trivial consequence of the fact that χ maps \mathbb{E}^* onto $\text{ri}(\mathcal{X})$.

2. Of course, this would not make sense in a real application, as M requires the knowledge of x^* . In practice L_r could be estimated as the spectral radius of $D(x_k)^{1/2} B^T B D(x_k)^{1/2}$ for a suitable k .

Part 2. Fix $z \in \text{ri}(\mathcal{X})$. Since $\chi = \nabla\psi^*$, by a well-known result on conjugate functions, see e.g. (Beck, 2017, Theorem 4.20), the relation $\chi(\zeta) = z$ is equivalent to $\zeta \in \partial\psi(z)$. It is then sufficient to prove that $\partial\psi(z) = \nabla\phi(z) + \mathcal{N}$. By definition, $g \in \partial\psi(z)$ means that, for each $x \in \mathbb{E}$, $\psi(x) \geq \psi(z) + \langle g, x - z \rangle$. This inequality trivially holds if $x \notin \mathcal{X}$, because then $\psi(x) = \phi(x) + \delta_{\mathcal{X}}(x) = \infty$. Therefore $g \in \partial\psi(z)$ if and only if, for $x \in \mathcal{X}$, $\psi(x) \geq \psi(z) + \langle g, x - z \rangle$, that is $\phi(x) \geq \phi(z) + \langle g, x - z \rangle$. On the other hand, since ϕ is differentiable at z , the linear approximation at z based on the use of $\nabla\phi(z)$ given by $\ell(x) = \phi(z) + \langle \nabla\phi(z), x - z \rangle$ is unique in satisfying $\phi(x) \geq \ell(x)$ for each $x \in \mathbb{E}$. In this way $g \in \partial\psi(z)$ if and only if $\langle g, x - z \rangle = \langle \nabla\phi(z), x - z \rangle$ for each $x \in \mathcal{X}$, which in turn is clearly equivalent to $g - \nabla\phi(z) \in \mathcal{N}$.

Part 3. We have $\zeta \in \chi^{-1}(\chi(\zeta))$, so that, by Part 2, $\zeta \in \nabla\phi(\chi(\zeta)) + \mathcal{N}$ or $\nabla\phi(\chi(\zeta)) - \zeta \in \mathcal{N}$.

Part 4. By Part 2, $\chi^{-1}(\chi(\zeta))$ is an affine set associated to the vector space \mathcal{N} . Therefore the relations $\zeta \in \chi^{-1}(\chi(\zeta))$ and $\eta \in \mathcal{N}$ imply $\zeta + \eta \in \chi^{-1}(\chi(\zeta))$, so that $\chi(\zeta) = \chi(\zeta + \eta)$. Differentiation leads to $\chi'(\zeta) = \chi'(\zeta + \eta)$.

6.2 Proof of Theorem 3

We first deal with the part of the Lyapunov function that involves the variable x . By the convexity and smoothness of f ,

$$f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L_f}{2} \|x_{k+1} - y_k\|^2,$$

and, from the definition of x_{k+1} in Step 3 of Algorithm 2, after shortening slightly the notation,

$$f(x_{k+1}) \leq f(y_k) + \langle \nabla_k, x_{k+1} - y_k \rangle + \frac{L_f}{2\gamma_k^2} \|\chi_{k+1} - \chi_k\|^2.$$

Here $\nabla_k = \nabla f(y_k)$ and $\chi_k = \chi(\zeta_k)$. We now use (23) to get

$$\begin{aligned} f(x_{k+1}) &\leq f(y_k) + \langle \nabla_k, \left(1 - \frac{1}{\gamma_k}\right) x_k + \frac{1}{\gamma_k} \chi_{k+1} - y_k \rangle + \frac{L_f}{2\gamma_k^2} \|\chi_{k+1} - \chi_k\|^2 \\ &= \left(1 - \frac{1}{\gamma_k}\right) \left(f(y_k) + \langle \nabla_k, x_k - y_k \rangle\right) \\ &\quad + \frac{1}{\gamma_k} \left(f(y_k) + \langle \nabla_k, \chi_{k+1} - y_k \rangle\right) + \frac{L_f}{2\gamma_k^2} \|\chi_{k+1} - \chi_k\|^2, \\ &= \left(1 - \frac{1}{\gamma_k}\right) \left(f(y_k) + \langle \nabla_k, x_k - y_k \rangle\right) + \frac{1}{\gamma_k} \left(f(y_k) + \langle \nabla_k, x^* - y_k \rangle\right) \\ &\quad + \frac{1}{\gamma_k} \langle \nabla_k, \chi_{k+1} - x^* \rangle + \frac{L_f}{2\gamma_k^2} \|\chi_{k+1} - \chi_k\|^2. \end{aligned}$$

Invoking again the convexity of f

$$\begin{aligned} f(x_{k+1}) &\leq \left(1 - \frac{1}{\gamma_k}\right) f(x_k) + \frac{1}{\gamma_k} f(x^*) + \frac{1}{\gamma_k} \langle \nabla_k, \chi_{k+1} - x^* \rangle \\ &\quad + \frac{L_f}{2\gamma_k^2} \|\chi_{k+1} - \chi_k\|^2, \end{aligned}$$

and therefore

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq \left(1 - \frac{1}{\gamma_k}\right) (f(x_k) - f(x^*)) + \frac{1}{\gamma_k} \langle \nabla_k, \chi_{k+1} - x^* \rangle \\ &\quad + \frac{L_f}{2\gamma_k^2} \|\chi_{k+1} - \chi_k\|^2. \end{aligned}$$

We now multiply across by $\gamma_k^2 h$ and take into account the formula in Step 2 of the algorithm,

$$\begin{aligned} \gamma_k^2 h (f(x_{k+1}) - f(x^*)) &\leq (\gamma_k^2 - \gamma_k) h (f(x_k) - f(x^*)) \\ &\quad - \langle \zeta_{k+1} - \zeta_k, \chi_{k+1} - x^* \rangle + \frac{L_f h}{2} \|\chi_{k+1} - \chi_k\|^2. \end{aligned}$$

The bounds (25) and (27) then yield

$$\begin{aligned} (\gamma_{k+1}^2 - \gamma_{k+1}) h (f(x_{k+1}) - f(x^*)) &\leq (\gamma_k^2 - \gamma_k) h (f(x_k) - f(x^*)) \\ &\quad - \langle \zeta_{k+1} - \zeta_k, \chi_{k+1} - x^* \rangle + \frac{1}{2L_\chi} \|\chi_{k+1} - \chi_k\|^2. \end{aligned} \tag{38}$$

We next address the part of the Lyapunov function involving the variable ζ . By using the three-points lemma for D_{ψ^*} as in (Krichene et al., 2015a, Lemma 5) and recalling that $\chi(\zeta^*) = x^*$,

$$D_{\psi^*}(\zeta_{k+1}, \zeta^*) = D_{\psi^*}(\zeta_k, \zeta^*) - D_{\psi^*}(\zeta_k, \zeta_{k+1}) + \langle \zeta_{k+1} - \zeta_k, \chi_{k+1} - x^* \rangle,$$

and the smoothness bound in (Krichene et al., 2015a, Lemma 5) yields

$$D_{\psi^*}(\zeta_{k+1}, \zeta^*) \leq D_{\psi^*}(\zeta_k, \zeta^*) - \frac{1}{2L_\chi} \|\chi_{k+1} - \chi_k\|^2 + \langle \zeta_{k+1} - \zeta_k, \chi_{k+1} - x^* \rangle.$$

It is now sufficient to add the last bound to (38).

6.3 Proof of Theorem 4

By differentiating we have,

$$\begin{aligned} \frac{d}{dt} \widehat{V} &= \frac{2t}{r^2} (f(x(t)) - f(x^*)) + \frac{t^2}{r^2} \langle \nabla f(x(t)), \dot{x}(t) \rangle \\ &\quad + \left\langle \frac{d}{dt} \nabla \phi(\chi(\zeta(t))), x^* - \chi(\zeta(t)) \right\rangle, \end{aligned}$$

and Lemma 1, Part 3, implies

$$\left\langle \frac{d}{dt} \nabla \phi(\chi(\zeta(t))), x^* - \chi(\zeta(t)) \right\rangle = \langle \dot{\zeta}(t), x^* - \chi(\zeta(t)) \rangle.$$

These two equalities and the differential equations (21a)–(21b) may be combined to yield:

$$\begin{aligned} \frac{d}{dt} \widehat{V} &= \left(\frac{2}{r} - 1 \right) \frac{t}{r} (f(x(t)) - f(x^*)) \\ &\quad - \frac{t}{r} \left(f(x^*) - f(x(t)) - \langle \nabla f(x(t)), x^* - x(t) \rangle \right). \end{aligned}$$

Both terms in the right hand-side are ≤ 0 , the first because $r \geq 2$ and the second because f is convex.

6.4 Proof of Theorems 5 and 6

We begin with Theorem 6. We reproduce the proof of Theorem 3 until we reach (38). For the part involving Bregman divergences, the three-points lemma for D_ϕ gives

$$D_\phi(x^*, \chi_{k+1}) = D_\phi(x^*, \chi_k) - D_\phi(\chi_{k+1}, \chi_k) + \langle \nabla\phi(\chi(\zeta_{k+1})) - \nabla\phi(\chi(\zeta_k)), \chi_{k+1} - x^* \rangle$$

and, as a consequence of Lemma 1, Part 3, the differences $\nabla\phi(\chi(\zeta_{k+1})) - \zeta_{k+1}$ and $\nabla\phi(\chi(\zeta_k)) - \zeta_k$ are in \mathcal{N} and we may alternatively write

$$D_\phi(x^*, \chi_{k+1}) = D_\phi(x^*, \chi_k) - D_\phi(\chi_{k+1}, \chi_k) + \langle \zeta_{k+1} - \zeta_k, \chi_{k+1} - x^* \rangle.$$

From (Beck, 2017, Lemma 9.4 (a))

$$D_\phi(x^*, \chi_{k+1}) \leq D_\phi(x^*, \chi_k) - \frac{1}{2L_\chi} \|\chi_{k+1} - \chi_k\|^2 + \langle \zeta_{k+1} - \zeta_k, \chi_{k+1} - x^* \rangle,$$

because $\psi = \phi + \delta_{\mathcal{X}}$ is $(1/L_\chi)$ -strongly convex (Beck, 2017, Theorem 5.26(a)). The proof concludes by adding the last bound to (38).

To prove Theorem 5 one may proceed similarly. The part of the Lyapunov function involving x is dealt with as in the proof of Lemma 2 in Krichene et al. (2015a) and the difference in Bregman divergences is treated exactly as in the proof of Theorem 6 just given.

Acknowledgments and Disclosure of Funding

JMS has been funded by Ministerio de Ciencia e Innovación (Spain), project PID2022-136585NB-C21, MCIN/AEI/10.13039/501100011033/FEDER, UE. PD and KCZ acknowledge support from the EPSRC grant EP/V006177/1.

References

- H.H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: First order methods revisited and applications. *Math. Oper. Res.*, pages 330–348, 2017.
- A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.
- S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4): 231–357, 2015.
- G. J. Cooper and A. Sayfy. Additive methods for the numerical solution of ordinary differential equations. *Math. Comput.*, 35, 1980.
- G. J. Cooper and A. Sayfy. Additive Runge-Kutta methods for stiff ordinary differential equations. *Math. Comput.*, 40, 1983.

- P. Dobson, J. M. Sanz-Serna, and K. C. Zygalakis. On the connections between optimization algorithms, Lyapunov functions, and differential equations: theory and insights. *arXiv:2305.08658*, 2023.
- M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems. *SIAM J. Optim.*, 28(3):2654–2689, 2018.
- W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Adv. Neural Inf. Process Syst.*, volume 28, pages 2845–2853, 2015a.
- W. Krichene, S. Krichene, and A. Bayen. Efficient Bregman projections onto the simplex. In *Proc. IEEE Conf. Decis. Control.*, pages 3291–3298. 2015b.
- L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM J. Optim.*, 26(1):57–95, 2016.
- H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, pages 333–354, 2018.
- A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.
- J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, second edition, 2006.
- B.T. Polyak. *Introduction to Optimization*. Optimization Software, New York, 1987.
- J. M. Sanz Serna and K. C. Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM J. Numer. Anal.*, 59(3):1542–1565, 2021.
- D. Scieur, A. d’Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Adv. Neural Inf. Process Syst.*, volume 30, pages 712–720, 2016.
- B. Shi, S. S. Du, W. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Adv. Neural Inf. Process Syst.*, volume 32, pages 5744–5752, 2019.
- B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Math. Program.*, 195(1):79–148, 2022.
- W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17(153):1–43, 2016.
- P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Technical report, MIT, Cambridge*, 2(3), 2008.

- A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proc. Natl. Acad. Sci.*, 113(47):E7351–E7358, 2016.
- A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22(113):1–34, 2021.
- S. J. Wright and B. Recht. *Optimization for Data Analysis*. Cambridge University Press, 2022.