A Lyapunov Approach to Accelerated First-Order Optimization In Continuous and Discrete Time

by

Walid Krichene

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Arts

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nikhil Srivastava, Chair
Professor Michael Christ
Professor Peter Bartlett
Professor Alexandre Bayen

Spring 2016

The thesis of Walid Krichene, titled A Lyapunov Approach to Accelerated First-Order Optimization In Continuous and Discrete Time, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

A Lyapunov Approach to Accelerated First-Order Optimization In Continuous and Discrete Time

Abstract

A Lyapunov Approach to Accelerated First-Order Optimization In Continuous and Discrete Time

by

Walid Krichene

Master of Arts in Mathematics

University of California, Berkeley

Professor Nikhil Srivastava, Chair

We study accelerated first-order dynamics for optimization, in continuous and discrete time. Combining the original continuous-time motivation of Nemirovski's mirror descent with a recent ODE interpretation of Nesterov's accelerated method, we propose a family of continuous-time dynamics for constrained minimization of convex functions with Lipschitz gradients, such that the solution trajectories are guaranteed to converge to the optimum at a $\mathcal{O}(1/t^2)$ rate. This family of continuous-time dynamics is naturally described as coupled dynamics of an unconstrained dual variable which accumulates gradient information, and a constrained primal variable which can be interpreted as an averaging of the mirrored dual trajectory. We then show that a large family of first-order accelerated methods can be obtained as a discretization of the ODE, and these methods converge at a $\mathcal{O}(1/k^2)$ rate. This connection between accelerated mirror descent and the ODE provides an intuitive approach to the design and analysis of accelerated first-order algorithms.

Contents

List of Figures

Chapter 1

Introduction

We consider a constrained convex optimization problem,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{X}, \end{array}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is convex and closed, f is a C^1 convex function, and its gradient, ∇f is assumed to be L_f-Lipschitz with respect to a pair of dual norms $(\|\cdot\|, \|\cdot\|_*)$, i.e. $\|\nabla f(x) - \nabla f(y)\|_* \leq L_f \|x - y\|$ for all $x, y \in \mathcal{X}$. Let $S \subset \mathcal{X}$ be the set of minimizers of f on $\mathcal{X}$, and $f^\star$ the value of f on S. Many convex optimization methods can be interpreted as the discretization of an ordinary differential equation, the solutions of which are guaranteed to converge to S. Perhaps the simplest such method is gradient descent for the unconstrained problem, given by the iteration $x^{(k+1)} = x^{(k)} - s\nabla f(x^{(k)})$ for some step size $s > 0$, which can be interpreted as the discretization of the ODE $\dot{X}(t) = -\nabla f(X(t))$, with discretization step s. The well-established theory of ordinary differential equations can provide guidance in the design and analysis of optimization algorithms, and has been used for unconstrained optimization [10, 9, 17], constrained optimization [31] and stochastic optimization [29]. It has also been applied to second-order methods for optimization, for example the Hessian-driven damping method in [4], and to more general problems, such as finding a zero of a monotone operator [2]. In particular, proving convergence of the solution trajectories of an ODE can often be achieved using simple and elegant Lyapunov arguments. The ODE can then be carefully discretized to obtain an optimization algorithm for which the convergence rate can be analyzed by using an analogous Lyapunov argument in discrete time.

In this thesis, we focus on two families of first-order methods: Nesterov's accelerated method [26], and Nemirovski's mirror descent method [23]. First-order methods have become increasingly important for large-scale optimization problems that arise in machine learning applications. Nesterov's accelerated method [26] has been applied to many problems and extended in a number of ways, see for example [27, 25, 24, 5]. The mirror descent method also provides an important generalization of the gradient descent method to constrained, non-Euclidean geometries, as discussed in [23, 6], and has many applications in convex optimization [8, 7, 14, 19], as well as online learning [11, 13]. An intuitive understanding of

these methods is of particular importance for the design and analysis of optimization algorithms. Although Nesterov's method has been notoriously hard to explain intuitively [18], progress has been made recently: in [32], Su et al. give an ODE interpretation of Nesterov's method. However, this interpretation is restricted to the original method [26], and does not apply to constrained, non-Euclidean geometries. In [1], Allen-Zhu and Orecchia give another interpretation of Nesterov's method, as performing, at each iteration, a convex combination of a mirror step and a gradient step. Although it covers a broader family of algorithms (including non-Euclidean geometries), this interpretation still requires an involved analysis, and lacks the simplicity and elegance of ODEs. We provide a new interpretation which has the benefits of both approaches: we show that a broad family of accelerated methods (which includes those studied in [32] and [1]) can be obtained as a discretization of a simple ODE, which is guaranteed to converges in $\mathcal{O}(1/t^2)$. This provides a unified interpretation, which could potentially simplify the design and analysis of first-order accelerated methods for constrained convex optimization.

The continuous-time interpretation [32] of Nesterov's method and the continuous-time motivation of mirror descent [23] both rely on a Lyapunov argument. They are reviewed in Chapter 2. By combining these ideas, we propose, in Chapter 3, a candidate Lyapunov function $V(X(t), Z(t), t)$ that depends on two state variables: $X(t)$, which evolves in the primal space $E = \mathbb{R}^n$ (more precisely, $X(t)$ evolves in the feasible set $\mathcal{X} \subset E$), and $Z(t)$, which evolves in the dual space E^*, and we design coupled dynamics of (X, Z) to guarantee that $\frac{d}{dt}V(X(t), Z(t), t) \leq 0$. Such a function is said to be a Lyapunov function in reference to [22]; see also [20]. This derivation leads to a new family of ODE systems, given by

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X) \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X) \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases} \tag{1.1}$$

where r is a positive parameter, and ψ^* is a distance generating function on E^* with Lipschitz gradient. We prove the existence and uniqueness of the solution to (1.1) in Theorem 1. Then we prove in Thereom 2, using the Lyapunov function V, that the solution trajectories are such that $f(X(t)) - f^\star = \mathcal{O}(1/t^2)$.

In Chapter 4, we derive equivalent formulations of the ODE. In particular, we show that the second equation is equivalent, in integral form, to $X(t) = \int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau / \int_0^t w(\tau)d\tau$, where $w(\tau) = \tau^{r-1}$, so that the primal variable X can be interpreted as a weighted average of the mirrored dual trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0, t]$. Here, the function $\nabla\psi^*$ is a mapping from the dual space E^* to the feasible set $\mathcal{X}$, which guarantees that the primal variable remains in the feasible set, by convexity. Motivated by this averaging interpretation, we generalize the ODE to allow other weight functions $w(\tau)$, and derive sufficient conditions on $w(\tau)$ for $t \mapsto V(X(t), Z(t), t)$ to be a Lyapunov function. We also give a formulation in terms of a second-order ODE only involving the primal variable X, which can be interpreted as describing the dynamics of a damped particle in a conservative potential field.

In Chapter 5, we give a discretization of these continuous-time dynamics, and obtain a family of accelerated mirror descent methods, for which we prove the same $\mathcal{O}(1/k^2)$ convergence rate (Theorem 5) using a Lyapunov argument analogous to the continuous-time case. We give, as an example, a new accelerated method on the simplex, which can be viewed as performing, at each step, a convex combination of two entropic projections with different step sizes. This ODE interpretation of accelerated mirror descent gives new insights and allows us to extend recent results, such as the adaptive restarting heuristics proposed by O'Donoghue and Candès in [28], which are known to empirically improve the convergence rate. We also propose a new restarting scheme for which the restarting condition is defined on the dual space. We test these methods on numerical examples in Chapter 6 and give comments on their theoretical and empirical performance.

Chapter 2

Nemirovski's mirror descent and Nesterov's accelerated method in continuous time

Proving convergence of the solution trajectories of an ODE often involves a Lyapunov argument. For example, to prove convergence of the solutions of the unconstrained gradient descent ODE, $\dot{X}(t) = -\nabla f(X(t))$, consider the Lyapunov function $D(x^\star, X(t)) = \frac{1}{2}\|X(t) - x^\star\|_2^2$ for some minimizer $x^\star \in S$. Then the time derivative of $D(x^\star, X(t))$ is given by

$$\begin{aligned}\frac{d}{dt}D(x^\star, X(t))(t) &= \left\langle \dot{X}(t), X(t) - x^\star \right\rangle \\ &= \langle -\nabla f(X(t)), X(t) - x^\star \rangle \\ &\leq -(f(X(t)) - f^\star),\end{aligned}$$

where the last inequality is by convexity of f. Integrating the inequality, we have $D(x^\star, X(t)) - D(x^\star, X(0)) \leq tf^\star - \int_0^t f(X(\tau))d\tau$, thus by Jensen's inequality, $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^\star \leq \frac{1}{t}\int_0^t f(X(\tau))d\tau - f^\star \leq \frac{D(x^\star, X(0))}{t}$, which proves that $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right)$ converges to the optimum at a $\mathcal{O}(1/t)$ rate. In fact one can also show the convergence of $X(t)$ to the set of minimizers S. Define the distance to the set of minimizers, $D(S, x) = \inf_{x^\star \in S} D(x^\star, x)$ (this is a continuous function of x whenever S is compact). We have shown that $D(x^\star, X(t))$ is a decreasing function of t for all $x^\star \in S$. Since $t \mapsto D(S, X(t))$ is the pointwise infimum of non-negative, decreasing functions, it is also decreasing and non-negative, therefore it has a limit as $t \to \infty$, and its limit is necessarily 0: By contradiction, suppose that its limit is strictly positive. Then there exists $d > 0$ and $T \geq 0$ such that for all $t \geq T$, $D(S, X(t)) > d$, and by continuity of f and $D(S, \cdot)$, $\delta \triangleq \inf_{\{x: D(S,x) > d\}} f(x) - f^* > 0$. Thus for all $t \geq T$, and

for all $x^\star \in S$,

$$\frac{d}{dt} D(x^\star, X(t)) \leq f^\star - f(X) \leq -\delta$$

Integrating, we would have $D(x^\star, X(t)) \leq D(x^\star, X(T)) - (t-T)\delta$ for all $t \geq T$, which contradicts the fact that D is non-negative. This proves that $D(S, X(t))$ converges to 0.

2.1 Mirror descent ODE

The previous argument was extended by Nemirovski and Yudin in [23] to a family of methods called mirror descent. The idea is to start from a non-negative function, then to design dynamics for which that function is a Lyapunov function. Nemirovski and Yudin argue that one can replace the Lyapunov function $D(x^\star, X(t)) = \frac{1}{2}\|X(t) - x^\star\|_2^2$ (used in gradient descent) by a function defined on the dual space, $D_{\psi^*}(Z(t), z^\star)$, where $Z(t) \in E^*$ is a dual variable for which we will design the dynamics, and the corresponding trajectory in the primal space is $X(t) = \nabla\psi^*(Z(t))$ and $x^\star = \nabla\psi^*(z^\star)$. Here, E^* is the dual space, i.e. the space of linear functionals on E (in our case, since $E = \mathbb{R}^n$, E^* can also be identified with $\mathbb{R}^n$, but we make this distinction since, conceptually, the spaces E and E^* are different), and ψ^* is a convex function assumed to be finite and differentiable on all of E^*, and such that $\nabla\psi^*$ maps from E^* to $\mathcal{X}$. Such a function ψ^* can be obtained by taking the Fenchel conjugate of a strongly convex function ψ with effective domain $\mathcal{X}$; See Appendix A for a more detailed discussion of duality properties of ψ and ψ^*, and the operator $\nabla\psi^*$, which we refer to as the mirror operator.

The function $D_{\psi^*}(\cdot,\cdot)$ is the Bregman divergence associated with ψ^*, given as follows: for all $z, y \in E^*$,

$$D_{\psi^*}(z, y) = \psi^*(z) - \psi^*(y) - \langle \nabla\psi^*(y), z - y\rangle .$$

The function ψ^* is said to be ℓ-strongly convex w.r.t. $\|\cdot\|_*$ if $D_{\psi^*}(z,y) \geq \frac{\ell}{2}\|z-y\|_*^2$ for all y, z, and it is said to be L-smooth (w.r.t. the norm $\|\cdot\|_*$) if $\nabla\psi^*$ is L-Lipschitz (equivalently, D_{ψ^*} is L-smooth if $D_{\psi^*}(z,y) \leq \frac{L}{2}\|z-y\|_*^2$, as shown in the appendix).

By definition of the Bregman divergence, we have

$$\begin{aligned}
\frac{d}{dt} D_{\psi^*}(Z(t), z^\star) &= \frac{d}{dt}\left(\psi^*(Z(t)) - \psi^*(z^\star) - \langle \nabla\psi^*(z^\star), Z(t) - z^\star\rangle\right) \\
&= \left\langle \nabla\psi^*(Z(t)) - \nabla\psi^*(z^\star), \dot{Z}(t)\right\rangle \\
&= \left\langle X(t) - x^\star, \dot{Z}(t)\right\rangle .
\end{aligned}$$

Therefore, if the dual variable Z obeys the dynamics $\dot{Z} = -\nabla f(X)$, then

$$\frac{d}{dt} D_{\psi^*}(Z(t), z^\star) = -\langle \nabla f(X(t)), X(t) - x^\star\rangle \leq -(f(X(t)) - f^\star)$$

and by the same argument as in the gradient descent ODE, $D_{\psi^*}(Z(t), z^\star)$ is a Lyapunov function and $f\left(\frac{1}{t}\int_0^t X(\tau)d\tau\right) - f^\star$ converges to 0 at a $\mathcal{O}(1/t)$ rate. The mirror descent ODE system can be summarized by

$$\begin{cases} X = \nabla\psi^*(Z) \\ \dot{Z} = -\nabla f(X) \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases} \tag{2.1}$$

This is illustrated in Figure 2.1.

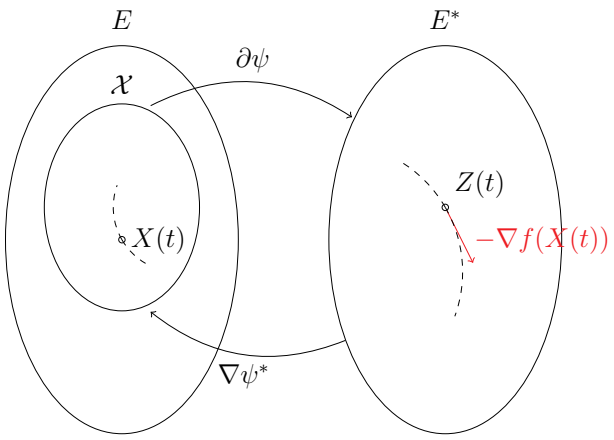


Figure 2.1: Illustration of the mirror descent ODE. The dual variable Z evolves in the (unconstrained) dual space E^*, and follows the flow of $-\nabla f(X(t))$. The primal trajectory $X(t)$ is obtained by applying the mirror operator $\nabla\psi^*$ to the dual trajectory $Z(t)$.

Note that since $X = \nabla\psi^*(Z)$, and the mirror operator $\nabla\psi^*$ maps into $\mathcal{X}$ by assumption, the solution trajectory $X(t)$ remains in $\mathcal{X}$. Therefore, the mirror descent ODE is a natural generalization of gradient descent to constrained optimization problems: if one can construct a mirror operator $\nabla\psi^*$ which maps into $\mathcal{X}$, the solution is guaranteed to remain in $\mathcal{X}$. We also observe that the unconstrained gradient descent ODE can be obtained as a special case of the mirror descent ODE (2.1) by taking $\psi^*(z) = \frac{1}{2}\|z\|_2^2$, for which $\nabla\psi^*$ is the identity, in which case X and Z coincide.

The family of mirror descent methods can then be obtained by discretizing the ODE (2.1), and can be analyzed by using an analogous Lyapunov function in discrete time [23]. The mirror descent method is of particular importance in convex optimization, since the appropriate choice of Bregman divergence D_{ψ^*} can lead to improving the dependence of the convergence rate on the dimension of the space, see for example [23] and [8].

2.2 ODE interpretation of Nesterov's accelerated method

In [32], Su et al. show that Nesterov's accelerated method [26] can be interpreted as a discretization of a second-order differential equation, given by

$$\begin{cases} \ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0, \\ X(0) = x_0,\ \dot{X}(0) = 0. \end{cases} \tag{2.2}$$

The argument uses the following function (up to reparameterization), $\mathcal{E}(t) = \frac{t^2}{r^2}(f(X) - f^\star) + \frac{1}{2}\|X + \frac{t}{r}\dot{X} - x^\star\|^2$, which is proved to be a Lyapunov function for the ODE (2.2) whenever $r \geq 2$. Since $\mathcal{E}$ is decreasing along trajectories of the system, it follows that for all $t > 0$, $\mathcal{E}(t) \leq \mathcal{E}(0) = \frac{1}{2}\|x_0 - x^\star\|^2$, therefore $f(X(t)) - f^\star \leq \frac{r^2}{t^2}\mathcal{E}(t) \leq \frac{r^2}{t^2}\mathcal{E}(0) \leq \frac{r^2}{t^2}\frac{\|x_0 - x^\star\|^2}{2}$, which proves that $f(X(t))$ converges to $f^\star$ at a $\mathcal{O}(1/t^2)$ rate.

One should note in particular that the squared Euclidean norm is used in the definition of $\mathcal{E}$ and, as a consequence, discretizing the ODE (2.2) leads to a family of unconstrained, Euclidean accelerated methods. In the next chapter, we show that by combining this argument with Nemirovski's idea of using a general Bregman divergence as a Lyapunov function, we can construct a much more general family of ODE systems which have the same $\mathcal{O}(1/t^2)$ convergence guarantee. And by discretizing the resulting dynamics, we obtain a general family of accelerated methods that are not restricted to the Euclidean geometry.

Chapter 3

Accelerated Mirror Descent in Continuous-time

3.1 Lyapunov design of the dynamics

Let $\|\cdot\|_*$ be a reference norm on the dual space E^*, and let ψ^* be a distance generating function on E^*, assumed to be L_{ψ^*}-smooth with respect to $\|\cdot\|_*$. Consider the function

$$V(X,Z,t) = \frac{t^2}{r^2}(f(X) - f^\star) + D_{\psi^*}(Z, z^\star) \tag{3.1}$$

where Z is a dual variable for which we will design the dynamics, and $z^\star$ is its value at equilibrium. Taking the time-derivative of $V(X(t), Z(t), t)$, we have

$$\frac{d}{dt}V(X(t),Z(t),t) = \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\left\langle \nabla f(X), \dot{X}\right\rangle + \left\langle \dot{Z}, \nabla\psi^*(Z) - \nabla\psi^*(z^\star)\right\rangle$$

Assume that $\dot{Z} = -\frac{t}{r}\nabla f(X)$. Then, the time-derivative becomes

$$\frac{d}{dt}V(X(t),Z(t),t) = \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}\left\langle \nabla f(X), -\frac{t}{r}\dot{X} + \nabla\psi^*(Z) - \nabla\psi^*(z^\star)\right\rangle.$$

Therefore, if X satisfies $X + \frac{t}{r}\dot{X} = \nabla\psi^*(Z)$, and $\nabla\psi^*(z^\star) = x^\star$, then,

$$\begin{aligned}\frac{d}{dt}V(X(t),Z(t),t) &= \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}\left\langle \nabla f(X), X - x^\star\right\rangle \\ &\leq \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}(f(X) - f^\star) \\ &\leq -t\frac{r-2}{r^2}(f(X) - f^\star)\end{aligned} \tag{3.2}$$

and it follows that V is a Lyapunov function whenever $r \geq 2$. The proposed ODE system is then given by

$$\begin{cases} \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t}{r}\nabla f(X), \\ X(0) = x_0,\ Z(0) = z_0,\ \text{with } \nabla\psi^*(z_0) = x_0. \end{cases}$$

In the Euclidean case, taking $\psi^*(z) = \frac{1}{2}\|z\|_2^2$, we have $\nabla\psi^*(z) = z$, thus $Z = X + \frac{t}{r}\dot{X}$, and the ODE system is equivalent to $\frac{d}{dt}\left(X + \frac{t}{r}\dot{X}\right) = -\frac{t}{r}\nabla f(X)$, i.e. $\frac{t}{r}\ddot{X} + \frac{r+1}{r}\dot{X} + \frac{t}{r}\nabla f(X) = 0$, which is equivalent to the ODE (2.2) studied in [32], which we recover as a special case.

It is also important to observe that since $\nabla\psi^*$ maps into $\mathcal{X}$, then any primal solution $X(t)$ is viable (i.e. remains in the feasible set $\mathcal{X}$). Intuitively, since $\dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X)$, then $\dot{X}(t)$ always points inside the feasible set $\mathcal{X}$. In particular, whenever $X(t)$ is on the boundary of E, $\dot{X}(t)$ towards the interior of $\mathcal{X}$, thus guaranteeing that X remains in $\mathcal{X}$. This argument is made more precise in the proof of Theorem 1.

3.2 Existence, uniqueness and viability of the solution

First, we prove existence and uniqueness of a solution to the ODE system (4.3), defined for all $t > 0$. By assumption, both ∇f and $\nabla\psi^*$ are Lipschitz-continuous functions. Unfortunately, due to the $\frac{r}{t}$ term in the expression of $\dot{X}$, the function $(X, Z, t) \mapsto (\dot{X}, \dot{Z})$ is not Lipschitz at $t = 0$. However, one can work around this by considering a sequence of approximating ODEs, similarly to the argument used in [32].

Theorem 1. *Suppose f is C^1, and that ∇f is L_f-Lipschitz, and let $x_0 \in \mathcal{X}$. Then the accelerated mirror descent ODE system (1.1) with initial condition (x_0, z_0) has a unique solution (X, Z), in $C^1([0,\infty), \mathbb{R}^n)$. Furthermore, the primal solution X is viable, that is $X(t) \in \mathcal{X}$ for all $t \geq 0$.*

We first show existence and uniqueness of a solution on any given interval $[0, T]$. Let $\delta > 0$, and consider the smoothed ODE system

$$\begin{cases} \dot{X} = \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t}{r}\nabla f(X), \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0. \end{cases} \tag{3.3}$$

Since the functions $(X, Z) \mapsto -\frac{t}{r}\nabla f(X)$ and $(X, Z) \mapsto \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X)$ are Lipschitz for all $t \in [0, T]$, by the Cauchy-Lipschitz theorem (Theorem 2.5 in [33]), the system (3.3) has a unique solution (X_δ, Z_δ) in $C^1([0, T])$. In order to show the existence of a solution to the original ODE, we use the following property of the solution to the smoothed ODE (proved in the appendix).

Lemma 1. *Let $t_0 = \frac{2}{\sqrt{L_f L_{\psi^*}}}$. Then the family of solutions $\left((X_\delta, Z_\delta)|_{[0,t_0]}\right)_{\delta \le t_0}$ is equi-Lipschitz-continuous and uniformly bounded. More precisely,*

$$\|\dot{Z}_\delta(t)\| \le \frac{3t}{r}\|\nabla f(x_0)\|,$$
$$\|\dot{X}_\delta(t)\| \le \frac{(3+r)L_{\psi^*}t}{2}\|\nabla f(x_0)\|.$$

Proof of existence. Consider the family of solutions $((X_{\delta_i}, Z_{\delta_i}), \delta_i = t_0 2^{-i})_{i\in\mathbb{N}}$ restricted to $[0, t_0]$. By Lemma 1, this family is equi-Lipschitz-continuous and uniformly bounded, thus by the Arzelà-Ascoli theorem, there exists a subsequence $((X_{\delta_i}, Z_{\delta_i}))_{i\in\mathcal{I}}$ that converges uniformly on $[0, t_0]$. Let $(\bar{X}, \bar{Z})$ be its limit. Then we prove that $(\bar{X}, \bar{Z})$ is a solution to the original ODE (1.1) on $[0, t_0]$.

First, since for all $i \in \mathcal{I}$, $X_{\delta_i}(0) = x_0$ and $Z_{\delta_i}(0) = z_0$, it follows that

$$\bar{X}(0) = \lim_{i\to\infty, i\in\mathcal{I}} X_{\delta_i}(0) = x_0,$$
$$\bar{Z}(0) = \lim_{i\to\infty, i\in\mathcal{I}} Z_{\delta_i}(0) = z_0,$$

thus $(\bar{X}, \bar{Z})$ satisfies the initial conditions. Next, let $t_1 \in (0, t_0)$, and let $(\tilde{X}, \tilde{Z})$ be the solution of the ODE (1.1) on $t \ge t_1$, with initial condition $(\bar{X}(t_1), \bar{Z}(t_1))$. Since $(X_{\delta_i}(t_1), Z_{\delta_i}(t_1))_{i\in\mathcal{I}} \to (\bar{X}(t_1), \bar{Z}(t_1))$ as $i \to \infty$, then by continuity of the solution w.r.t. initial conditions, we have that for some $\epsilon > 0$, $X_{\delta_i} \to \tilde{X}$ uniformly on $[t_1, t_1 + \epsilon)$. But we also have $X_{\delta_i} \to \bar{X}$ uniformly on $[0, t_0]$, therefore $\bar{X}$ and $\tilde{X}$ coincide on $[t_1, t_1 + \epsilon)$, therefore $\bar{X}$ satisfies the ODE on $[t_1, t_1 + \epsilon)$. And since t_1 is arbitrary in $(0, t_0)$, this concludes the proof of existence. □

Proof of uniqueness. It suffices to prove uniqueness on an open neighborhood of 0, since away from 0, uniqueness is guaranteed by the Cauchy-Lipschitz theorem.

Let (X, Z) and $(\bar{X}, \bar{Z})$ be two solutions of the ODE (1.1), and let $\Delta_Z = Z - \bar{Z}$ and $\Delta_X = X - \bar{X}$. Then Δ_X, Δ_Z are C^1, and we have

$$\begin{cases} \dot{\Delta}_Z = -\frac{t}{r}\left(\nabla f(X) - \nabla f(\bar{X})\right) \\ \dot{\Delta}_X = \frac{r}{t}\left(\nabla\psi^*(Z) - \nabla\psi^*(\bar{Z}) - \Delta_X\right) \\ \Delta_Z(0) = \Delta_X(0) = 0 \end{cases}$$

Let $A(t) = \sup_{[0,t]} \frac{\|\dot{\Delta}_Z(u)\|}{u}$, and $B(t) = \sup_{[0,t]} \|\Delta_X\|$. Note that $B(t)$ is finite since Δ_X is continuous on $[0, t]$. The finiteness of $A(t)$ will be established below. We have

$$\|\dot{\Delta}_Z(t)\| = \frac{t}{r}\|\nabla f(X(t)) - \nabla f(\bar{X}(t))\| \le \frac{L_f t}{r}\|\Delta_X(t)\| \le \frac{L_f t}{r}B(t).$$

Dividing by t and taking the supremum, we have

$$A(t) \le \frac{L_f}{r}B(t). \tag{3.4}$$

Next, since $t^r\dot{\Delta}_X + rt^{r-1}\Delta_X = rt^{r-1}\left(\nabla\psi^*(Z) - \nabla\psi^*(\bar{Z})\right)$, we have

$$\frac{d}{dt}(t^r\Delta_X) = rt^{r-1}\left(\nabla\psi^*(Z) - \nabla\psi^*(\bar{Z})\right).$$

Therefore, integrating and taking norms

$$\begin{aligned}
t^r\|\Delta_X(t)\| &\leq \int_0^t r\tau^{r-1}\|\nabla\psi^*(Z(\tau)) - \nabla\psi^*(\bar{Z}(\tau))\|d\tau \\
&\leq rt^{r-1}\int_0^t L_{\psi^*}\|\Delta_Z(\tau)\|d\tau \\
&\leq L_{\psi^*}rt^{r-1}A(t)\int_0^t \frac{\tau^2}{2}d\tau \\
&= \frac{L_{\psi^*}rt^{r-1}t^3A(t)}{6},
\end{aligned}$$

where we used the fact that $\|\Delta_Z(\tau)\| = \|\int_0^\tau \dot{\Delta}_Z(u)du\| \leq \int_0^\tau uA(t)du = A(t)\frac{\tau^2}{2}$. Dividing by t^r and taking the supremum,

$$B(t) \leq \frac{L_{\psi^*}rt^2}{6}A(t). \tag{3.5}$$

Combining (3.4) and (3.5), we have $A(t) \leq \frac{L_fL_{\psi^*}t^2}{6}A(t)$. It follows that $A(t) = 0$ for $0 \leq t < \sqrt{\frac{6}{L_fL_{\psi^*}}}$, which in turn implies that $B(t) = 0$ on the same interval. This concludes the proof. □

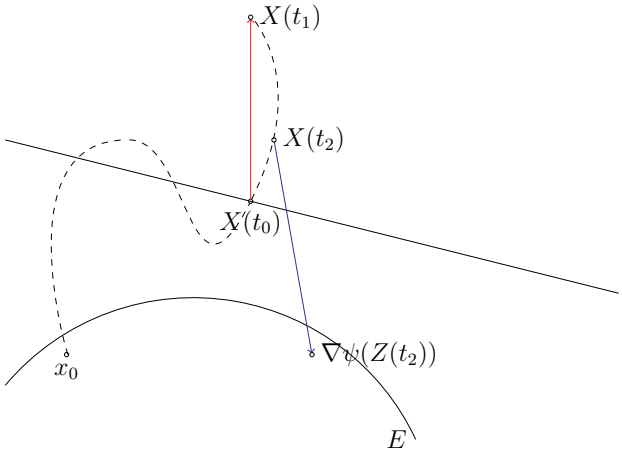


Figure 3.1: Illustration of the proof of viability.

Proof of viability. We now prove that the primal solution X remains in $\mathcal{X}$ for all t. Intuitively, since $\dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X)$, the derivative $\dot{X}$ will point towards $\mathcal{X}$, keeping $X(t)$ inside the feasible set.

Suppose by contradiction that there exists $t_1 > 0$ such that $x_1 = X(t_1) \notin \mathcal{X}$. Since $\mathcal{X}$ is convex and compact, by the separation theorem, there exists a hyperplane that strictly separates x_1 and $\mathcal{X}$. That is, there exists $u, a \in \mathbb{R}^n$ such that $\langle u, x_1 - a\rangle > 0$ and $\langle u, x - a\rangle < 0$ for all $x \in \mathcal{X}$. Now let $d(x) = \langle u, x - a\rangle$. Since the solution trajectory $X(t)$ is C^1, $t \mapsto d(X(t))$ is also C^1, and $\dot{d}(X(t)) = \left\langle u, \dot{X}(t)\right\rangle$.

We have $d(X(0)) < 0$ (since $x_0 \in \mathcal{X}$) and $d(X(t_1)) > 0$, thus there exists t_0 such that $d(X(t_0)) = 0$ and $d(X(t)) > 0$ for all $t \in (t_0, t_1]$, that is, t_0 is the last time $X(t)$ crosses the separating hyperplane (t_0 is simply $\sup\{t : d(X(t)) \le 0\}$). Then by definition, $d(X(t_1)) - d(X(t_0)) > 0$, but by Taylor's theorem, there exists $t_2 \in [t_0, t_1]$ such that

$$\begin{aligned} d(X(t_1)) - d(X(t_0)) = \dot{d}(X(t_2)) &= \left\langle u, \dot{X}(t_2)\right\rangle \\ &= \frac{r}{t}\left\langle u, \nabla\psi^*(Z(t_2)) - X(t_2)\right\rangle \\ &= \frac{r}{t}\left(d(\nabla\psi^*(Z(t_2))) - d(X(t_2))\right) < 0 \end{aligned}$$

since $\nabla\psi^*(Z(t_2)) \in \mathcal{X}$. This is a contradiction, which concludes the proof. □

3.3 Convergence rate

It is straightforward to establish the convergence rate of the function values.

Theorem 2. *Suppose that ∇f and $\nabla\psi^*$ are Lipschitz. Let $(X(t), Z(t))$ be the solution to the accelerated mirror descent ODE (1.1) with $r \ge 2$. Then for all $t > 0$,*

$$f(X(t)) - f^\star \le \frac{r^2 D_{\psi^*}(z_0, z^\star)}{t^2} \tag{3.6}$$

Furthermore, if $r > 2$, then $\int_0^\infty t(f(X(t)) - f^\star)dt$ is finite.

Proof. By construction of the ODE, we have $V(X(t), Z(t), t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$ is a Lyapunov function. It follows that for all $t > 0$,

$$\frac{t^2}{r^2}(f(X(t)) - f^\star) \le V(X(t), Z(t), t) \le V(x_0, z_0, 0) = D_{\psi^*}(z_0, z^\star),$$

which proves the first inequality. Furthermore, we have that

$$\frac{d}{dt}V(X(t), Z(t), t) \le -\frac{r-2}{r^2}t(f(X(t)) - f^\star),$$

thus, integrating from 0 to T and rearranging, we have

$$\int_0^T t(f(X(t)) - f^\star)dt \leq \frac{r^2}{r-2}V(x_0, z_0, 0) = \frac{r^2}{r-2}D_{\psi^*}(z_0, z^\star),$$

which proves the second part of the claim. □

Remark 1. *The second part of the theorem indicates that the convergence rate is in fact better than $\Omega(1/t^2)$. Indeed, if $f(X(t)) - f^\star \geq \frac{c}{t^2}$ for some positive constant c, then $\int_1^T t(f(X(t)) - f^\star)dt \geq c \ln T$, which would contradict the theorem. We also observe that, although it seems from the bound* (3.6) *that smaller values of the parameter r are better, the upper bound on the integral diverges as r approaches 2, which indicates that smaller values of r are not necessarily better. In Section 4.3, we will give another interpretation of the parameter r as a damping coefficient, and we will further discuss its effect on convergence.*

3.4 Restarting the ODE in the strongly convex case

When the objective function is strongly convex with a known parameter, faster convergence can be obtained by restarting the ODE at fixed intervals. That is, for some period T which depends on the function parameters, if we call $T_k = kT$, we can define a trajectory (X, Z) to be the union of the solutions, on each interval $[T_k, T_{k+1})$, of the ODE

$$\begin{cases} \dot{Z} = -\frac{t-T_k}{r}\nabla f(X) \\ \dot{X} = \frac{r}{t-T_k}\left(\nabla\psi^*(Z) - X\right) \\ X(T_k) = x_{T_k},\ Z(T_k) = z_{T_k},\ \text{where } x_{T_k} = X(T_k^-), \text{ and } \nabla\psi^*(z_{T_k}) = x_{T_k}. \end{cases} \tag{3.7}$$

That is, we solve a sequence of ODEs, one on each interval, and choose the initial conditions at the start of each interval so that the primal trajectory is continuous. The dual variable Z is reinitialized at T_k in order to satisfy $\nabla\psi^*(Z(T_k)) = X(T_k^-)$. As we will see in Section 4.1, the primal variable can be written as the weighted average $X(t) = \frac{\int_{T_k}^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_{T_k}^t w(\tau)d\tau}$ with $w(\tau) = \tau^{r-1}$. Thus, setting $\nabla\psi^*(Z(T_k)) = X(T_k^-)$ ensures continuity of the primal trajectory (but the dual trajectory is in general, discontinuous at T_k). Also note that as a consequence of restarting, $\dot{X}(T_k) = 0$.

We now present a simple restarting strategy in the strongly convex case.

Theorem 3. *Suppose that ψ^* is ℓ_{ψ^*}-strongly convex. Then the restarted ODE with period $T = \sqrt{\frac{er^2}{\ell_{\psi^*}\ell_f}}$ guarantees that for all $k \geq 0$ and all $t \in [T_k + 1, T_{k+1}]$,*

$$f(X(t)) - f^\star \leq T^2 e^{-\frac{t}{T}}(f(x_0) - f^\star).$$

In other words, $f(X)$ converges exponentially to $f^\star$, at a rate $\frac{1}{T} = \sqrt{\frac{er^2}{\ell_{\psi^*}\ell_f}}$. Note, however, that this method requires previous knowledge of the strong convexity parameter ℓ_f.

Proof. We have for all $t \geq T_k$,

$$\begin{aligned} f(X(t)) - f^\star &\leq \frac{r^2}{(t-T_k)^2} D_{\psi^*}(Z(T_k), z^\star) && \text{by Theorem 2,} \\ &\leq \frac{r^2}{\ell_{\psi^*} 2(t-T_k)^2} \|X(T_k) - x^\star\|^2 && \text{by strong convexity of } \psi^*, \\ &\leq \frac{r^2}{\ell_{\psi^*}\ell_f (t-T_k)^2} (f(X(T_k)) - f^\star) && \text{by strong convexity of } f. \end{aligned}$$

Thus using our choice $T = \sqrt{\frac{er^2}{\ell_{\psi^*}\ell_f}}$, we have $f(X(T_{k+1})) - f^\star \leq \frac{f(X(T_k))-f^\star}{e}$, so by induction $f(X(T_k)) - f^\star \leq e^{-k}(f(x_0) - f^\star)$, and for $t \in [T_k + 1, T_{k+1}]$,

$$\begin{aligned} f(X(t)) - f^\star &\leq \frac{r^2}{\ell_{\psi^*}\ell_f} \left(f(X(T_k)) - f^\star\right) \\ &\leq \frac{r^2}{\ell_{\psi^*}\ell_f} e^{-k} \left(f(x_0) - f^\star\right) \\ &\leq \frac{er^2}{\ell_{\psi^*}\ell_f} e^{-\frac{t}{T}} \left(f(x_0) - f^\star\right) && \text{since } t \leq (k+1)T \\ &= T^2 e^{-\frac{t}{T}} (f(x_0) - f^\star), \end{aligned}$$

□

3.5 Non-differentiable objective functions

In this section, we consider the case in which the objective function is non-differentiable. One such case of particular interest is composite optimization, in which the objective function can be decomposed into the sum of two terms $f = f_1 + f_2$ where f_1 is differentiable with Lipschitz gradient, and f_2 is a general convex function; this model covers many problems in machine learning, such as ℓ_1-regularized regression, and many algorithms have been developed for composite optimization in discrete time, for example [24], as well as continuous time, for example [4]. In this section, we discuss how the Lyapunov argument can be extended to non-differentiable functions. More precisely, assume that f is a closed and proper convex function (not necessarily differentiable), and denote by $\partial f(x)$ the subdifferential of f at x (a closed and convex set). A natural way to extend the ODE (1.1) to this non-differentiable case is to replace the dual differential equation $\dot{Z}(t) = -\frac{t}{r}\nabla f(X(t))$ by the differential inclusion $\dot{Z}(t) \in -\frac{t}{r}\partial f(X(t))$. As we will see, this may not suffice to guarantee that the energy function V decreases along continuous solution trajectories. As observed by [32], the directional derivative $f'(X; \dot{X})$ plays a central role in deriving the correct dynamics in the non-differentiable case.

The directional derivative of f at x in the direction y is defined by

$$f'(x;y) = \lim_{\epsilon \to 0,\ \epsilon > 0} \frac{f(x+\epsilon y) - f(x)}{\epsilon},$$

where the limit can be $+\infty$. It exists at any point x in the effective domain of f, and is a positively homogeneous convex function of y, see Theorem 23.1 in [30]. Additionally, we have the following connection between the directional derivative and the subdifferential: By Theorem 23.4 in [30] we have that for all x in the interior of the domain of f (denoted $\operatorname{int} \operatorname{dom} f$), $\partial f(x)$ is a non-empty bounded set, and

$$f'(x;y) = \sup_{g \in \partial f(x)} \langle g, y \rangle . \tag{3.8}$$

Thus we can associate to $f'(x;y)$ the set of subgradients which achieve the maximum (the supremum is attained since $\partial f(x)$ is a compact set in this case). We will denote this set

$$d(x;y) = \arg\max_{g \in \partial f(x)} \langle g, y \rangle .$$

Theorem 4. *Consider the energy function $t \mapsto V(X(t), Z(t), t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$, and suppose that $(X(t), Z(t))$ is a continuous and differentiable solution trajectory of the ODE*

$$\begin{cases} \dot{Z} \in -\frac{t}{r} d(X, \dot{X}) \\ \dot{X} = \frac{r}{t}(\nabla \psi^*(Z) - X). \end{cases}$$

Then the energy function is differentiable and $\frac{d}{dt} V(X(t), Z(t), t) \leq 0$.

Since the energy function is decreasing, any continuous and differentiable solution will satisfy $f(X(t)) - f^\star = \mathcal{O}(1/t^2)$ by a similar argument to Theorem 2. Note however that we do not discuss existence of such solutions in this case.

Proof. To prove that the energy function is differentiable, consider the difference quotient, defined for $\epsilon > 0$,

$$\begin{aligned} \Delta_t(\epsilon) &= \frac{V(t+\epsilon) - V(t)}{\epsilon} \\ &= \frac{t^2}{r^2} \frac{f(X(t+\epsilon)) - f(X(t))}{\epsilon} + \frac{2t+\epsilon}{r^2}(f(X(t+\epsilon)) - f^\star) + \frac{D_{\psi^*}(Z(t+\epsilon), z^\star) - D_{\psi^*}(Z(t), z^\star)}{\epsilon}. \end{aligned}$$

Using the fact that a convex function is locally Lipschitz (so that $f(x + o(\epsilon)) = f(x) + o(\epsilon)$), and that $D_{\psi^*}(Z(t), z^\star)$ is differentiable, we have

$$\begin{aligned} \Delta_t(\epsilon) &= \frac{t^2}{r^2} \frac{f(X(t) + \epsilon \dot{X}(t)) + o(\epsilon) - f(X(t))}{\epsilon} + \frac{2t+\epsilon}{r^2}(f(X(t)) + o(1) - f^\star) + \frac{d}{dt} D_{\psi^*}(Z(t), z^\star) + o(1) \\ &= \frac{t^2}{r^2} \frac{f(X(t) + \epsilon \dot{X}(t)) - f(X(t))}{\epsilon} + \frac{2t}{r^2}(f(X(t)) - f^\star) + \frac{d}{dt} D_{\psi^*}(Z(t), z^\star) + o(1). \end{aligned} \tag{3.9}$$

The derivative of the Bregman divergence in (3.9) is

$$\frac{d}{dt}D_{\psi^*}(Z(t), z^\star) = \left\langle \dot{Z}(t), \nabla\psi^*(Z(t)) - \nabla\psi^*(z^\star) \right\rangle = \left\langle \dot{Z}(t), X(t) + \frac{t}{r}\dot{X}(t) - x^\star \right\rangle.$$

The first term in (3.9) converges, as $\epsilon \to 0$, to $f'(X; \dot{X})$. Combining the two limits, we have that the limit of $\Delta_t(\epsilon)$ exists and

$$\lim_{\epsilon\to 0,\ \epsilon>0} \Delta_t(\epsilon) = \frac{t^2}{r^2} f'(X(t); \dot{X}(t)) + \frac{2t}{r^2}(f(X(t)) - f^\star) + \left\langle \dot{Z}(t), X(t) + \frac{t}{r}\dot{X}(t) - x^\star \right\rangle,$$

and if we let $\dot{Z}(t) = -\frac{t}{r}g(t)$, then

$$\lim_{\epsilon\to 0,\ \epsilon>0} \Delta_t(\epsilon) \leq \frac{t^2}{r^2}\left(f'(X; \dot{X}) - \left\langle g, \dot{X} \right\rangle\right) + \frac{t}{r}(f(X) - f^\star - \langle g, X - x^\star\rangle), \tag{3.10}$$

where we used the assumption that $r \geq 2$. Note that if $\dot{Z}$ satisfies the differential inclusion $\dot{Z}(t) \in -\frac{t}{r}\partial f(X(t))$ (in other words, $g(t)$ is a subgradient of f at $X(t)$), then the second term in inequality (3.10) is non-positive by definition of a subgradient, but the first term $f'(X; \dot{X}) - \left\langle g, \dot{X} \right\rangle$ is non-negative by (3.8), and one cannot conclude that the energy is decreasing. This motivates our choice of the subgradient. Indeed, when $\dot{Z}(t) \in -\frac{t}{r}d(X; \dot{X})$ (in other words, $g(t)$ is a subgradient of f at $X(t)$ *that maximizes the linear functional* $\langle \cdot, \dot{X}(t)\rangle$), the first term in inequality (3.10) is non-positive, therefore $\lim_{\epsilon\to 0,\ \epsilon>0} \Delta_t(\epsilon) \leq 0$, which concludes the proof. □

Chapter 4

Equivalent formulations

In this chapter, we give equivalent formulations of ODE (1.1), which allows us to give different interpretations.

4.1 Averaging formulation

Starting from the equation $\dot{X} = \frac{r}{t}(\nabla\psi^*(Z(t)) - X(t))$, we can multiply both sides by $\frac{t^r}{r}$ and rearrange to obtain $\frac{t^r}{r}\dot{X}(t) + t^{r-1}X(t) = t^{r-1}\nabla\psi^*(Z(t))$. Integrating from 0 to t, and observing that $\frac{t^r}{r}\dot{X}(t) + t^{r-1}X(t)$ is the time derivative of $\frac{t^r}{r}X(t)$, we have

$$\frac{t^r}{r}X(t) = \int_0^t \tau^{r-1}\nabla\psi^*(Z(\tau))d\tau.$$

Finally, dividing by $\frac{t^r}{r}$, we have

$$X(t) = \frac{r}{t^r}\int_0^t \tau^{r-1}\nabla\psi^*(Z(\tau))d\tau = \frac{\int_0^t \tau^{r-1}\nabla\psi^*(Z(\tau))d\tau}{\int_0^t \tau^{r-1}d\tau}.$$

Therefore the primal variable $X(t)$ can be interpreted as a weighted average of the trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0,t]$, with time-varying weights $w(\tau) = \tau^{r-1}$. This interpretation formalizes a connection between acceleration and averaging, as observed in [16] for the unconstrained quadratic case. This also provides an intuitive interpretation of the parameter r: it controls the weights in the expression of X. A higher value of r puts larger weights on the recent points $\nabla\psi^*(Z(t))$.

The accelerated mirror descent ODE can then be written in the equivalent form:

$$\begin{cases} \dot{Z} = -\eta(t)\nabla f\left(X(t)\right), \ \eta(t) = \frac{t}{r} \\ X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}, \ w(\tau) = \tau^{r-1} \\ X(0) = x_0, \ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0 \end{cases} \tag{4.1}$$

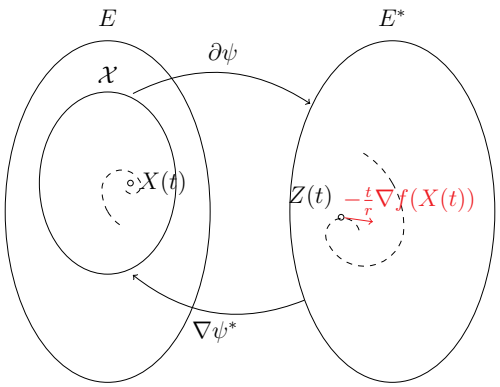


Figure 4.1: Illustration of the accelerated mirror descent ODE. The dual variable Z evolves in the dual space E^*, and accumulates gradient at a rate $\eta(t) = \frac{t}{r}$, and the primal variable $X(t)$ is obtained by averaging the mirrored trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0, t]$.

Here Z is a dual variable which accumulates the gradient of f, at a rate $\eta(t) = \frac{t}{r}$, and X is a weighted average of the "mirrored" dual trajectory $\nabla\psi^*(Z(\tau))$, $\tau \in [0, t]$, with weight function $w(\tau) = \tau^{r-1}$. This is illustrated in Figure 4.1. We also note that since $\nabla\psi^*(Z(\tau))$ remains in $\mathcal{X}$ for all τ, so does X, by convexity of the feasible set $\mathcal{X}$. This provides an alternate, simple proof of the viability of the solution (last part of Theorem 1).

Generalized weighting

Motivated by the averaging representation (4.1), it is natural to ask whether a different averaging scheme can guarantee the same $\mathcal{O}(1/t^2)$ convergence rate. One way to achieve this is to start from a given averaging scheme, $X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)}$ for some positive, continuous, increasing function of time $w : \mathbb{R}_+ \to \mathbb{R}_+$, and design the dynamics of Z in order to guarantee that the energy function

$$V(X(t), Z(t), t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star),$$

is decreasing. Taking the time-derivative of V, we have

$$\frac{d}{dt}V(X(t), Z(t), t) = \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\left\langle \nabla f(X), \dot{X}\right\rangle + \left\langle \dot{Z}, \nabla\psi^*(Z) - x^\star\right\rangle.$$

Let $W(t) = \int_0^t w(\tau)d\tau$. Then using the assumption on X, we have $X(t)W(t) = \int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau$, and taking time derivatives, we have $\dot{X}W + Xw = w\nabla\psi^*(Z)$, that is, for all $t > 0$, $\dot{X} = \frac{w}{W}(\nabla\psi^*(Z) - X)$. Using this expression of $\dot{X}$, the energy derivative becomes

$$\frac{d}{dt}V(X(t), Z(t), t) = \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\frac{w}{W}\left\langle \nabla f(X), \nabla\psi^*(Z) - X\right\rangle + \left\langle \dot{Z}, \nabla\psi^*(Z) - x^\star\right\rangle.$$

In order to eliminate the $\nabla\psi^*(Z)$ term, we choose $\dot{Z} = -\frac{w(t)}{W(t)}\frac{t^2}{r^2}\nabla f(X)$. Then,

$$\begin{aligned}\frac{d}{dt}V(X(t),Z(t),t) &= \frac{2t}{r^2}(f(X)-f^\star) - \frac{t^2}{r^2}\frac{w(t)}{W(t)}\langle\nabla f(X), X-x^\star\rangle \\ &\leq \frac{2t}{r^2}(f(X)-f^\star) - \frac{t^2}{r^2}\frac{w(t)}{W(t)}(f(X)-f^\star) \\ &= -\frac{t^2}{r^2}\left(\frac{w(t)}{W(t)} - \frac{2}{t}\right)(f(X)-f^\star).\end{aligned}$$

Therefore, if w satisfies the condition

$$\frac{w(t)}{W(t)} \geq \frac{2}{t}, \tag{4.2}$$

then V is a Lyapunov function. The resulting generalized accelerated mirror descent dynamics can be given in these two equivalent forms:

$$\begin{cases} \dot{Z} = -\frac{w(t)}{W(t)}\frac{t^2}{r^2}\nabla f(X) \\ X = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{W(t)} \\ \nabla\psi^*(Z(0)) = x_0 \end{cases} \qquad \begin{cases} \dot{Z} = -\frac{w(t)}{W(t)}\frac{t^2}{r^2}\nabla f(X) \\ \dot{X} = \frac{w(t)}{W(t)}(\nabla\psi^*(Z) - X) \\ X(0) = x_0,\ \nabla\psi^*(Z(0)) = x_0 \end{cases} \tag{4.3}$$

where the functions $w(t)$ and $W(t) = \int_0^t w(\tau)d\tau$ satisfy the following assumption:

Assumption 1. *The weight function $w : \mathbb{R}_+ \to \mathbb{R}_+$ is assumed to be non-negative, continuous and increasing, and such that the function $a(t) = \frac{w(t)}{W(t)}$ is decreasing, bounded below by $\frac{2}{t}$ for all $t > 0$.*

Proving the existence and uniqueness of the solution for this generalized weighting scheme requires additional assumptions on the weight function (to guarantee that the derivatives are well behaved in the neighborhood of 0). One sufficient condition is that $a(t) = o(t^{-\frac{3}{2}})$ at 0 (the argument is a straightforward generalization of the proof of Theorem 1).

Next, we describe a method to construct a weight function w that satisfies the conditions of Assumption 1. Let $\frac{w(t)}{W(t)} = a(t)$. Then writing $\frac{d}{dt}\ln W(t) = a(t)$ and integrating from 1 to t, we have $\frac{W(t)}{W(1)} = e^{\int_1^t a(\tau)d\tau}$, and $\frac{w(t)}{w(1)} = \frac{a(t)}{a(1)}e^{\int_1^t a(\tau)d\tau}$. The time derivative of $w(t)$ is $w(1)e^{\int_1^t a(\tau)d\tau}(a'(t)+a^2(t))$. Therefore the conditions of Assumption 1 are satisfied whenever $w(t)$ is of the form

$$w(t) = w(1)\frac{a(t)}{a(1)}e^{\int_1^t a(\tau)d\tau},$$

and $a : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-negative, decreasing, differentiable function with $a(t) \geq \frac{2}{t}$ and $a(t) = o\left(\frac{1}{t^2}\right)$ as t tends to 0. Note that the expression of w is defined up to the constant $w(1)$, which reflects the fact that the conditions of Assumption 1 are scale-invariant (if the conditions hold for a function w, then they hold for αw for all $\alpha > 0$).

Example 1. *Take* $a(t) = \frac{\beta}{t}$ *with* $\beta \geq 2$. *Then* $a(t)$ *satisfies the conditions discussed above, and* $w(t) = \frac{a(t)}{a(1)} e^{\int_1^t a(\tau)d\tau} = \frac{\beta/t}{\beta} e^{\beta \ln t} = t^{\beta-1}$. *The dynamics are then given by one of the following equivalent systems:*

$$\begin{cases} \dot{Z} = -\frac{\beta t}{r^2} \nabla f(X) \\ X = \frac{\int_0^t \tau^{\beta-1} \nabla \psi^*(Z(\tau)) d\tau}{\int_0^t \tau^{\beta-1} d\tau} \\ \nabla \psi^*(Z(0)) = x_0 \end{cases} \qquad \begin{cases} \dot{Z} = -\frac{\beta t}{r^2} \nabla f(X) \\ \dot{X} = \frac{\beta}{t}(\nabla \psi^*(Z) - X) \\ X(0) = x_0, \ \nabla \psi^*(Z(0)) = x_0. \end{cases}$$

4.2 Primal representation and damped nonlinear oscillators

In this section, we will assume that ψ^* is twice differentiable on all of E^*, and we will denote its Hessian at a point $z \in E^*$ by $\nabla^2 \psi^*(z)$, defined as $\nabla^2 \psi^*(z)_{i,j} = \frac{\partial^2 \psi^*(z)}{\partial z_i \partial z_j}$. This assumption is not particularly restrictive, see Appendix A for examples. Writing $\frac{t}{r}\dot{X} + X = \nabla \psi^*(Z)$ and taking the time-derivative, we have

$$\frac{t}{r}\ddot{X} + \frac{1}{r}\dot{X} + \dot{X} = \nabla^2 \psi^*(Z)\dot{Z} = -\frac{t}{r}\nabla^2 \psi^*(Z) \nabla f(X).$$

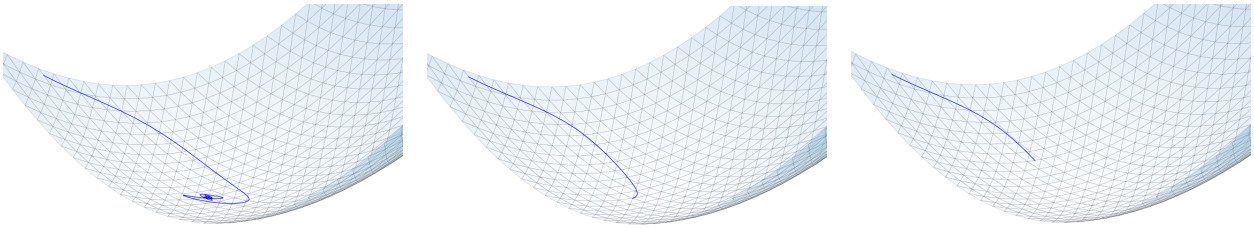
Multiplying both sides by $\frac{r}{t}$, we have

$$\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla^2 \psi^*(Z) \nabla f(X) = 0. \tag{4.4}$$

The initial condition for $\dot{X}$ is $\dot{X}(0) = 0$. To prove this, one can argue that for all $\delta > 0$, the solution to the smoothed ODE (3.3) satisfies $\dot{X}_\delta(0) = \frac{r}{\delta}(\nabla \psi^*(z_0) - x_0) = 0$, thus $\dot{X}(0)$ is also equal to zero since the solution X is a limit point of the equi-Lipschitz family of solutions (X_δ).

ODE (4.4) can be interpreted as a damped nonlinear oscillator: If we ignore the Hessian term (in other words when $\nabla^2 \psi^*(Z)$ is the identity, which corresponds to the unconstrained Euclidean case), then the ODE becomes $\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0$. It can be interpreted as describing the evolution of a particle with position X, velocity $\dot{X}$ and acceleration $\ddot{X} = -\nabla f(X) - \frac{r+1}{t}\dot{X}$. The first term is a conservative force due to the scalar potential f, and the second term is a dissipative force proportional to the velocity, which can be thought of as a viscous friction term. Some properties of this system have been recently studied in [3]. Note that the damping constant $\frac{r+1}{t}$ is time-dependent, and vanishes as time tends to infinity. The parameter r can then be interpreted as a damping coefficient. Intuitively, the larger r, the more energy is dissipated. This is illustrated in Figure 4.2, which shows the solution trajectory of the ODE on a finite time interval, in a simplex-constrained example, with different values of r. The Hessian term $\nabla^2 \psi^*(Z)$ is a non-linear transformation that applies to the gradient, in order to keep the trajectory in the feasible set. Remarkably, this

transformation is not static, it depends on the value of the dual variable, hence varies with time. Intuitively, whenever $\nabla\psi^*(Z)$ approaches the (relative) boundary of the feasible set, the term $\nabla^2\psi^*(Z)$ should transform the gradient so that it points inside the feasible set.



(a) $r = 2$ (b) $r = 20$ (c) $r = 200$

Figure 4.2: Solution trajectories of the accelerated mirror descent ODE on a finite time interval $t \in [0, T]$, for simplex-constrained quadratic minimization, with different values of the parameter r. Larger values of r result in more energy dissipation, and suppress oscillations, but because the time-horizon is finite, too much energy dissipation means that the trajectory does not make enough progress within $[0, T]$, as can be seen in plot (c). This example shows that the "best damping" is not necessarily obtained for smaller values of r, as one could think from the bound of Theorem 2.

In order to have a fully primal representation of the ODE, we seek to eliminate the dual variable Z. By duality of the subdifferentials (Appendix A), we have $x = \nabla\psi^*(z)$ if and only if $z \in \partial\psi(x)$. We have $\dot{X} = \frac{t}{r}(\nabla\psi^*(Z) - X)$, thus $Z \in \partial\psi(\frac{t}{r}\dot{X} + X)$. Note that, in general, X and $\dot{X}$ do not entirely determine Z when ψ is not differentiable. However, as discussed in Proposition 3 in the Appendix, if we assume that ψ is the restriction to $\mathcal{X}$ of a differentiable function Ψ, i.e. $\psi(x) = \Psi(x) + \delta_{\mathcal{X}}(x)$, then, $\nabla^2\psi^*(\partial\psi(x))$ does not depend on the choice of subgradient in $\partial\psi(x)$, therefore

$$\nabla^2\psi^*(Z) = \nabla^2\psi^*\left(\partial\psi\left(\frac{t}{r}\dot{X} + X\right)\right) = \nabla^2\psi^*\left(\nabla\Psi\left(\frac{t}{r}\dot{X} + X\right)\right).$$

Plugging this expression in (4.4), we have

$$\begin{cases} \ddot{X} + \frac{r+1}{t}\dot{X} + \nabla^2\psi^* \circ \nabla\Psi(\frac{t}{r}\dot{X} + X)\nabla f(X) = 0 \\ X(0) = x_0, \ \dot{X}(0) = 0 \end{cases} \tag{4.5}$$

This ODE has been studied independently by Wibisono and Wilson in [35]. In particular, they give an interpretation of the ODE as the Euler-Lagrange equation associated to a particular Lagrangian function.

If we define $\tilde{Z} = \frac{t}{r}\dot{X} + X = \nabla\psi^*(Z)$, then this ODE can also be written more concisely as

$$\begin{cases} \dot{\tilde{Z}} + \frac{t}{r}\nabla^2\psi^* \circ \nabla\Psi(\tilde{Z})\nabla f(X) = 0 \\ X(t) = \frac{\int_0^t \tau^{r-1}\tilde{Z}(\tau)d\tau}{\int_0^t \tau^{r-1}d\tau} \\ X(0) = \tilde{Z}(0) = x_0 \end{cases} \tag{4.6}$$

where X and $\tilde{Z}$ are both primal variables.

A similar derivation can be made for the mirror descent ODE (2.1), as follows: writing $X = \nabla\psi^*(Z)$ and taking the time derivative, we have $\dot{X} = \nabla^2\psi^*(Z)\dot{Z} = -\nabla^2\psi^*(Z)\nabla f(X) = -\nabla^2\psi^* \circ \nabla\Psi(X)\nabla f(X)$, which leads to the ODE

$$\begin{cases} \dot{X} + \nabla^2\psi^* \circ \nabla\Psi(X)\nabla f(X) = 0 \\ X(0) = x_0 \end{cases} \tag{4.7}$$

The operator $\nabla^2\psi^* \circ \nabla\Psi$ appears in both primal representations (4.5) and (4.7). For some choices of distance generating functions, $\nabla^2\psi^* \circ \nabla\Psi$ has a simple expression. In the next section, we give three such examples: one for unconstrained Euclidean optimization, one for positive-orthant-constrained optimization, and one for simplex-constrained optimization, using the negative entropy as a distance generating function.

4.3 Examples of accelerated mirror descent dynamics

Unconstrained Euclidean optimization Suppose that $\mathcal{X} = \mathbb{R}^n$ and take $\psi^*(z) = \frac{1}{2}\|z\|_2^2$. Then $\nabla\psi^*(z) = z$ and the Hessian at any point is equal to the identity, thus

$$\nabla^2\psi^* \circ \nabla\psi(x) = I_n$$

for all $x \in \mathbb{R}^n$. Therefore the mirror descent ODE (4.7) reduces to the gradient descent ODE $\dot{X} + \nabla f(X) = 0$, and the accelerated mirror descent ODE (4.5) reduces to the Nesterov ODE studied in [32], $\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0$. The latter can be interpreted as a damped non-linear oscillator (the non-linearity comes from the gradient term), with vanishing damping coefficient $\frac{r+1}{t}$.

Positive-orthant-constrained dynamics Now suppose that $\mathcal{X}$ is the positive orthant $\mathbb{R}^n_+$, and consider the negative entropy function $\psi(x) = \sum_i x_i \ln x_i$. Then its dual is $\psi^*(z) = \sum_i e^{z_i - 1}$, and

$$\nabla\psi(x)_i = 1 + \ln x_i, \qquad \nabla^2\psi^*(z)_{i,j} = \delta_i^j e^{z_i - 1}.$$

Thus for all $x \in \mathbb{R}^n_+$,

$$\nabla^2\psi^* \circ \nabla\psi(x) = \mathrm{diag}(x).$$

Therefore, the mirror descent ODE in its primal form (4.7) reduces to

$$\begin{cases} \forall i, \dot{X}_i = -X_i \nabla f(X)_i \\ X(0) = x_0 \end{cases}$$

and the accelerated mirror descent ODE (4.6) can be written in one of its equivalent forms:

$$\begin{cases} \forall i,\ \ddot{X}_i + \frac{r+1}{t}\dot{X}_i = -(\frac{t}{r}\dot{X}_i + X_i)\nabla_i f(X) \\ X(0) = x_0,\ \dot{X}(0) = 0 \end{cases} \qquad \begin{cases} \forall i,\ \dot{\tilde{Z}}_i = -\tilde{Z}_i \nabla_i f(X) \\ X(t) = \frac{\int_0^t \tau^{r-1}\tilde{Z}(\tau)d\tau}{\int_0^t \tau^{r-1}d\tau} \\ X(0) = \tilde{Z}(0) = x_0 \end{cases}$$

For the mirror descent ODE, one can verify that the solution remains in the positive orthant since $\dot{X}$ tends to 0 as X_i approaches the boundary of the feasible set. Similarly for the accelerated version, $\dot{\tilde{Z}}$ tends to 0 as $\tilde{Z}$ approaches the boundary, thus $\tilde{Z}$ remains feasible, and so does X by convexity.

Simplex-constrained entropic optimization: the replicator dynamics. Now suppose that $\mathcal{X}$ is the n-simplex, $\mathcal{X} = \Delta = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$. Consider the distance-generating function $\psi(x) = \Psi(x) + \delta_{\mathcal{X}}(x)$, where Ψ is the negative entropy function, $\Psi(x) = \sum_{i=1}^n x_i \ln x_i$. Then its dual is $\psi^*(z) = \ln(\sum_{i=1}^n e^{z_i})$ (see Appendix A.4), defined on E^*, and we have

$$\nabla\Psi(x)_i = 1 + \ln x_i, \qquad \nabla\psi^*(z)_i = \frac{e^{z_i}}{\sum_k e^{z_k}}, \qquad \nabla^2\psi^*(z)_{ij} = \frac{\delta^j_i e^{z_i}}{\sum_k e^{z_k}} - \frac{e^{z_i}e^{z_j}}{(\sum_k e^{z_k})^2}.$$

This example can be used to illustrate the role of the Hessian term in equation (4.4). Suppose that $\nabla\psi^*(Z)$ approaches the relative boundary of the feasible set, say $e^{Z_{i_0}}$ approaches 0. Then $(\nabla^2\psi^*(Z)\nabla f(X))_{i_0} = \frac{e^{Z_{i_0}}}{\sum_k e^{Z_k}}\left(\nabla_{i_0} f(X) - \left\langle \nabla f(X), \frac{e^Z}{\sum_{e^{Z_k}}} \right\rangle\right)$, also approaches 0. Figure 4.3 displays the vector field $\nabla^2\psi^*(Z)\nabla f(X)$ (which we can think of as the modified potential) for different values of Z.

To derive the primal representation of the ODE, it is simple to calculate the expression

$$\nabla^2\psi^* \circ \nabla\Psi(x)_{ij} = \frac{\delta^j_i x_i}{\sum_k x_k} - \frac{x_i x_j}{(\sum_k x_k)^2} = \delta^j_i x_i - x_i x_j.$$

Therefore, the mirror descent ODE in its primal form (4.7) reduces to the following

$$\begin{cases} \forall i,\ \dot{X}_i + X_i\left(\nabla_i f(X) - \langle X, \nabla f(X)\rangle\right) = 0 \\ X(0) = x_0 \end{cases} \tag{4.8}$$

This ODE is known as the replicator dynamics in the evolutionary game theory literature, see for example [34]. It is used to model large population dynamics, that is, one considers

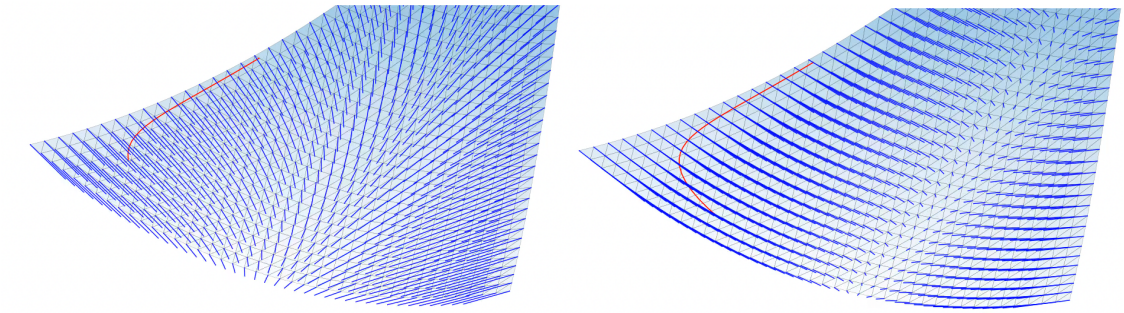


Figure 4.3: Vector field $X \mapsto \nabla^2\psi^*(Z)\nabla f(X)$ for different values of Z (taken along a solution trajectory for an example problem with solution on the relative boundary of the simplex). As $\nabla\psi^*(Z)$ approaches the relative boundary, the vector field changes in such a way that the component that is orthogonal to the boundary vanishes.

a population of players, and a finite action set $\{1, \dots, n\}$, such that at time t, $X_i(t)$ is the proportion of players who adopt action i. Then $\nabla f_i(X)$ is the cost of action i given the distribution X. This can be used to model competition for resources in, e.g., biological systems [34] or traffic systems [15]: if a resource i is used by a larger proportion of players, its cost increases.

The ODE is called replicator as it can be obtained using a simple model of adaptive play as follows: at time t, players are randomly matched in pairs, and if their current actions are, respectively, i and j, then the first player will switch to j (i.e. replicate the action of the second player) with a rate proportional to $\nabla_j f(X) - \nabla_i f(X)$, and similarly for the second player. As a consequence, the rate of increase of X_i is simply the sum over all actions j of $X_i X_j$ (the probability of the match (i, j)) multiplied by the difference in costs $\nabla_j f(X) - \nabla_i f(X)$, i.e.

$$\begin{aligned}
\dot{X}_i &= \sum_{j=1}^{n} X_i X_j (\nabla_j f(X) - \nabla_i f(X)) \\
&= X_i \left(\sum_{j=1}^{n} X_j (\nabla_j f(X) - \nabla_i f(X)) \right) \\
&= X_i \left(\langle X, \nabla f(X) \rangle - \nabla_i f(X) \right).
\end{aligned}$$

Similarly, the accelerated mirror descent ODE in its primal forms (4.5) and (4.6) reduces

to the following

$$\begin{cases} \forall i,\ \ddot{X}_i + \frac{r+1}{t}\dot{X}_i + (\frac{t}{r}\dot{X}_i + X_i)\left(\nabla_i f(X) - \left\langle \tilde{Z}, \nabla f(X)\right\rangle\right) = 0 \\ X(0) = x_0,\ \dot{X}(0) = 0 \end{cases}$$

$$\begin{cases} \forall i, \dot{\tilde{Z}}_i + \tilde{Z}_i\left(\nabla_i f(X) - \left\langle \tilde{Z}, \nabla f(X)\right\rangle\right) = 0 \\ X = \frac{\int_0^t \tau^{r-1}\tilde{Z}(\tau)d\tau}{\int_0^t \tau^{r-1}d\tau} \\ \tilde{Z}(0) = x_0 \end{cases}$$

Using the second primal form, the accelerated version of the replicator dynamics can be interpreted as follows: the usual replicator update is performed on the variable $\tilde{Z}$, but the gradient is evaluated at $X(t)$, which is obtained by averaging.

Chapter 5

Discrete Optimization

5.1 Forward-backward Euler discretization

Next, we show that with a careful discretization of the continuous-time dynamics, we can obtain a general family of accelerated mirror descent methods for constrained optimization. Using a mixed Euler scheme (forward in the Z variable, and backward in the X variable), see e.g. Chapter 2 in [12], we can discretize the ODE system (1.1) using a step size $\sqrt{s}$ as follows. Given a solution (X, Z) of the ODE (1.1), let $t_k = k\sqrt{s}$, and $x^{(k)} = X(t_k) = X(k\sqrt{s})$. Approximating $\dot{X}(t_k)$ with $\frac{X(t_k+\sqrt{s})-X(t_k)}{\sqrt{s}}$, and, similarly, $\dot{Z}(t_k)$ with $\frac{Z(t_k+\sqrt{s})-Z(t_k)}{\sqrt{s}}$, we propose the discretization

$$\begin{cases} \frac{z^{(k+1)}-z^{(k)}}{\sqrt{s}} = -\frac{k\sqrt{s}}{r}\nabla f(x^{(k)}), \\ \frac{x^{(k+1)}-x^{(k)}}{\sqrt{s}} = \frac{r}{(k+1)\sqrt{s}}\left(\nabla\psi^*(z^{(k+1)}) - x^{(k+1)}\right). \end{cases} \tag{5.1}$$

The second equation can be rewritten as $x^{(k+1)} = \left(x^{(k)} + \frac{r}{k+1}\nabla\psi^*(z^{(k+1)})\right) / \left(1 + \frac{r}{k+1}\right)$ (note the independence on s, due to the time-scale invariance of the first ODE). In other words, $x^{(k+1)}$ is a convex combination of $\nabla\psi^*(z^{(k)})$ and $x^{(k)}$ with coefficients $\lambda_{k+1} = \frac{r}{r+k+1}$ and $1 - \lambda_{k+1} = \frac{k+1}{r+k+1}$. To summarize, our first discrete scheme can be written as

$$\begin{cases} z^{(k+1)} = z^{(k)} - \frac{ks}{r}\nabla f(x^{(k)}), \\ x^{(k+1)} = \lambda_{k+1}\nabla\psi^*(z^{(k+1)}) + (1-\lambda_{k+1})x^{(k)}, \ \lambda_k = \frac{r}{r+k}. \end{cases} \tag{5.2}$$

Note that since $\nabla\psi^*$ maps into $\mathcal{X}$, starting from $x^{(0)} \in \mathcal{X}$ guarantees that $x^{(k)}$ remains in $\mathcal{X}$ for all k.

An equivalent form of the mirror descent update When the primal distance generating function ψ is differentiable, one has that $z = \nabla\psi(\tilde{z})$ if and only if $\tilde{z} = \nabla\psi^*(z)$. Thus, letting $\tilde{z}^{(k)} = \nabla\psi^*(z^{(k)})$, we can rewrite the dual variable update $z^{(k)}$ in terms of $\tilde{z}^{(k)}$ as

follows:

$$\begin{aligned}\tilde{z}^{(k+1)} &= \nabla\psi^*(z^{(k+1)}) \\ &= \nabla\psi^*\left(z^{(k)} - \frac{ks}{r}\nabla f(x^{(k)})\right) \\ &= \nabla\psi^*\left(\nabla\psi(\tilde{z}^{(k)}) - \frac{ks}{r}\nabla f(x^{(k)})\right)\end{aligned}$$

Then by duality of subgradients, (Theorem 23.5 in [30]), we have that $\nabla\psi^*(z) = \arg\max_{x\in\mathcal{X}} \langle z, x\rangle - \psi(x) = \arg\min_{x\in\mathcal{X}} \psi(x) - \langle z, x\rangle$, thus

$$\begin{aligned}\tilde{z}^{(k+1)} &= \arg\min_{x\in\mathcal{X}} \psi(x) - \left\langle \nabla\psi(\tilde{z}^{(k)}) - \frac{ks}{r}\nabla f(x^{(k)}), x\right\rangle \\ &= \arg\min_{x\in\mathcal{X}} \frac{ks}{r}\left\langle \nabla f(x^{(k+1)}), x\right\rangle + D_\psi(x, \tilde{z}^{(k)}).\end{aligned}$$

In this case, the discretization can be written purely in terms of the primal variables $x^{(k)}$ and $\tilde{z}^{(k)}$ as follows

$$\begin{cases} x^{(k+1)} = \lambda_{k+1}\tilde{z}^{(k+1)} + (1-\lambda_{k+1})x^{(k)},\ \lambda_k = \frac{r}{r+k}, \\ \tilde{z}^{(k+1)} = \arg\min_{x\in\mathcal{X}} \frac{ks}{r}\left\langle \nabla f(x^{(k+1)}), x\right\rangle + D_\psi(x, \tilde{z}^{(k)}). \end{cases}$$

We will eventually modify this scheme in order to be able to prove the desired $\mathcal{O}(1/k^2)$ convergence rate. However, we start by analyzing this version. Motivated by the continuous-time Lyapunov function (3.1), and using the correspondence $t \approx k\sqrt{s}$, consider the potential function, defined for $k \geq 1$,

$$E^{(k)} = \frac{k^2 s}{r^2}(f(x^{(k-1)}) - f^\star) + D_{\psi^*}(z^{(k)}, z^\star).$$

Then we have

$$\begin{aligned}E^{(k+1)} - E^{(k)} &= \frac{(k+1)^2 s}{r^2}(f(x^{(k)}) - f^\star) - \frac{k^2 s}{r^2}(f(x^{(k-1)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\ &= \frac{k^2 s}{r^2}(f(x^{(k)}) - f(x^{(k-1)})) + \frac{s(1+2k)}{r^2}(f(x^{(k)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star).\end{aligned}$$

And through simple algebraic manipulation, the last term can be bounded as follows

$$\begin{aligned}&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\ &= D_{\psi^*}(z^{(k+1)}, z^{(k)}) + \left\langle \nabla\psi^*(z^{(k)}) - \nabla\psi^*(z^\star), z^{(k+1)} - z^{(k)}\right\rangle \\ &= D_{\psi^*}(z^{(k+1)}, z^{(k)}) + \left\langle \frac{k}{r}(x^{(k)} - x^{(k-1)}) + x^{(k)} - x^\star, -\frac{ks}{r}\nabla f(x^{(k)})\right\rangle \\ &\leq D_{\psi^*}(z^{(k+1)}, z^{(k)}) + \frac{k^2 s}{r^2}(f(x^{(k-1)}) - f(x^{(k)})) + \frac{ks}{r}(f^\star - f(x^{(k)})).\end{aligned}$$

where the first equality is by definition of the Bregman divergence, the second equality is by the discretization (5.2), and the last inequality is by convexity of f. Therefore we have

$$E^{(k+1)} - E^{(k)} \leq -\frac{s[(r-2)k-1]}{r^2}(f(x^{(k)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^{(k)}),$$

where the first term is negative whenever $r \geq 3$, for all $k \geq 1$. Comparing this expression with the bound (3.2) on $\frac{d}{dt}V(X(t), Z(t), t)$ in the continuous-time case, we see that we obtain an analogous bound, except for the additional Bregman divergence term $D_{\psi^*}(z^{(k+1)}, z^{(k)})$, and we cannot immediately conclude that $E^{(k)}$ is a Lyapunov function. This can be remedied by a modification of the discretization scheme, described next.

5.2 Discrete-time accelerated mirror descent

In the expression (5.2) of $x^{(k+1)} = \lambda_k \nabla\psi^*(z^{(k+1)}) + (1-\lambda_k)x^{(k)}$, we propose to replace $x^{(k)}$ with $\tilde{x}^{(k+1)}$, obtained as a solution to a minimization problem

$$\tilde{x}^{(k+1)} = \arg\min_{x \in \mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), x \right\rangle + R(x, x^{(k)}),$$

where R is regularization function that satisfies the following assumptions: there exist $0 < \ell_R \leq L_R$ such that for all $x, x' \in E$, $\frac{\ell_R}{2}\|x - x'\|^2 \leq R(x, x') \leq \frac{L_R}{2}\|x - x'\|^2$.

In the Euclidean case, one can take R to be the squared Euclidean distance, $R(x, x') = \frac{\|x-x'\|_2^2}{2}$, in which case $\ell_R = L_R = 1$ and the $\tilde{x}$ update becomes a prox-update, or take $R(x, x') = D_\phi(x, x')$ for some distance generating function ϕ which is ℓ_R-strongly convex and L_R-smooth, in which case the $\tilde{x}$ update becomes a mirror update. The resulting method is summarized in Algorithm 1, and illustrated in Figure 5.1. This algorithm is a generalization of Allen-Zhu and Orecchia's interpretation of Nesterov's method in [1], where $x^{(k+1)}$ is a convex combination of a mirror descent update and a gradient descent update.

Algorithm 1 Accelerated mirror descent with distance generating function ψ^*, regularizer R, step size s, and parameter $r \geq 3$

1: Initialize $\tilde{x}^{(0)} = \tilde{z}^{(0)} = x_0$.
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad \tilde{z}^{(k+1)} = \arg\min_{\tilde{z} \in \mathcal{X}} \frac{ks}{r}\left\langle \nabla f(x^{(k)}), \tilde{z} \right\rangle + D_\psi(\tilde{z}, \tilde{z}^{(k)}) = \nabla\psi^*(\nabla\Psi(\tilde{z}^{(k)}) - \frac{ks}{r}\nabla f(x^{(k)}))$
$\quad$ (equivalently, $z^{(k+1)} = z^{(k)} - \frac{ks}{r}\nabla f(x^{(k)})$ and $\tilde{z}^{(k+1)} = \nabla\psi^*(z^{(k+1)})$)
4: $\quad \tilde{x}^{(k+1)} = \arg\min_{\tilde{x} \in \mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(k)})$
5: $\quad x^{(k+1)} = \lambda_{k+1}\tilde{z}^{(k+1)} + (1-\lambda_{k+1})\tilde{x}^{(k+1)}$, with $\lambda_k = \frac{r}{r+k}$
6: **end for**

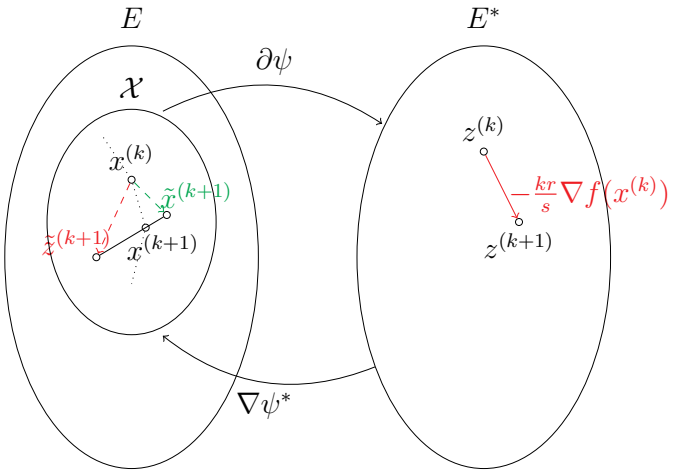


Figure 5.1: Illustration of the accelerated mirror descent method in discrete time. The dual variable $z^{(k)}$ is updated by taking a step in the direction of the negative gradient $-\nabla f(x^{(k)})$, with a rate $\frac{k}{r}$. The corresponding primal variable is $\tilde{z}^{(k+1)} = \nabla\psi^*(z^{(k+1)})$. The variable $\tilde{x}^{(k+1)}$ is obtained by performing a prox update from $x^{(k)}$, then $x^{(k+1)}$ is updated by taking a convex combination of $\tilde{z}^{(k+1)}$ and $\tilde{x}^{(k+1)}$.

5.3 Consistency of the modified scheme

One can show that given our assumptions on R, $\tilde{x}^{(k+1)} = x^{(k)} + \mathcal{O}(s)$. Indeed, we have

$$\begin{aligned}\frac{\ell_R}{2}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 &\le R(\tilde{x}^{(k+1)}, x^{(k)}) \le R(x^{(k)}, x^{(k)}) + \gamma s \left\langle \nabla f(x^{(k)}), x^{(k)} - \tilde{x}^{(k+1)} \right\rangle \\ &\le \gamma s \|\nabla f(x^{(k)})\|_* \|\tilde{x}^{(k+1)} - x^{(k)}\|\end{aligned}$$

therefore $\|\tilde{x}^{(k+1)} - x^{(k)}\| \le s\frac{2\gamma\|\nabla f(x^{(k)})\|_*}{\ell_R}$, which proves the claim. Using this observation, we can show that the modified discretization scheme is consistent with the original ODE (1.1), that is, the difference equations defining $x^{(k)}$ and $z^{(k)}$ converge, as s tends to 0, to the ordinary differential equations of the continuous-time system (1.1). The difference equations of Algorithm 1 are equivalent to (5.1) in which $x^{(k)}$ is replaced by $\tilde{x}^{(k+1)}$, i.e.

$$\begin{cases} \frac{z^{(k+1)} - z^{(k)}}{\sqrt{s}} = -\frac{k\sqrt{s}}{r}\nabla f(x^{(k)}) \\ \frac{x^{(k+1)} - \tilde{x}^{(k+1)}}{\sqrt{s}} = \frac{r}{(k+1)\sqrt{s}}(\nabla\psi^*(z^{(k+1)}) - x^{(k+1)}). \end{cases}$$

Now suppose there exist C^1 functions (X, Z), defined on $\mathbb{R}^+$, such that $X(t_k) \approx x^{(k)}$ and $Z(t_k) \approx z^{(k)}$ for $t_k = k\sqrt{s}$. Then, using the fact that $\tilde{x}^{(k)} = x^{(k)} + \mathcal{O}(s)$, we have $\frac{x^{(k+1)} - \tilde{x}^{(k+1)}}{\sqrt{s}} = \frac{x^{(k+1)} - x^{(k)}}{\sqrt{s}} + \mathcal{O}(\sqrt{s}) \approx \frac{X(t_k+\sqrt{s}) - X(t_k)}{\sqrt{s}} + \mathcal{O}(\sqrt{s}) = \dot{X}(t_k) + \mathcal{O}(\sqrt{s})$, and similarly,

$\frac{z^{(k+1)}-z^{(k)}}{\sqrt{s}} \approx \dot{Z}(t_k) + o(1)$, therefore the difference equation system can be written as

$$\begin{cases} \dot{Z}(t_k) + o(1) = -\frac{t_k}{r}\nabla f(X(t_k)) \\ \dot{X}(t_k) + \mathcal{O}(\sqrt{s}) = \frac{r}{t_k+\sqrt{s}}(\nabla\psi^*(Z(t_k+\sqrt{s})) - X(t_k+\sqrt{s})) \end{cases}$$

which converges to the ODE (1.1) as $s \to 0$.

5.4 Convergence rate

To prove convergence of the algorithm, consider the modified potential function

$$\tilde{E}^{(k)} = \frac{k^2 s}{r^2}(f(\tilde{x}^{(k)}) - f^\star) + D_{\psi^*}(z^{(k)}, z^\star).$$

Lemma 2. *If $\gamma \geq L_R L_{\psi^*}$ and $s \leq \frac{\ell_R}{2L_f\gamma}$, then for all $k \geq 0$,*

$$\tilde{E}^{(k+1)} - \tilde{E}^{(k)} \leq \frac{(2k+1-kr)s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star).$$

As a consequence, if $r \geq 3$, $\tilde{E}$ is a Lyapunov function for $k \geq 1$.

This lemma is proved in the appendix.

Theorem 5. *The discrete-time accelerated mirror descent Algorithm 1 with parameter $r \geq 3$ and step sizes $\gamma \geq \frac{L_R}{L_{\psi^*}}$, $s \leq \frac{\ell_R}{2L_f\gamma}$, guarantees that for all $k > 0$,*

$$f(\tilde{x}^{(k))}) - f^\star \leq \frac{r^2}{sk^2}\tilde{E}^{(1)} \leq \frac{r^2 D_{\psi^*}(z_0, z^\star)}{sk^2} + \frac{f(x_0) - f^\star}{k^2}.$$

Proof. The first inequality follows immediately from Lemma 2: $\frac{k^2 s}{r^2}(f(\tilde{x}^{(k))}) - f^\star) \leq \tilde{E}^{(k)} \leq \tilde{E}^{(1)}$. The second inequality follows from a simple bound on $\tilde{E}^{(1)}$, proved below. By Lemma 2 again, we have

$$\begin{aligned} \tilde{E}^{(1)} &\leq \tilde{E}^{(0)} + \frac{s}{r^2}(f(\tilde{x}^{(1)}) - f^\star) \\ &= D_{\psi^*}(z^{(0)}, z^\star) + \frac{s}{r^2}(f(\tilde{x}^{(1)}) - f^\star) \end{aligned}$$

and to conclude, it suffices to show that $f(\tilde{x}^{(1)}) \leq f(x^{(0)})$. Recall that $\tilde{x}^{(1)} = \arg\min_{\tilde{x}\in\mathcal{X}} \gamma s \left\langle \nabla f(x^{(0)}), \tilde{x}\right\rangle + R(\tilde{x}, x^{(0)})$, thus

$$\gamma s \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)}\right\rangle + R(\tilde{x}^{(1)}, x^{(0)}) \leq \gamma s \left\langle \nabla f(x^{(0)}), x^{(0)}\right\rangle \tag{5.3}$$

Therefore,

$$\begin{aligned}
&f(\tilde{x}^{(1)}) - f(x^{(0)}) \\
&\quad \leq \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} - x^{(0)} \right\rangle + \frac{L_f}{2}\|\tilde{x}^{(1)} - x^{(0)}\|^2 && \text{by Lemma 4} \\
&\quad \leq \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} - x^{(0)} \right\rangle + \frac{L_f}{\ell_R} R(\tilde{x}^{(1)}, x^{(0)}) && \text{by assumption on } R \\
&\quad \leq \left\langle \nabla f(x^{(0)}), \tilde{x}^{(1)} - x^{(0)} \right\rangle + \frac{1}{\gamma s} R(\tilde{x}^{(1)}, x^{(0)}) && \text{using that } \frac{L_f}{\ell_R} \leq \frac{1}{\gamma s}
\end{aligned}$$

and we conclude by applying (5.3). □

5.5 Example: accelerated entropic descent

We give an instance of Algorithm 1 for simplex-constrained problems. Suppose that $\mathcal{X} = \Delta^n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$ is the n-simplex. Taking ψ to be the negative entropy on Δ, we have for $x \in \mathcal{X}$, $z \in E^*$,

$$\psi(x) = \sum_{i=1}^n x_i \ln x_i, \quad \psi^*(z) = \ln\left(\sum_{i=1}^n e^{z_i}\right), \quad \nabla\psi(x)_i = 1 + \ln x_i, \quad \nabla\psi^*(z)_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}.$$

The resulting mirror descent update is a simple entropy projection and can be computed exactly in $\mathcal{O}(n)$ operations, and ψ^* can be shown to be 1-smooth w.r.t. $\|\cdot\|_\infty$, see for example [6, 8]. For the second update, we take $R(x, y) = D_\phi(x, y)$ where ϕ is a smoothed negative entropy function defined as follows: let $\epsilon > 0$, and let

$$\phi(x) = \epsilon \sum_{i=1}^n (x_i + \epsilon)\ln(x_i + \epsilon) + \delta_\Delta(x).$$

Although no simple expression is known for the mirror operator $\nabla\phi^*(z) = \arg\max_x \langle z, x\rangle - \phi(x)$, it can be solved efficiently, in $\mathcal{O}(n \log n)$ time using a deterministic algorithm, or $\mathcal{O}(n)$ expected time using a randomized algorithm, see [21]. Additionally, D_ϕ satisfies our assumptions: ϕ is $\frac{\epsilon}{1+n\epsilon}$-strongly convex w.r.t. $\|\cdot\|_1$, and 1-smooth w.r.t. $\|\cdot\|_\infty$. The resulting accelerated mirror descent method on the simplex can then be implemented efficiently, and by Theorem 5 it is guaranteed to converge in $\mathcal{O}(1/k^2)$ whenever $\gamma \geq 1$ and $s \leq \frac{\epsilon}{2(1+n\epsilon)L_f\gamma}$.

5.6 Restarting the discrete algorithm

In this section, we adapt the restarting heuristics proposed by O'Donoghue and Candès in [28], and Su et al. in [32], and propose new restarting heuristics. In Section 3.4, we motivated restarting for strongly convex functions, by observing that restarting at fixed intervals (determined by the strong convexity parameter of the objective), allows us to

recover linear convergence. Even when the function is not strongly convex, restarting can be intuitively motivated by the observation that because of the "memory" in the solution (both in the dual variable $Z(t) = Z(0) + \int_0^t -\tau \nabla f(X(\tau))$, which accumulates gradients, and the primal variable due to averaging), the trajectory may point in a "bad direction" at a given time t. Thus, one can restart the ODE whenever a given condition is met, by resetting time to zero and reinitializing it at the current point, effectively wiping the memory of the solution.

Recall that in continuous-time, the algorithm is restarted at a given time T_k, by solving a new ODE given by (3.7), in which time is shifted by $-T_k$, and the dual variable is reinitialized to have $\nabla\psi^*(Z(T_k)) = X(T_k^-)$ (to ensure continuity of the primal trajectory).

We define restarting in discrete time similarly to the continuous time: The algorithm is restarted at time K simply by shifting future time by $-K$, and setting the dual variable $z^{(k+1)}$ such that $\nabla\psi^*(z^{(k+1)})$ coincides with the current iterate $x^{(k+1)}$. This is summarized in Algorithm 2, where we give a general template for restarted algorithms; specific restarting conditions are discussed below.

Algorithm 2 Accelerated mirror descent with restart

Initialize $K = 0$, $\tilde{x}^{(0)} = \tilde{z}^{(0)} = x_0$.
for $k \in \mathbb{N}$ **do**
  $\tilde{z}^{(k+1)} = \arg\min_{\tilde{z}\in\mathcal{X}} \frac{(k-K)s}{r} \left\langle \nabla f(x^{(k)}), \tilde{z} \right\rangle + D_\psi(\tilde{z}, \tilde{z}^{(k)})$
  (equivalently, $z^{(k+1)} = z^{(k)} - \frac{(k-K)s}{r}\nabla f(x^{(k)})$ and $\tilde{z}^{(k+1)} = \nabla\psi^*(z^{(k+1)})$)
  $\tilde{x}^{(k+1)} = \arg\min_{\tilde{x}\in\mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(k+1)})$
  $x^{(k+1)} = \lambda_{k-K+1}\tilde{z}^{(k+1)} + (1 - \lambda_{k-K+1})\tilde{x}^{(k+1)}$, with $\lambda_k = \frac{r}{r+k}$
  if Restart condition **then**
    $K \leftarrow k$
    $\tilde{z}^{(k+1)} \leftarrow x^{(k+1)}$
  end if
end for

Many restarting conditions have been proposed in recent literature, motivated by unconstrained continuous-time optimization. We review and briefly discuss some of these conditions, then propose a new condition motivated by the primal-dual form of ODE (1.1).

(i) Gradient restart condition [28]: $\left\langle x^{(k+1)} - x^{(k)}, \nabla f(x^{(k)}) \right\rangle > 0$. Intuitively, the algorithm is restarted whenever the trajectory makes an acute angle with the gradient.

(ii) Function restart condition [28]: $f(x^{(k+1)}) \geq f(x^{(k)})$. This condition is similar to the gradient condition, since one has $f(x^{(k+1)}) \geq f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle$ by convexity of f, thus the second condition is implied by the first.

(iii) Speed restart condition [32]: $\|x^{(k+1)} - x^{(k)}\| < \|x^{(k)} - x^{(k-1)}\|$. This condition was proposed by Su et al. in [32], and is motivated by the unconstrained Euclidean case:

intuitively, the speed starts to decrease whenever (and the system starts losing momentum) whenever the trajectory points in a bad direction.

(iv) Dual restarting: $\left\langle z^{(k+1)}, \nabla f(x^{(k)}) \right\rangle > 0$. Intuitively, the algorithm is restarted whenever the dual variable (which cumulates gradients), points in a bad direction with respect to the current gradient.

We test these conditions numerically in Chapter 6, and discuss their qualitative differences.

Chapter 6

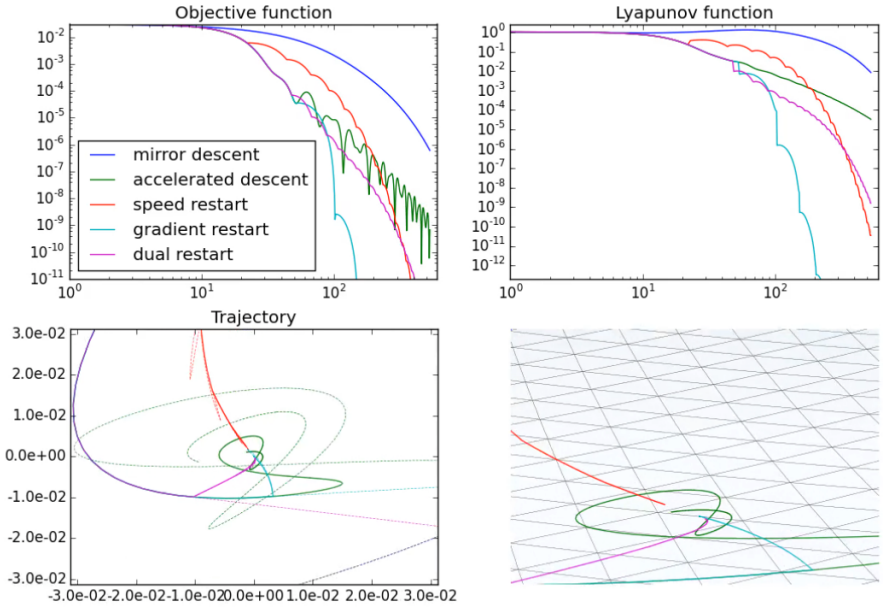
Numerical experiments

To illustrate our results, we implement the accelerated mirror descent method proposed in Algorithm 1, on simplex-constrained problems in $\mathbb{R}^n$, $n = 3$, to be able to visualize the simplex. We run the algorithm on a simple quadratic $f(x) = \langle x - x^\star, Q(x - x^\star)\rangle$ for a random positive definite matrix Q, and a weakly convex function given by $f(x) = g(x_1 - x_1^\star)^2 + (x_2 - x_3)^2$, where $g(x) = \min(x + \alpha, \max(0, x - \alpha))$. The solution set of the second problem is the segment given by $\{x \in \Delta : x_1 \in [x_1^\star - \alpha, x_1^\star + \alpha] \text{ and } x_2 = x_3\}$. We implement the accelerated entropic descent algorithm proposed in Section 5.5, and include the (non-accelerated) entropic descent for reference. We also implement the restarting heuristics discussed in Section 5.6. The corresponding code and videos are available at the following url: `http://www.github.com/walidk/AcceleratedMirrorDescent`.

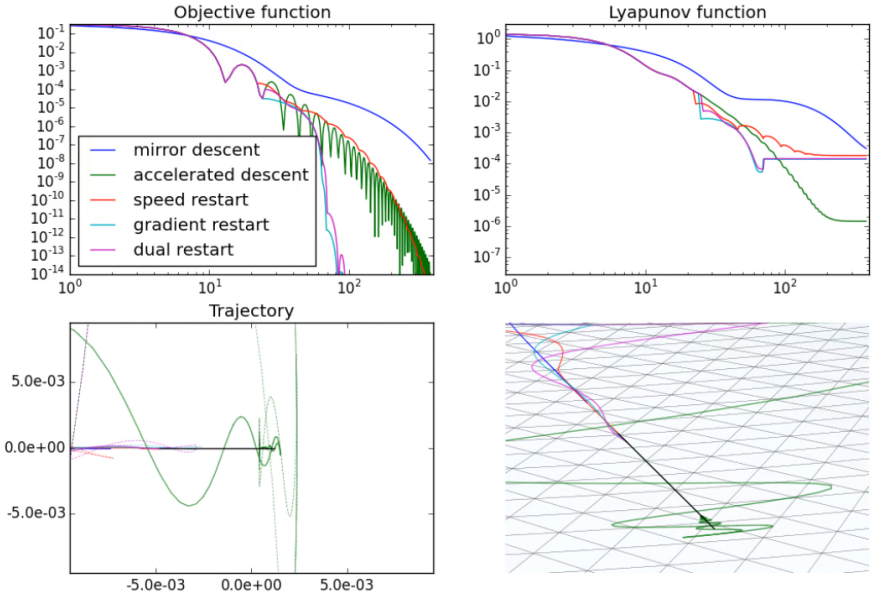
The results are given in Figures 6.1, 6.2, and 6.3. Each figure contains four plots: the first shows the value of the objective $f(x^{(k)}) - f^\star$ as a function of k, the second shows that value of the Lyapunov function $V(x^{(k)}, z^{(k)}, k)$ as a function of k, and the bottom plots show the trajectory of the solution $x^{(k)}$ on the simplex viewed as a subset of $\mathbb{R}^2$ (bottom left) and the trajectory $(x^{(k)}, f(x^{(k)}))$ on top of the surface of function values (bottom right). In the trajectory plots, the dotted lines show the mirrored dual trajectory $\nabla\psi^*(z^{(k)})$. Finally, in the weakly convex case, the set of minimizers is visualized as a solid black segment.

The accelerated mirror descent method exhibits a polynomial convergence rate, which is empirically faster than the $\mathcal{O}(1/k^2)$ rate predicted by Theorem 5, both in the strongly and weakly convex cases. The experiments confirm that the Lyapunov function is decreasing for the accelerated method, but not for the plain mirror descent method. It is worth noting that restarting sometimes increases the value of the energy function, thus a different argument is needed to analyze the convergence of these heuristics.

The method also exhibits oscillations around the set of minimizers. We observe that increasing the parameter r seems to reduce the period of the oscillations, and results in a trajectory that is initially slower, but faster for large k, see Figure 6.2. The restarting heuristics alleviate the oscillation and empirically speed up the convergence. This observation also holds when the solution is on the boundary of the feasible set, see Figure 6.3-a for an example.



(a) Strongly convex quadratic.



(b) Weakly convex function.

Figure 6.1: Accelerated mirror descent on the simplex, and restarting heuristics.

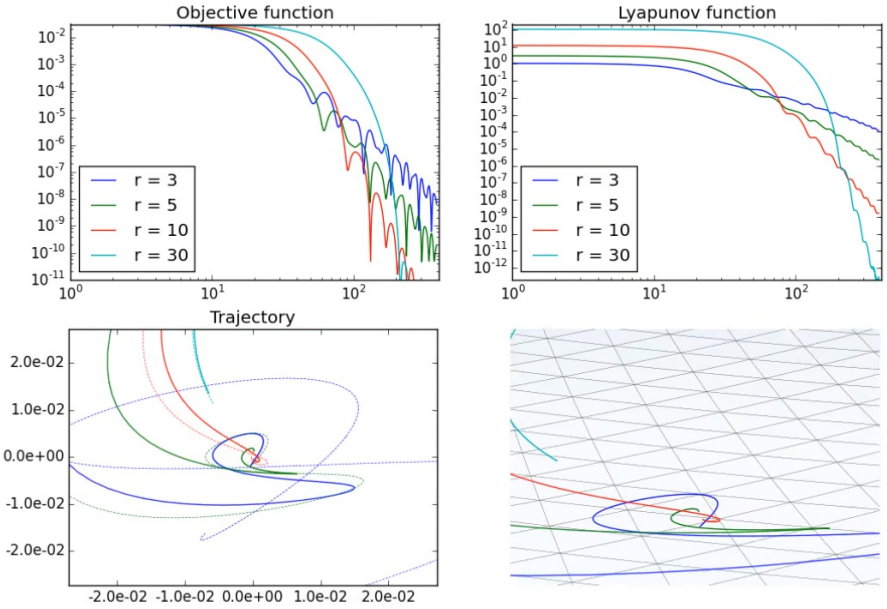


Figure 6.2: Effect of the parameter r.

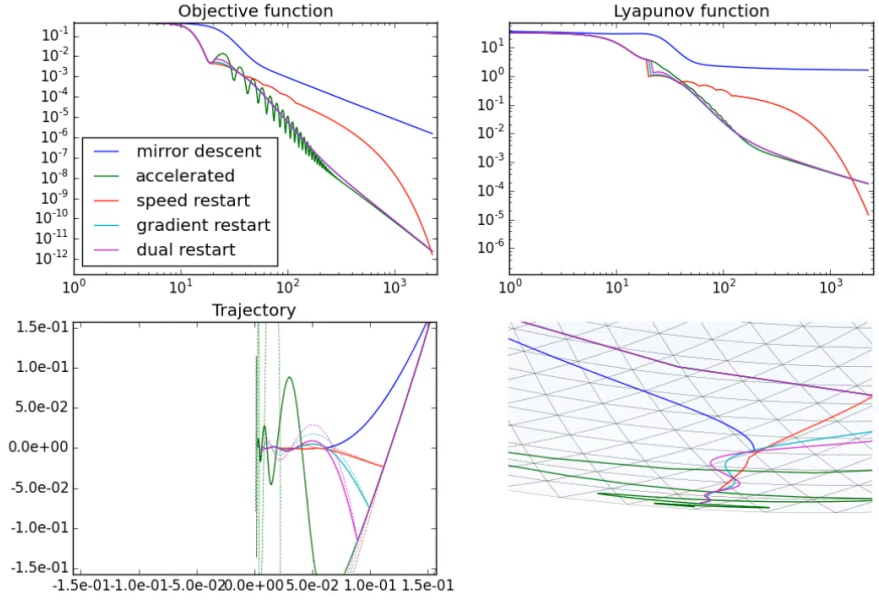


Figure 6.3: Effect of restarting when the solution is on the boundary.

Finally, we observe that in the weakly convex case, different instances of the algorithm converge to different limit points in the solution set, see Figure 6.2-b and 6.3-b. Our theoretical results on ly prove convergence in function values, which proves that the distance to the solution set $d(x^{(k)}, S)$ converges to 0, but does not prove that the trajectory $x^{(k)}$ converges. The experiments suggest that it does, and that the limit depends on the parameters of the algorithm (initial condition and value of the parameter r).

Chapter 7

Conclusion

By combining the Lyapunov argument that motivated mirror descent, and a recent ODE interpretation [32] of Nesterov's method, we propose a method to construct an energy function tailored to a given constrained convex optimization problem. The energy function combines a term that encodes the desired convergence rate, and a term that encodes the constraints (the Bregman divergence term). We then derive an ODE which is tailored to that energy function, and show existence, uniqueness and viability of its solutions. It turns out that this ODE also has a simple interpretation as a coupling between a dual variable $Z(t)$ which cumulates gradients (similar to the original mirror descent method, but with an *increasing* rate), and a primal variable $X(t)$ obtained by averaging the mirrored dual $\nabla\psi^*(Z(t))$.

By construction of the energy function, the solution trajectories converge to the set of minimizers at a $\mathcal{O}(1/t^2)$ rate. By discretizing the continuous-time dynamics, we obtain a family of accelerated mirror descent methods and proved, using the same energy function, an analogous $\mathcal{O}(1/k^2)$ rate when the step size is small enough. This connection with the continuous-time dynamics can provide intuition and insights into the behavior of the discrete-time algorithm, and motivates a more detailed study of properties of the ODE system (1.1) and its solutions, such as a characterization of the oscillatory behavior of the solution and the effect of the parameter r, the convergence rates under additional assumptions such as strong convexity, the convergence of trajectories when the minimizer is not unique, and a rigorous study of the restarting heuristics.

Appendix A

Bregman projections

A.1 Dual distance generating functions

We consider a closed, convex set $\mathcal{X}$, and a pair of conjugate convex functions ψ, ψ^* such that ψ is closed and proper, and the effective domain of ψ is $\mathcal{X}$. We denote $\mathcal{X}^*$ the effective domain of ψ^*. By Fenchel's duality theorem, ψ^{**} coincides with ψ, and we have for all $x \in E$ and $z \in E^*$:

$$\psi^*(z) = \sup_{x \in E} \langle z, x \rangle - \psi(x), \qquad \psi(x) = \sup_{z \in E^*} \langle z, x \rangle - \psi^*(z).$$

Since ψ and ψ^* are proper convex functions, they are both subdifferentiable on the relative interior of their respective domains (Theorem 23.4 in [30]). And if we denote $\partial\psi(x)$ the subdifferential of ψ at x, then we have, by definition of a subgrdient,

$$\begin{aligned} z \in \partial\psi(x) &\Leftrightarrow \psi(x') - \langle z, x' \rangle \geq \psi(x) - \langle z, x \rangle \ \forall x' \in E \\ &\Leftrightarrow x \in \arg\max_{x' \in E} \langle z, x' \rangle - \psi(x') \\ &\Leftrightarrow \psi^*(z) = \langle z, x \rangle - \psi(x) \end{aligned}$$

and switching the roles of ψ and ψ^* (and using the fact that $\psi^{**} = \psi$), we have the equivalence

$$\psi^*(z) + \psi(x) = \langle z, x \rangle \Leftrightarrow z \in \partial\psi(x) \Leftrightarrow x \in \partial\psi^*(z). \tag{A.1}$$

In other words, $\partial\psi$ and $\partial\psi^*$ are inverses of each other (in the sense of set valued mappings).

A.2 The mirror operator $\nabla\psi^*$

Recall that in mirror descent, we defined the dynamics of the variables X, Z as follows

$$\begin{cases} \dot{Z} = -\nabla f(X) \\ X = \nabla\psi^*(Z) \end{cases}$$

where $X \in \mathcal{X}$, and $Z \in E^*$ is, a priori, unconstrained. We now discuss how to obtain such an operator $\nabla\psi^*$, which maps E^* into $\mathcal{X}$. By the previous observation, we have for all $z \in E^*$,

$$\partial\psi^*(z) = \arg\max_{x \in E} \langle z, x\rangle - \psi(x).$$

And since $\operatorname{dom}\psi = \mathcal{X}$, we have that $\partial\psi^*(z) \subset \mathcal{X}$. Thus we have a set-valued function $\partial\psi^*(\cdot)$ which maps E^* into $\mathcal{X}$. For the mirror dynamics to be well-defined, we need $\partial\psi^*(z)$ to be single-valued for all $z \in E^*$, in other words, we need ψ^* to be differentiable on all of E^*. The following proposition gives a necessary and sufficient condition. First, we review some definitions.

Definition 1. *A convex function ψ is cofinite if its epigraph does not contain any non-vertical half-line.*

Definition 2. *A convex function ψ is essentially strictly convex if it is strictly convex on all convex subsets where it is subdifferentiable.*

Definition 3. *A convex function ψ is essentially smooth if it is differentiable on the interior of its domain, and $\|\nabla\psi(x)\| \to \infty$ as x tends to the boundary of the domain.*

Proposition 1. *Let ψ, ψ^* be a pair of convex, closed function which are conjugates of each other. Then ψ^* is finite and differentiable on all of E^* if and only if ψ is essentially strictly convex and cofinite.*

Proof. By Theorem 13.3 in [30], $\operatorname{dom}\psi^* = E^*$ if and only if ψ is cofinite. And by Theorem 25.3 in [30], ψ^* is essentially smooth if and only if ψ is essentially strictly convex. But when $\operatorname{dom}\psi^* = E^*$, essential smoothness and differentiability are equivalent. Therefore,

$$\begin{aligned}\psi^* \text{ is finite and differentiable on } E^* &\Leftrightarrow \operatorname{dom}\psi^* = E^* \text{ and } \psi^* \text{ is essentially smooth} \\ &\Leftrightarrow \psi \text{ is cofinite and } \psi \text{ is essentially strictly convex.}\end{aligned}$$

□

Note that, in general, ψ may not be differentiable. In fact, differentiability of ψ is very restrictive: By definition, ψ is differentiable at x if and only if there exists z such that $\lim_{\|x'-x\|\to 0} \frac{\psi(x')-\psi(x)-\langle z, x'-x\rangle}{\|x'-x\|} = 0$; in particular, ψ can only be differentiable on the interior of $\mathcal{X}$ since ψ needs to be finite in a neighborhood of x for the limit to be 0. Therefore, if $\mathcal{X}$ has empty interior, ψ is nowhere differentiable.

Finally, note that we required ψ^* to be differentiable on all of E^* since in the general case, the dynamics of the dual variable $\dot{Z} = -\nabla f(X)$ can evolve anywhere in E^*. However, for some problems, one may have a particular structure of ∇f which guarantees that Z remains in a subset of E^*. For example, suppose that there exists a convex cone $\mathcal{K}$ such that $\nabla f(x) \in \mathcal{K}$ for all $x \in \mathcal{X}$. Then Z remains in $-\mathcal{K}$, and it suffices that ψ^* is differentiable on $-\mathcal{K}$, not necessarily all of E^*. We give an example in Section A.4.

A.3 Bregman divergences and projections

Next, we define the Bregman divergences generated by the functions ψ and ψ^*. Suppose that ψ^* is differentiable on E^*. Then for $(z, z') \in E^* \times E^*$, let

$$D_{\psi^*}(z, z') = \psi^*(z) - \psi^*(z') - \langle \nabla\psi^*(z'), z - z' \rangle .$$

Assuming ψ is differentiable on $\mathcal{X}$ (this will be relaxed in Remark 2), we also define the dual Bregman divergence

$$D_{\psi}(x, x') = \psi(x) - \psi(x') - \langle \nabla\psi(x'), x - x' \rangle .$$

Then we have the following identity that relates the dual Bregman divergences.

Proposition 2. *Let $z, z' \in E^*$, and let $x = \nabla\psi^*(z)$ and $x' = \nabla\psi^*(z')$. Then*

$$D_{\psi^*}(z, z') = D_{\psi}(x', x).$$

Proof. Using equivalence (A.1), we have

$$\begin{aligned} D_{\psi^*}(z, z') &= \psi^*(z) - \psi^*(z') - \langle \nabla\psi^*(z'), z - z' \rangle \\ &= [\langle z, x \rangle - \psi(x)] - [\langle z', x' \rangle - \psi(x')] - \langle x', z - z' \rangle \\ &= \psi(x') - \psi(x) - \langle z, x' - x \rangle \\ &= D_{\psi}(x', x), \end{aligned}$$

which proves the claim. □

In the following examples, we will motivate a relaxation of the differentiability assumption on ψ, and show that one can in fact define a Bregman divergence on $\mathcal{X}$ when ψ is not differentiable but is a restriction of a differentiable function.

A.4 Examples

Entropy projection onto the positive orthant

Let $\mathcal{X}$ be the positive orthant $\mathcal{X} = \mathbb{R}^d_+$, and consider the negative (generalized) entropy $\psi(x) = -H(x) = \sum_i x_i \ln x_i$. Then ψ is differentiable on the interior of $\mathcal{X}$, $\nabla\psi(x) = (1 + \ln x_i)_i$, and a simple calculation shows that $D_{\psi}(x, x') = \sum_i x_i \ln \frac{x_i}{x'_i} - \sum_i (x_i - x'_i)$, the generalized I-divergence of x to x'.

Writing the definition of ψ^*, we have

$$\psi^*(z) = \sup_{x \in \mathbb{R}^d_+} \langle z, x \rangle - \sum_i x_i \ln x_i.$$

The maximization can be solved explicitly by writing the Lagrangian of the problem: for $\lambda \in \mathbb{R}^d_+$, let $L(x, \lambda) = \langle z, x\rangle - \sum_i x_i \ln x_i + \sum_i \lambda_i x_i$. Its gradient with respect to x is $z - (1 + \ln x_i)_i + \lambda$. Then by the KKT optimality conditions, x is optimal if and only if there exist $\lambda \in \mathbb{R}^d_+$ such that

$$\begin{cases} z - (1 + \ln x_i)_i + \lambda = 0 \\ x \geq 0, \\ x_i \lambda_i = 0 \ \forall i. \end{cases}$$

The first condition is equivalent to $x_i = e^{z_i + \lambda_i - 1}$, and since any solution of this form is strictly positive, the complementary slackness condition requires that $\lambda = 0$, thus the solution is simply

$$\nabla \psi^*(z) = x = \left(e^{z_i - 1}\right)_i$$

and $\psi^*(z) = \langle z, x\rangle - \psi(x) = \sum_i e^{z_i - 1}$, defined and differentiable on all of E^*.

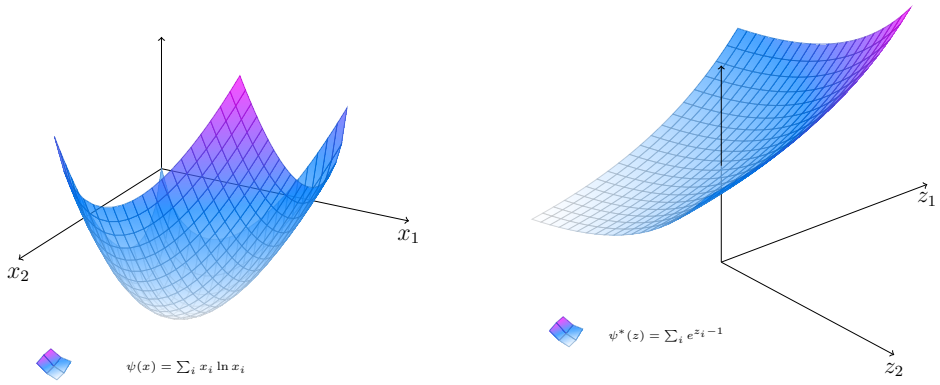


Figure A.1: Illustration of the generalized negative entropy function $\psi(x) = -H(x)$, and its conjugate $\psi^*(z) = \sum_i e^{z_i - 1}$.

Entropy projection onto the simplex

Let $\mathcal{X}$ be the probability simplex on $\mathbb{R}^n$, i.e. $\mathcal{X} = \Delta = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$, and let ψ be the negative entropy $-H$ restricted to Δ. Formally, $\psi(x) = -H(x) + \delta_\Delta(x)$, where $\delta_\mathcal{X}(\cdot)$ is the indicator function of the convex set $\mathcal{X}$, defined as follows: $\delta_\mathcal{X}(x) = 0$ if $x \in \mathcal{X}$, and $+\infty$ otherwise.

Remark 2. *Strictly speaking, ψ is nowhere differentiable, since Δ has empty interior in $\mathbb{R}^n$. In fact, the subdifferential of ψ at $x \in \operatorname{ri} \Delta$ is*

$$\partial \psi(x) = -\nabla H(x) + \mathbb{R}u,$$

where $-\nabla H(x)$ is the gradient of H at x computed in the previous section, and u is a normal vector to the affine hull of Δ at x, i.e. u satisfies $\langle u, y - x\rangle = 0$ for all $y \in \operatorname{aff} \Delta$. One such

normal vector is the vector with all entries equal to 1. *Next, we argue that in this case, if we replace* $\nabla\psi(x)$ *by any vector in* $\partial\psi(x)$ *in the definition of* $D_\psi(y,x)$, *the choice of subgradient vector does not affect the value of* $D_\psi(y,x)$. *Formally, we write*

$$D_\psi(y,x) = \psi^*(y) - \psi^*(x) - \langle \partial\psi(x), y - x\rangle\,.$$

Indeed, we have for all $z \in \partial\psi(x)$, $z = \nabla H(x) + \alpha u$ *for some* α, *and since* u *is normal to* Δ, *we have for all* $y \in \Delta$,

$$\psi(y)-\psi(x)-\langle z, y-x\rangle = \psi(y)-\psi(x)-\langle \nabla H(x)+\alpha u, y-x\rangle = \psi^*(y)-\psi^*(x)-\langle \nabla H(x), y-x\rangle\,,$$

and the definition is unambiguous. More generally, we have the following proposition.

Proposition 3. *Suppose that* ψ *is the restriction of a differentiable function* Ψ *to a convex set* $\mathcal{X}$ *of affine dimension* $m < n$ *(the affine dimension of a convex set is the affine dimension of its affine hull),*

$$\psi(x) = \Psi(x) + \delta_{\mathcal{X}}(x).$$

Then for all $x \in \operatorname{ri}\mathcal{X}$, *the subdifferential of* ψ *is given by*

$$\partial\psi(x) = \left\{\nabla\Psi(x) + U\alpha,\ \alpha \in \mathbb{R}^{n-m}\right\}$$

where $(u_i)_{i\in\{m+1,\dots,n\}}$ *forms a basis for* N, *the subspace of normal vectors to the affine hull of* $\mathcal{X}$, *and* $U = \begin{pmatrix} u_{m+1} \mid \dots \mid u_n \end{pmatrix}$. *As a consequence, for all* $(x,x') \in \mathcal{X} \times \operatorname{ri}\mathcal{X}$,

$$D_\psi(x',x) = \psi(x') - \psi(x) - \langle \partial\psi(x), x' - x\rangle$$

is defined unambiguously, and does not depend on the choice of subgradient $z \in \partial\psi(x)$. *Furthermore,* $\partial\psi^*(z + U\alpha) = \partial\psi^*(z)$ *for all* z, *and if* ψ^* *is twice differentiable on* E^*, *then for all* $x \in \operatorname{ri}\mathcal{X}$,

$$\nabla^2\psi^*(\partial\psi(x))$$

is defined unambiguously, and does not depend on the choice of subgradient $z \in \partial\psi(x)$.

Finally, for all $z, z' \in E^*$, *if* $x \in \partial\psi^*(z)$ *and* $x' \in \partial\psi^*(z')$, *then*

$$D_{\psi^*}(z,z') = D_\psi(x',x).$$

Proof. First, we have

$$\partial\psi(x) = \partial\Psi(x) + \partial\delta_{\mathcal{X}}(x) = \nabla\Psi(x) + \partial\delta_{\mathcal{X}}(x),$$

since Ψ is differentiable. The subdifferential of $\delta_{\mathcal{X}}$ at x is the normal cone to $\mathcal{X}$ at x, and since $x \in \operatorname{ri}\mathcal{X}$, the normal cone coincides with the subspace N of vectors that are normal to the affine hull of $\mathcal{X}$, which proves the first part of the proposition.

To prove that the Bregman divergence is defined unambiguously, let $z \in \partial\psi(x)$. Then $\exists\alpha \in \mathbb{R}^{n-m}$ such that $z = \nabla\Psi(x) + \sum_{i=m+1}^{n} \alpha_i u_i$, and

$$\begin{aligned}\psi(x') - \psi(x) - \langle z, x' - x\rangle &= \psi(x') - \psi(x) - \left\langle \nabla\psi(x) + \sum_i \alpha_i u, x' - x \right\rangle \\ &= \psi(x') - \psi(x) - \langle \nabla\psi(x), x' - x\rangle\end{aligned}$$

which does not depend on the choice of $z \in \partial\psi(x)$. Next, to show that $\partial\psi^*(z+U\alpha) = \partial\psi^*(z)$, we have

$$\begin{aligned}x \in \partial\psi^*(z) &\Leftrightarrow z \in \partial\psi(x) \\ &\Leftrightarrow z \in \partial\psi(x) - U\alpha \ \forall\alpha \in \mathbb{R}^{n-m} \\ &\Leftrightarrow z + U\alpha \in \partial\psi(x) \ \forall\alpha \in \mathbb{R}^{n-m} \\ &\Leftrightarrow x \in \partial\psi^*(z + U\alpha) \ \forall\alpha \in \mathbb{R}^{n-m}.\end{aligned}$$

Furthermore, if ψ^* is twice differentiable, since $\nabla\psi^*(z+U\alpha) = \nabla\psi^*(z)$ for all $\alpha \in \mathbb{R}^{n-m}$, we also have $\nabla^2\psi^*(z + U\alpha) = \nabla^2\psi^*(z)$, and it follows that

$$\nabla^2\psi^*(\partial\psi(x)) = \nabla^2\psi^*(\nabla\Psi(x))$$

which does not depend on the choice of subgradient. Finally, the Bregman duality identity follows from the proof of Proposition 2. □

Given the previous convention, and writing $\psi(x) = -H(x) + \delta_\Delta(x)$, we have $-\nabla H(x) = (1 + \ln x_i)_i$, and a simple calculation shows that $D_\psi(x, x') = \sum_i x_i \ln \frac{x_i}{x'_i}$ is the Kullback Leibler divergence between the distribution vectors x, x'. Similarly to the previous section, we can write the definition of ψ^*,

$$\psi^*(z) = \max_{x\in\Delta} \langle x, z\rangle - \sum_i x_i \ln x_i,$$

and solve the maximization problem by writing the Lagrangian: for $\mu \in \mathbb{R}$ and $\lambda \in \mathbb{R}^d_+$, let $L(x, \nu, \mu) = \langle x, z\rangle - \sum_i x_i \ln x_i + \nu(\sum_i x_i - 1) + \sum_i \lambda_i x_i$. Its gradient with respect to x is $z - (1 + \ln x_i)_i - \boldsymbol{\nu} + \lambda$. Then by the KKT optimality conditions, x is optimal if and only if there exist $\lambda \in \mathbb{R}^d_+$ and ν such that

$$\begin{cases} z - (1 + \ln x_i)_i - \boldsymbol{\nu} + \lambda = 0 \\ x \geq 0, \ \sum_i x_i = 1 \\ x_i\lambda_i = 0 \ \forall i. \end{cases}$$

The first condition can be rewritten $x_i = e^{z_i+\lambda_i}/e^{\nu+1}$. Thus the third condition (complementary slackness), requires λ_i to be 0, and the expression of x simplifies to $x_i = e^{z_i}/e^{\nu+1}$.

Finally, the primal feasibility condition $\sum_i x_i = 1$ requires that $\sum_i e^{z_i}/e^{\nu+1} = 1$. Therefore, the unique solution of the maximization problem is

$$\nabla\psi^*(z)_i = x_i = \frac{e^{z_i}}{\sum_j e^{z_j}},$$

and simple algebra shows that $\psi^*(z) = \langle z, x\rangle - \psi(x) = \ln\sum_i e^{z_i}$, defined on $E^* = \mathbb{R}^n$. Note that we can verify the observation of Remark 2: if u is a normal vector to aff Δ, then $\nabla\psi^*(z) = \nabla\psi^*(z + \alpha u)$ for all u, so by duality of the subdifferentials, $z \in \partial\psi(x)$ if and only if $z + \alpha u \in \partial\psi(x)$ for all α. It also follows that ψ^* is linear in the direction u, i.e. $\psi^*(x + \alpha u) = \alpha + \psi^*(x)$.

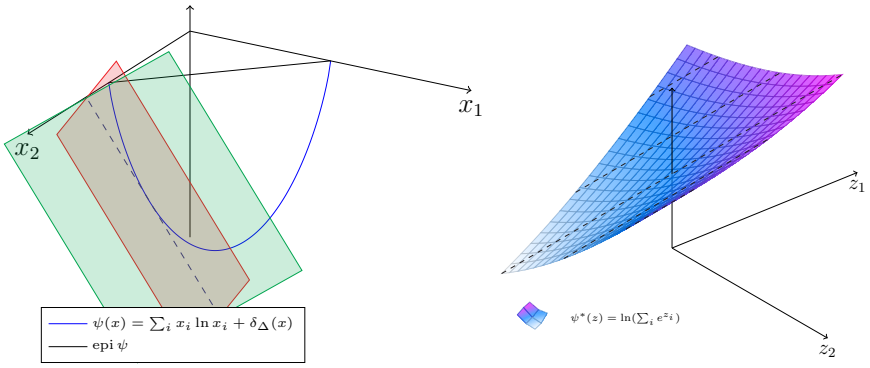


Figure A.2: Illustration of the negative entropy function restricted to the simplex $\psi(x) = -H(x) + \delta_\Delta(x)$, and its conjugate $\psi^*(z) = \ln(\sum_i e^{z_i})$. The function ψ is subdifferentiable on the interior of Δ, but nowhere differentiable. The figure illustrates this fact by showing two supporting hyperplanes at the same point. The conjugate function ψ^* is linear in the direction normal to the simplex (shown in dashed lines on the right).

Itakura-Saito divergence on the positive orthant

Let $\mathcal{X}$ be the positive orthant $\mathcal{X} = \mathbb{R}^d_+$, and let $\psi(x) = -\sum_i \ln x_i$. Then $\nabla\psi(x) = \left(-\frac{1}{x_i}\right)_i$, and a simple calculation shows that $D_\psi(x, x') = \sum_i \left(\frac{x_i}{x'_i} - \ln\frac{x_i}{x'_i} - 1\right)$, the Itakura-Saito divergence of x and x'.

Writing the expression of ψ^*, we have

$$\psi^*(z) = \sup_{x\in\mathbb{R}^d_+} \langle z, x\rangle + \sum_i \ln x_i,$$

it is finite on $\mathcal{X}^* = \mathbb{R}^d_-$. The maximization can be solved using the same approach as the previous examples. Define the Lagrangian, for $\lambda \in \mathbb{R}^d_+$, $L(x, \lambda) = \langle z, x\rangle + \sum_i \ln x_i + \sum_i \lambda_i x_i$.

Its gradient with respect to x is $z + \left(\frac{1}{x_i}\right)_i + \lambda$, and x is optimal if and only if there exists $\lambda \in \mathbb{R}^d_+$ such that

$$\begin{cases} z + \left(\frac{1}{x_i}\right)_i + \lambda = 0 \\ x \geq 0 \\ \lambda_i x_i = 0, \end{cases}$$

and the first condition can be rewritten as $x_i = \frac{1}{z_i + \lambda_i}$. Since any solution of this form is non-zero, the complementary slackness condition requires that $\lambda = 0$, and the first condition becomes $x_i = -\frac{1}{z_i}$. Therefore

$$\nabla \psi^*(z) = \left(-\frac{1}{z_i}\right)_i$$

and simple calculation shows that $\psi^*(z) = \langle z, x \rangle + \sum_i \ln x_i = -\sum_i [1 + \ln(-z_i)]$, defined on $\mathcal{X}^* = \mathbb{R}^d_-$.

$\mathcal{X}$	$\mathcal{X}^*$	$\psi^*(z)$	$\psi(x)$	$\nabla\psi^*(z)$	$\nabla\psi(x)$	$D_{\psi^*}(x, x')$
Δ	$\mathbb{R}^n$	$\ln \sum_i e^{z_i}$	$\sum_i x_i \ln x_i$	$\left(\frac{e^{z_i}}{\sum_j e^{z_j}}\right)_i$	$(1 + \ln x_i)_i$	$\sum_i x_i \ln \frac{x_i}{x'_i}$
$\mathbb{R}^n_+$	$\mathbb{R}^n$	$\sum_i e^{z_i - 1}$	$\sum_i x_i \ln x_i$	$\left(e^{z_i - 1}\right)_i$	$(1 + \ln x_i)_i$	$\sum_i x_i \ln \frac{x_i}{x'_i} - \sum_i (x_i - x'_i)$
$\mathbb{R}^n_+$	$\mathbb{R}^d_-$	$-\sum_i [1 + \ln(-z_i)]$	$-\sum_i \ln x_i$	$\left(-\frac{1}{z_i}\right)_i$	$\left(-\frac{1}{x_i}\right)_i$	$\sum_i \left(\frac{x_i}{x'_i} - \ln \frac{x_i}{x'_i} - 1\right)$

Table A.1: Examples of dual distance generating functions and the corresponding Bregman projections.

Appendix B

Proof of Lemma 1

Let us rewrite the smoothed accelerated mirror descent ODE system

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X) \\ \dot{X} = \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X) \\ X(0) = x_0,\ Z(0) = z_0 \text{ with } \nabla\psi^*(z_0) = x_0. \end{cases}$$

By the Cauchy-Lipschitz theorem, there exists a unique solution (X_δ, Z_δ) defined on $[0, t_{\max})$, and the solution is C^1. Define, for $t > 0$,

$$\begin{aligned} A_\delta(t) &= \sup_{u\in[0,t]} \frac{\|\dot{Z}_\delta(u)\|}{u} \\ B_\delta(t) &= \sup_{u\in[0,t]} \frac{\|X_\delta(u) - x_0\|}{u} \\ C_\delta(t) &= \sup_{u\in[0,t]} \|\dot{X}_\delta(u)\| \end{aligned}$$

These quantities are finite for the following reasons:

- $\frac{\|X_\delta(u)-x_0\|}{u} = \|\dot{X}_\delta(0)\| + o(1)$ near 0, thus B_δ is finite.
- $\|\dot{X}_\delta\|$ is continuous thus bounded on $[0, t]$, thus C_δ is finite.
- Finiteness of A_δ is a consequence of the following lemma.

To prove Lemma 1, we first need the auxiliary lemma below, that provides bounds on $A_\delta, B_\delta, C_\delta$.

Lemma 3. *For all t,*

$$rA_\delta(t) \le \|\nabla f(x_0)\| + L_f t B_\delta(t), \tag{B.1}$$

$$B_\delta(t) \le \frac{L_{\psi^*} rt}{6} A_\delta(t), \tag{B.2}$$

$$C_\delta(t) \le r\left(\frac{tL_{\psi^*}}{2} A_\delta(t) + B_\delta(t)\right). \tag{B.3}$$

Proof. By definition of A_δ and B_δ, we have

$$\|Z_\delta(t) - z_0\| \le \int_0^t \|\dot{Z}_\delta(v)\| dv \le A_\delta(t) \int_0^t v dv = \frac{t^2}{2} A_\delta(t), \tag{B.4}$$
$$\|X_\delta(t) - x_0\| \le t B_\delta(t).$$

Now, from the first equation in (3.3), we have for all $t \le t_0$

$$\begin{aligned}
\frac{r}{t}\|\dot{Z}_\delta(t)\| &= \|\nabla f(X_\delta(t))\| \\
&\le \|\nabla f(x_0)\| + \|\nabla f(X_\delta(t)) - \nabla f(x_0)\| \\
&\le \|\nabla f(x_0)\| + L_f \|X_\delta(t) - x_0\| && \nabla f \text{ is } L_f\text{-Lipschitz} \\
&\le \|\nabla f(x_0)\| + L_f t B_\delta(t).
\end{aligned}$$

Thus,

$$rA_\delta(t) \le \|\nabla f(x_0)\| + L_f t B_\delta(t).$$

To prove inequality (B.2), we show that $\|X_\delta(t) - x_0\| \le \frac{r}{\max(\delta,t)} \int_0^t \|\nabla\psi^*(Z_\delta(s)) - \nabla\psi^*(z_0)\| ds$. We consider the two cases $t \le \delta$ and $t \ge \delta$.

- Let $t \le \delta$. From the second equation in (3.3), we have

$$e^{\frac{rt}{\delta}}\left(\dot{X}_\delta + \frac{r}{\delta}(X_\delta - x_0)\right) = \frac{r}{\delta} e^{\frac{rt}{\delta}} (\nabla\psi^*(Z_\delta) - \nabla\psi^*(z_0)),$$

  i.e.,

$$\frac{d}{dt}\left((X_\delta(t) - x_0) e^{\frac{rt}{\delta}}\right) = \frac{r}{\delta} e^{\frac{rt}{\delta}} (\nabla\psi^*(Z_\delta(t)) - \nabla\psi^*(z_0)),$$

  thus integrating

$$(X_\delta(t) - x_0) e^{\frac{rt}{\delta}} = \frac{r}{\delta} \int_0^t e^{\frac{rs}{\delta}} (\nabla\psi^*(Z_\delta(s)) - \nabla\psi^*(z_0)) ds$$

  dividing by $e^{\frac{rt}{\delta}}$ and taking norms we obtain the desired inequality.

- Let $t \geq \delta$. From the second equation in (3.3), we have

$$t^r \left(\dot{X}_\delta + \frac{r}{t}(X_\delta - x_0) \right) = r t^{r-1} (\nabla \psi^*(Z_\delta) - \nabla \psi^*(z_0)),$$

  i.e.

$$\frac{d}{dt} \left(t^r (X_\delta(t) - x_0) \right) = r t^{r-1} (\nabla \psi^*(Z_\delta) - \nabla \psi^*(z_0)),$$

  thus integrating

$$t^r (X_\delta(t) - x_0) = \int_0^t r s^{r-1} (\nabla \psi^*(Z_\delta(s)) - \nabla \psi^*(z_0)) ds$$

  dividing by t^r and taking norms, we obtain the desired inequality.

Now we have

$$\begin{aligned}
\|X_\delta(t) - x_0\| &\leq \frac{r}{\max(\delta, t)} \int_0^t \|\nabla \psi^*(Z_\delta(s)) - \nabla \psi^*(z_0)\| ds \\
&\leq \frac{L_{\psi^*} r}{\max(\delta, t)} \int_0^t \|Z_\delta(s) - z_0\| ds && \nabla \psi^* \text{ is } L_{\psi^*}\text{-Lipschitz} \\
&\leq \frac{L_{\psi^*} r}{\max(\delta, t)} \int_0^t \frac{s^2}{2} A_\delta(t) ds && \text{by (B.4)} \\
&= \frac{L_{\psi^*} r}{\max(\delta, t)} A_\delta(t) \frac{t^3}{6} \\
&\leq \frac{L_{\psi^*} r t^2 A_\delta(t)}{6}.
\end{aligned}$$

Dividing by t and taking the supremum, we have (B.2).

Finally, to bound C_δ, we have from the second equation in (3.3), for all $t \leq t_0$,

$$\begin{aligned}
\|\dot{X}_\delta(t)\| &= \frac{r}{\max(\delta, t)} \|\nabla \psi^*(Z_\delta(t)) - X_\delta(t)\| \\
&\leq \frac{r}{\max(\delta, t)} (\|\nabla \psi^*(Z_\delta(t)) - \nabla \psi^*(z_0)\| + \|X_\delta(t) - x_0\|) \\
&\leq \frac{r}{\max(\delta, t)} (L_{\psi^*} \|Z_\delta(t) - z_0\| + \|X_\delta(t) - x_0\|) \\
&\leq \frac{r}{\max(\delta, t)} \left(\frac{t^2}{2} L_{\psi^*} A_\delta(t) + t B_\delta(t) \right) \\
&\leq r \left(\frac{L_{\psi^*} t}{2} A_\delta(t) + B_\delta(t) \right),
\end{aligned}$$

which conclude the proof. □

Proof of Lemma 1. First, we show that $A_\delta, B_\delta, C_\delta$ are bounded on $[0, t_0]$, uniformly in δ. Combining (B.1) and (B.2), we have

$$rA_\delta(t) \leq \|\nabla f(x_0)\| + L_f t B_\delta(t) \leq \|\nabla f(x_0)\| + L_f t \frac{L_{\psi^*} r t}{6} A_\delta(t).$$

Thus $A_\delta(t)\left(1 - \frac{L_{\psi^*} L_f}{6} t^2\right) \leq \frac{\|\nabla f(x_0)\|}{r}$. And when $t \leq \frac{2}{\sqrt{L_f L_{\psi^*}}}$, $1 - \frac{L_{\psi^*} L_f}{6} t^2 \geq \frac{1}{3}$, thus

$$A_\delta(t) \leq \frac{3}{r} \|\nabla f(x_0)\|. \tag{B.5}$$

Next, we have

$$\begin{aligned} C_\delta(t) &\leq r\left(\frac{t L_{\psi^*}}{2} A_\delta(t) + B_\delta(t)\right) && \text{by (B.3)} \\ &\leq r\left(\frac{t L_{\psi^*}}{2} A_\delta(t) + \frac{L_{\psi^*} r t}{6} A_\delta(t)\right) && \text{by (B.2)} \\ &\leq \frac{(3+r) L_{\psi^*} t}{2} \|\nabla f(x_0)\| && \text{by (B.5)} \end{aligned}$$

To conclude, we have for all $t \in [0, t_0]$

$$\begin{aligned} \|\dot{Z}_\delta(t)\| &\leq t A_\delta(t) \leq \frac{3t}{r} \|\nabla f(x_0)\|, \\ \|\dot{X}_\delta(t)\| &\leq C_\delta(t) \leq \frac{(3+r) L_{\psi^*} t}{2} \|\nabla f(x_0)\|, \end{aligned}$$

which are bounded uniformly in δ on $[0, t_0]$, thus the family is equi-Lipschitz-continuous on $[0, t_0]$. It also follows that it is uniformly bounded on the same interval. □

Appendix C

Proof of Lemma 2

In our analysis, we will use the following lemmas.

Lemma 4. *Let f be a convex function and suppose that ∇f is L_f-Lipschitz w.r.t. $\|\cdot\|$. Then for all x, x', x^+,*

$$f(x^+) \leq f(x') + \langle \nabla f(x), x^+ - x' \rangle + \tfrac{L_f}{2}\|x^+ - x\|^2$$

Proof. Since ∇f is L_f-Lipschitz, we have

$$f(x^+) \leq f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{L_f}{2}\|x^+ - x\|^2$$

and by convexity of f,

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle$$

Summing the two inequalities, we obtain the result. □

Lemma 5. *For all u, v, w*

$$D_{\psi^*}(u, v) - D_{\psi^*}(w, v) = -D_{\psi^*}(w, u) + \langle \nabla\psi^*(u) - \nabla\psi^*(v), u - w \rangle$$

Proof. By definition of the Bregman divergence, we have

$$\begin{aligned}
&D_{\psi^*}(u, v) - D_{\psi^*}(w, v) \\
&= \psi^*(u) - \psi^*(v) - \langle \nabla\psi^*(v), u - v \rangle - (\psi^*(w) - \psi^*(v) - \langle \nabla\psi^*(v), w - v \rangle) \\
&= \psi^*(u) - \psi^*(w) - \langle \nabla\psi^*(v), u - w \rangle \\
&= -(\psi^*(w) - \psi^*(u) - \langle \nabla\psi^*(u), w - u \rangle) + \langle \nabla\psi^*(u) - \nabla\psi^*(v), u - w \rangle \\
&= -D_{\psi^*}(w, u) + \langle \nabla\psi^*(u) - \nabla\psi^*(v), u - w \rangle
\end{aligned}$$

□

Lemma 6. *For all $u, v \in E^*$,*

$$\frac{1}{2L_{\psi^*}}\|\tilde{u} - \tilde{v}\|^2 \leq D_{\psi^*}(u, v) \leq \frac{L_{\psi^*}}{2}\|u - v\|_*^2$$

where $\tilde{u} = \nabla\psi^*(u)$ and $\tilde{v} = \nabla\psi^*(v)$.

Proof. We have

$$\begin{aligned}
D_{\psi^*}(u, v) &= \psi^*(u) - \psi^*(v) - \langle \nabla\psi^*(v), u - v\rangle \\
&= \int_0^1 \nabla \langle \psi^*(v + t(u - v)) - \nabla\psi^*(v), u - v\rangle \, dt \\
&\leq \|u - v\|_* \int_0^1 \|\psi^*(v + t(u - v)) - \nabla\psi^*(v)\| dt \quad \text{by the Cauchy-Schwartz inequality} \\
&\leq L_{\psi^*}\|u - v\|_* \int_0^1 \|v + t(u - v) - v\|_* dt \qquad \text{since } \psi^* \text{ is } L_{\psi^*}\textit{Lipschitz} \\
&= L_{\psi^*}\|u - v\|_*^2 \int_0^1 t dt
\end{aligned}$$

which proves the second inequality. The first inequality will be proved by dualizing this inequality. Fix $v \in E^*$ and define

$$\begin{aligned}
h(u) &= D_{\psi^*}(u, v) = \psi^*(u) - \psi^*(v) - \langle \nabla\psi^*(v), u - v\rangle , \\
d(u) &= \frac{L_{\psi^*}}{2}\|u - v\|_*^2.
\end{aligned}$$

Then by the previous inequality, $h(u) \leq d(u)$ for all $u \in E^*$, and taking duals, we have $h^*(u^*) \geq d^*(u^*)$ for all u^*. We now derive the duals. Let $\tilde{v} = \psi^*(v)$. Then,

$$\begin{aligned}
h^*(u^*) &= \sup_u \langle u^*, u\rangle - h(u) \\
&= \sup_u \langle u^*, u\rangle - \psi^*(u) + \psi^*(v) + \langle \tilde{v}, u - v\rangle \\
&= \psi^*(v) - \langle v, \tilde{v}\rangle + \sup_u \langle u^* + \tilde{v}, u\rangle - \psi^*(u) \\
&= \psi^*(v) - \langle v, \tilde{v}\rangle + \psi(u^* + \tilde{v}),
\end{aligned}$$

and

$$\begin{aligned} d^*(u^*) &= \sup_u \langle u^*, u\rangle - d(u) \\ &= \sup_u \langle u^*, u\rangle - \frac{L_{\psi^*}}{2}\|u-v\|_*^2 \\ &= \sup_w \langle u^*, v+w\rangle - \frac{L_{\psi^*}}{2}\|w\|_*^2 \\ &= \langle u^*, v\rangle + \sup_w \langle u^*, w\rangle - \frac{L_{\psi^*}}{2}\|w\|_*^2 \\ &= \langle u^*, v\rangle + \frac{1}{2L_{\psi^*}}\|u^*\|^2, \end{aligned}$$

where the last equality uses Cauchy-Schwartz. Therefore combining the two inequalities,

$$\psi^*(v) - \langle v, u^* + \tilde{v}\rangle + \psi(u^*+\tilde{v}) \geq \frac{1}{2L_{\psi^*}}\|u^*\|^2.$$

In particular, for all $u \in E^*$, if we call $\tilde{u} = \nabla\psi^*(u)$, and take $u^* = \tilde{u} - \tilde{v}$, then

$$\psi^*(v) - \langle v, \tilde{u}\rangle + \psi(\tilde{u}) \geq \frac{1}{2L_{\psi^*}}\|\tilde{u}-\tilde{v}\|^2,$$

and by Theorem 23.5 in Rockafellar, $\psi(\tilde{u}) = \langle u, \tilde{u}\rangle - \psi^*(\tilde{u})$, thus

$$\psi^*(v) - \psi^*(u) - \langle \tilde{u}, v-u\rangle \geq \frac{1}{2L_{\psi^*}}\|\tilde{u}-\tilde{v}\|^2.$$

which proves the claim. □

Proof of Lemma 2. We start by bounding the difference in Bregman divergences

$$\begin{aligned} &D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\ &\quad = -D_{\psi^*}(z^{(k)}, z^{(k+1)}) + \left\langle \nabla\psi^*(z^{(k+1)}) - \nabla\psi^*(z^\star), z^{(k+1)} - z^{(k)}\right\rangle \quad \text{by Lemma 5,} \\ &\quad \leq -\frac{1}{2L_{\psi^*}}\|\tilde{z}^{(k+1)} - \tilde{z}^{(k)}\|^2 + \left\langle \tilde{z}^{(k+1)} - x^\star, -\frac{ks}{r}\nabla f(x^{(k)})\right\rangle \qquad \text{by Lemma 6.} \quad \text{(C.1)} \end{aligned}$$

Now using the step from $x^{(k)}$ to $\tilde{x}^{(k+1)}$, we have

$$\tilde{x}^{(k+1)} = \arg\min_{x\in\mathcal{X}} \gamma s\left\langle \nabla f(x^{(k)}), x\right\rangle + R(x, x^{(k)})$$

with $\frac{\ell_R}{2}\|x-y\|^2 \leq R(x,y) \leq \frac{L_R}{2}\|x-y\|^2$. Therefore, for any x, $R(x, x^{(k)}) \geq R(\tilde{x}^{(k+1)}, x^{(k)}) + \gamma s\left\langle \nabla f(x^{(k)}), \tilde{x}^{(k+1)} - x\right\rangle$. We can write

$$\tilde{z}^{(k+1)} - \tilde{z}^{(k)} = \frac{1}{\lambda_k}\left(\lambda_k\tilde{z}^{(k+1)} + (1-\lambda_k)\tilde{x}^{(k)} - x^{(k)}\right) = \frac{1}{\lambda_k}\left(d^{(k+1)} - x^{(k)}\right),$$

where we have defined $d^{(k+1)}$ in the obvious way. Thus

$$\begin{aligned}
&\|\tilde{z}^{(k+1)} - \tilde{z}^{(k)}\|^2 \\
&= \frac{1}{\lambda_k^2}\|d^{(k+1)} - x^{(k)}\|^2 \\
&\geq \frac{1}{\lambda_k^2}\frac{2}{L_R}R(d^{(k+1)}, x^{(k)}) \\
&\geq \frac{1}{\lambda_k^2}\frac{2}{L_R}\left(R(\tilde{x}^{(k+1)}, x^{(k)}) + \gamma s\left\langle \nabla f(x^{(k)}), \tilde{x}^{(k+1)} - d^{(k+1)}\right\rangle\right) \\
&\geq \frac{1}{\lambda_k^2}\frac{2}{L_R}\left(\frac{\ell_R}{2}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 + \gamma s\left\langle \nabla f(x^{(k)}), \tilde{x}^{(k+1)} - \lambda_k\tilde{z}^{(k+1)} - (1-\lambda_k)\tilde{x}^{(k)}\right\rangle\right).
\end{aligned}$$

Thus

$$\lambda_k\frac{kL_R}{2r\gamma}\|\tilde{z}^{(k+1)} - \tilde{z}^{(k)}\|^2 \geq \frac{k\ell_R}{2r\lambda_k\gamma}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 + \left\langle \frac{ks}{r}\nabla f(x^{(k)}), \frac{1}{\lambda_k}\tilde{x}^{(k+1)} - \tilde{z}^{(k+1)} - \frac{1-\lambda_k}{\lambda_k}\tilde{x}^{(k)}\right\rangle. \tag{C.2}$$

Subtracting (C.2) from (C.1),

$$\begin{aligned}
&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq -\alpha_k\|\tilde{z}^{(k+1)} - \tilde{z}^{(k)}\|^2 - \frac{k\ell_R}{2r\lambda_k\gamma}\|\tilde{x}^{(k+1)} - x^{(k)}\|^2 \\
&\qquad + \left\langle -\frac{ks}{r}\nabla f(x^{(k)}), -x^\star + \frac{1}{\lambda_k}\tilde{x}^{(k+1)} - \frac{1-\lambda_k}{\lambda_k}\tilde{x}^{(k)}\right\rangle,
\end{aligned}$$

where

$$\alpha_k = \frac{1}{2L_{\psi^*}} - \frac{k\lambda_k L_R}{2r\gamma}.$$

Defining $D_1^{(k+1)} = \|\tilde{x}^{(k+1)} - x^{(k)}\|^2$ and $D_2^{(k+1)} = \|\tilde{z}^{(k+1)} - \tilde{z}^{(k)}\|^2$, we can rewrite the last inequality as

$$\begin{aligned}
&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&= -\alpha_k D_2^{(k+1)} - \frac{k\ell_R}{2r\lambda_k\gamma}D_1^{(k+1)} + \frac{sk}{r}\left\langle -\nabla f(x^{(k)}), \tilde{x}^{(k+1)} - x^\star\right\rangle \\
&\qquad + \frac{1-\lambda_k}{\lambda_k}\frac{sk}{r}\left\langle -\nabla f(x^{(k)}), \tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right\rangle
\end{aligned}$$

By Lemma 4, we can bound the inner products as follows

$$\begin{aligned}
\left\langle \tilde{x}^{(k+1)} - \tilde{x}^{(k)}, -\nabla f(x^{(k)})\right\rangle &\leq f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2}D_1^{(k+1)}, \\
\left\langle \tilde{x}^{(k+1)} - x^\star, -\nabla f(x^{(k)})\right\rangle &\leq f^* - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2}D_1^{(k+1)}.
\end{aligned}$$

Combining these inequalities, and using the fact that $\frac{1-\lambda_k}{\lambda_k} = \frac{k}{r}$, we have

$$\begin{aligned}
&D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq -\alpha_k D_2^{(k+1)} + \frac{k^2 s}{r^2}\left(f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2} D_1^{(k+1)}\right) + \frac{ks}{r}\left(f^\star - f(\tilde{x}^{(k+1)}) + \frac{L_f}{2} D_1^{(k+1)}\right) \\
&\quad - \frac{k\ell_R}{2r\lambda_k\gamma} D_1^{(k+1)} \\
&= \frac{k^2 s}{r^2}\left(f(\tilde{x}^{(k)}) - f(\tilde{x}^{(k+1)})\right) + \frac{ks}{r}\left(f^* - f(\tilde{x}^{(k+1)})\right) - \alpha_k D_2^{(k+1)} - \beta_k D_1^{(k+1)},
\end{aligned}$$

where

$$\beta_k = \frac{k\ell_R}{2r\lambda_k\gamma} - \frac{L_f k^2 s}{2r^2} - \frac{L_f k s}{2r}.$$

Finally, we obtain a bound on the difference $\tilde{E}^{(k+1)} - \tilde{E}^{(k)}$:

$$\begin{aligned}
&\tilde{E}^{(k+1)} - \tilde{E}^{(k)} \\
&= \frac{(k+1)^2 s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) - \frac{k^2 s}{r^2}(f(\tilde{x}^{(k)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&= \frac{k^2 s}{r^2}(f(\tilde{x}^{(k+1)}) - f(\tilde{x}^{(k)})) + \frac{(2k+1)s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) + D_{\psi^*}(z^{(k+1)}, z^\star) - D_{\psi^*}(z^{(k)}, z^\star) \\
&\leq \frac{(2k+1-kr)s}{r^2}(f(\tilde{x}^{(k+1)}) - f^\star) - \alpha_k D_2^{(k+1)} - \beta_k D_1^{(k+1)}
\end{aligned}$$

For the desired inequality to hold, it suffices that $\alpha_k, \beta_k \geq 0$, i.e.

$$\begin{aligned}
&\frac{1}{2L_{\psi^*}} - \frac{kL_R}{2(r+k)\gamma} \geq 0 \\
&\frac{k(r+k)\ell_R}{2r^2\gamma} - \frac{L_f k^2 s}{2r^2} - \frac{L_f k s}{2r} \geq 0,
\end{aligned}$$

i.e.

$$\begin{aligned}
&\gamma \geq \frac{k}{k+r} L_R L_{\psi^*} \\
&s \leq \frac{\ell_R}{L_f \gamma}.
\end{aligned}$$

So it is sufficient that

$$\gamma \geq L_R L_{\psi^*} \qquad\qquad s \leq \frac{\ell_R}{L_f \gamma}$$

which concludes the proof. □

Bibliography

[1] Zeyuan Allen-Zhu and Lorenzo Orecchia. "Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent". In: *CoRR* (2014).

[2] Felipe Alvarez and Hedy Attouch. "An Inertial Proximal Method for Maximal Monotone Operators via Discretization of a Nonlinear Oscillator with Damping". In: *Set-Valued Analysis* 9.1-2 (2001), pp. 3–11. ISSN: 0927-6947.

[3] Hedy Attouch, Juan Peypouquet, and Patrick Redont. "Fast Convergence of an Inertial Gradient-like System with Vanishing Viscosity". In: *CoRR* abs/1507.04782 (2015).

[4] Hédy Attouch, Juan Peypouquet, and Patrick Redont. "A Dynamical Approach to an Inertial Forward-Backward Algorithm for Convex Minimization". In: *SIAM Journal on Optimization* 24.1 (2014), pp. 232–256.

[5] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202.

[6] Amir Beck and Marc Teboulle. "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization". In: *Oper. Res. Lett.* 31.3 (May 2003), pp. 167–175.

[7] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. Society for Industrial and Applied Mathematics, 2001.

[8] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. "The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography". In: *SIAM J. on Optimization* 12.1 (Jan. 2001), pp. 79–108. ISSN: 1052-6234.

[9] Anthony Bloch, ed. *Hamiltonian and gradient flows, algorithms, and control*. American Mathematical Society, 1994.

[10] A. A. Brown and M. C. Bartholomew-Biggs. "Some Effective Methods for Unconstrained Optimization Based on the Solution of Systems of Ordinary Differential Equations". In: *Journal of Optimization Theory and Applications* 62.2 (1989), pp. 211–224. ISSN: 0022-3239.

[11] Sébastien Bubeck and Nicolò Cesa-Bianchi. "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122.

[12] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Ltd, 2008.

[13] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[14] Ofer Dekel et al. "Optimal Distributed Online Prediction". In: *Proceedings of the 28th International Conference on Machine Learning (ICML)*. June 2011.

[15] Simon Fischer and Berthold Vöcking. "On the evolution of selfish routing". In: *Algorithms–ESA 2004*. Springer, 2004, pp. 323–334.

[16] Nicolas Flammarion and Francis R. Bach. "From Averaging to Acceleration, There is Only a Step-size". In: *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*. 2015, pp. 658–695.

[17] U. Helmke and J.B. Moore. *Optimization and dynamical systems*. Communications and control engineering series. Springer-Verlag, 1994.

[18] Anatoli Juditsky. *Convex Optimization II: Algorithms, Lecture Notes*. 2013.

[19] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. "Solving variational inequalities with stochastic mirror-prox algorithm". In: *Stoch. Syst.* 1.1 (2011), pp. 17–58.

[20] H.K. Khalil. *Nonlinear systems*. Macmillan Pub. Co., 1992.

[21] Walid Krichene, Syrine Krichene, and Alexandre Bayen. "Efficient Bregman Projections onto the Simplex". In: *54th IEEE Conference on Decision and Control*. 2015.

[22] A.M. Lyapunov. *General Problem of the Stability Of Motion*. Control Theory and Applications Series. Taylor & Francis, 1992.

[23] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1983.

[24] Yu. Nesterov. "Gradient methods for minimizing composite functions". English. In: *Mathematical Programming* 140.1 (2013), pp. 125–161.

[25] Yu. Nesterov. "Smooth minimization of non-smooth functions". English. In: *Mathematical Programming* 103.1 (2005), pp. 127–152.

[26] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.

[27] Yurii Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Springer Science & Business Media, 2004.

[28] Brendan O'Donoghue and Emmanuel Candès. "Adaptive Restart for Accelerated Gradient Schemes". English. In: *Foundations of Computational Mathematics* 15.3 (2015), pp. 715–732. ISSN: 1615-3375.

[29] M. Raginsky and J. Bouvrie. "Continuous-time stochastic Mirror Descent on a network: Variance reduction, consensus, convergence". In: *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*. 2012, pp. 6793–6800.

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

[31] J. Schropp and I. Singer. "A dynamical systems approach to constrained minimization". In: *Numerical Functional Analysis and Optimization* 21.3-4 (2000), pp. 537–551.

[32] Weijie Su, Stephen Boyd, and Emmanuel Candès. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *NIPS*. 2014.

[33] Gerald Teschl. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Soc., 2012.

[34] Jörgen W Weibull. *Evolutionary game theory*. MIT press, 1997.

[35] Andre Wibisono and Ashia Wilson. "On Accelerated Methods in Optimization". In: *CoRR* (2015).