Linear Coupling:
An Ultimate Unification of Gradient and Mirror Descent

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
MIT CSAIL

Lorenzo Orecchia
orecchia@mit.edu
MIT Math

January 2, 2015

Abstract

First-order methods play a central role in large-scale convex optimization. Even though many variations exist, each suited to a particular problem form, almost all such methods fundamentally rely on two types of algorithmic steps and two corresponding types of analysis: gradient-descent steps, which yield primal progress, and mirror-descent steps, which yield dual progress. In this paper, we observe that the performances of these two types of step are complementary, so that faster algorithms can be designed by linearly coupling the two steps.

In particular, we obtain a simple accelerated gradient method for the class of smooth convex optimization problems. The first such method was proposed by Nesterov back to 1983 [Nes83, Nes04, Nes05], but to the best of our knowledge, the proof of the fast convergence of accelerated gradient methods has not found a clear interpretation and is still regarded by many as crucially relying on "algebraic tricks"[Jud13]. We apply our novel insights to construct a new accelerated gradient method as a natural linear coupling of gradient descent and mirror descent and to write its proof of convergence as a simple combination of the convergence analyses of the two underlying descent steps.

We believe that the complementary view and the linear coupling technique in this paper will prove very useful in the design of first-order methods as it allows us to design fast algorithms in a conceptually easier way. For instance, our technique greatly facilitates the recent breakthroughs in solving packing and covering linear programs [AO15, AO14].

1 Introduction

The study of fast iterative methods for approximately solving linear programs and, more generally, convex programming problems is a central focus of research in convex optimization, with important applications in Machine Learning, Combinatorial Optimizations and many other areas of Computer Science and Mathematics. The crowning jewel of this field of research has been the development of interior point methods, iterative methods that produce ε-additive approximations to the optimum with a small number of iterations and a logarithmic $\log\left(\frac{1}{\varepsilon}\right)$ dependence on the accuracy ε.

The fast rate of convergence of interior point methods comes at the cost of more expensive iterations, typically requiring the solution of a system of linear equations in the input variables. As a consequence, the cost of each iteration typically grows at least quadratically with the problem dimension, making interior point methods impractical for very-large-scale convex programs where the problem dimension is on the magnitude of millions or billions [BN13]. In such a regime, the methods of choice are first-order algorithms. These are modeled as accessing the target convex-optimization problem $\min_{x\in Q} f(x)$ in a black-box fashion: the algorithm queries a point $y \in Q$ at every iteration and receives the pair $\big(f(y), \nabla f(y)\big)$.[1] The convergence of the algorithm is measured in the number of queries necessary to produce a feasible solution which achieves an additive ε-approximation to the optimum.

Because of the restricted interaction with the input, first-order methods only require very cheap and often highly parallelizable iterations, which makes them well-suited to massive optimization problems. At the same time, first-order methods often require a number of iterations inversely polynomial to the accuracy ε, i.e. exponentially larger than required by interior-point algorithms.

Recently, first-order methods have experienced a renaissance in the design of fast algorithms for fundamental combinatorial problems. In particular, gradient-descent techniques play a crucial role in recent breakthroughs on the complexity of approximate maximum flow problems [LRS13, She13, KLOS14, Mad13]. At the same time, multiplicative weight updates, another first-order method and a cornerstone technique in online learning, have become a standard tool in the design of fast algorithms and have been applied with success to a variety of problems, including approximately solving linear and semidefinite relaxations of fundamental combinatorial problems [PST95, FS95, AHK05, AHK12] as well as spectral algorithms for graph problems [CKM$^+$11, OSV12].

Despite the myriad of applications, first-order methods with provable convergence guarantees can be mostly classified as instantiations of two fundamental algorithmic ideas: *gradient descent* and the *mirror descent*.[2]

A method with provable guarantees must provide both a solution x_{out} and an implicit or explicit certificate that x_{out} in the form of a lower bound on the optimum. We refer to the task of constructing a solution x_{out} of small objective as the primal side of the problem and to that of constructing a lower bound on the optimum as the dual side.

We will argue that gradient descent takes a fundamentally primal approach, while mirror descent follows a complementary dual approach. In our main result, we will show how these two approaches blend in a natural manner to yield a new and simple accelerated gradient method for smooth convex optimization problems.

[1] Here, variable x is constrained to lie in a convex set $Q \subseteq \mathbb{R}^n$, which is known as the *constraint set* of the problem.

[2] We emphasize here that these two terms are sometimes used ambiguosly in the literature; in this paper, we attempt to stick as close as possible to the conventions of the Optimization community and in particular in the textbooks [Nes04, BN13] with one exception: we extend the definition of gradient descent to non-Euclidean norms in a natural way, following [KLOS14].

1.1 Understanding First-Order Methods: Gradient Descent and Mirror Descent

In this section, we provide high-level descriptions of the gradient-descent and the mirror-descent algorithms and their analysis. While much of this material is classical in the field of optimization, our intuitive presentation of these ideas forms the basis for our main result. For a more detailed survey of gradient descent and mirror descent, we recommend the textbooks [Nes04, BN13].

For the purpose of this section, we only consider the case of unconstrained minimization (i.e. $Q = \mathbb{R}^n$), but, as we will see in Section 2, the same intuition and a similar analysis extend to the constrained case. In the following, we will also be using generic dual norms $\|\cdot\|$ and $\|\cdot\|_*$. At a first reading, they can be both replaced with the Euclidean norm $\|\cdot\|_2$.

1.1.1 Primal Approach: Gradient Descent for Smooth Convex Optimization

A natural approach to iterative optimization is to decrease the objective function as much as possible at every iteration. To formalize the effectiveness of this idea, one has to introduce an additional smoothness assumption on the objective function $f(x)$; specifically, this is achieved by considering the class of objectives that are L-smooth (i.e., that have L-Lipschitz continuous gradient):

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

The smoothness condition immediately yields a global quadratic upper bound on the function around a query point x:

$$\forall y, \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\| \ . \tag{1.1}$$

The gradient-descent algorithm exploits this bound by taking a step that maximizes the guaranteed objective decrease (i.e., the primal progress) $f(x_k) - f(x_{k+1})$ at every iteration k. More precisely,

$$x_{k+1} \leftarrow \arg\min_y \left\{ \frac{L}{2}\|y - x_k\|^2 + \langle \nabla f(x), y - x_k \rangle \right\} \ .$$

Notice that here $\|\cdot\|$ is a generic norm. When this is the Euclidean ℓ_2-norm, the step takes the familiar additive form $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. However, in other cases, e.g., for the non-Euclidean ℓ_1 or ℓ_∞ norms, the update step will not follow the direction of the gradient $\nabla f(x_k)$ (see for instance [Nes05, KLOS14]).

Under the smoothness assumption above, the magnitude of this primal progress is at least

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|_*^2 \ . \tag{1.2}$$

In general, this quantity will be larger when the gradient $\nabla f(x_k)$ has large norm.

Inequality (1.2) ensures that at every iteration the objective value of the current solution x_k decreases by at least $\frac{1}{2L}\|\nabla f(x_k)\|_*^2$. The proof of convergence of gradient descent is completed by using a basic convexity argument to relate $f(x_k) - f(x^*)$ and $\|\nabla f(x_k)\|_*$ (where x^* is the minimizer of $f(x)$). The final bound shows that the algorithm converges to an ε-approximate solution in $O\left(\frac{L}{\varepsilon}\right)$ iterations [Nes04]. More details on the gradient-descent algorithm and its analysis are given in Section 2.1 and in Nesterov's book [Nes04].

In conclusion, it is useful to think of gradient descent as choosing query points in a greedy way to ensure the largest possible primal progress at every iteration. The limitation of this strategy is that it does not make any attempt to construct a good lower bound to the optimum value, i.e., it essentially ignores the dual problem. In the next subsection, we will see a method that takes the opposite approach by focusing completely on the dual side. This method is suitable when there is no guarantee on the smoothness of the objective.

1.1.2 Dual Approach: Mirror Descent for Nonsmooth Convex Optimization

In non-smooth convex optimization, we are given an upper bound ρ on the Lipschitz constant of $f(x)$, rather than $\nabla f(x)$. When f is differentiable, this means that the gradient could change arbitrarily fast, but its norm remains bounded, i.e., $\|\nabla f(x)\| \leq \rho$ for every $x \in Q$. The possibility that the gradient varies quickly seriously undermines the performance of gradient descent, which relies on making a certain amount of primal progress at every iteration. In this case, it is not possible to guarantee that an update step of a predetermined length would result in an improved objective value, as the gradient may be very large even at points very near the optimum. At the same time, we cannot afford to take too small steps as this limits our rate of convergence.

Dual-averaging methods (see for instance [NY78, Nes07, DSST10, Xia10, BN13]) bypass this obstacle by tackling the dual problem of constructing a lower bound to the optimum. They interpret each queried gradient as a hyperplane lower bounding the objective function $f(x)$ and attempt to carefully construct a convex combination of these hyperplanes that yields a stronger lower bound. Intuitively, the flatter the queried gradients are (i.e. the smaller $\|\nabla f(x_k)\|_* \leq \rho$ is), the fewer iterations will be needed to combine them into an approximately optimal solution.

Formally, at each iteration k, using the convexity of $f(x)$, we can consider the following lower bound implied by the gradient $\nabla f(x_k)$:

$$\forall u, \quad f(u) \geq f(x_k) + \langle \nabla f(x_k), u - x_k \rangle \ .$$

To get a stronger lower bound, we can form a linear combination of the lower bounds given by all the queried gradients, and obtain[3]

$$\forall u, \quad f(u) \geq \tfrac{1}{T} \textstyle\sum_{t=0}^{T-1} f(x_k) + \tfrac{1}{T} \sum_{t=0}^{T-1} \langle \nabla f(x_k), u - x_k \rangle \ . \tag{1.3}$$

On the upper bound side, we consider the point $\overline{x} = \frac{1}{T} \sum_{k=0}^{T-1} x_k$, i.e., the mean of the queried points. By straightforward convexity argument, we have $f(\overline{x}) \leq \frac{1}{T} \sum_{k=0}^{T-1} f(x_k)$. As a result, we can upper bound the distance between $f(\overline{x})$ and $f(u)$ for any arbitrary u using (1.3):

$$\forall u, \quad f(\overline{x}) - f(u) \leq \tfrac{1}{T} \textstyle\sum_{k=0}^{T-1} \langle \nabla f(x_k), x_k - u \rangle \stackrel{\text{def}}{=} R_T(u) \ . \tag{1.4}$$

Borrowing terminology from online learning, the righthand side $R_T(u)$ is known as the *regret* of the sequence $(x_k)_{k=0}^{T-1}$ with respect to point u.

Dual Averaging via Regularization: Mirror Descent. We are aware of two main algorithmic instantiations of dual averaging: *Nemirovski's mirror descent* [NY78] and *Nesterov's dual averaging* [Nes07].[4] Both these algorithm make use of a *regularizer* $w(\cdot)$, also known as the distance-generating function (DGF), which is a strongly convex function over Q with respect to some norm $\|\cdot\|$. The two methods are very similar, differing only in how the constraint set is integrated in the update step [McM11]. In fact, they are exactly identical in the unconstrained case $Q = \mathbb{R}^n$ and, more generally, when $w(\cdot)$ enjoys some nice properties (see Appendix A.3). Below, we focus on the unconstrained case.

Both algorithms consider a regularized version $\widetilde{R}_k$ of the regret in (1.4):

$$\widetilde{R}_k(u) \stackrel{\text{def}}{=} \frac{1}{\alpha k} \cdot \Big(-w(u) + \alpha \sum_{i=0}^{k-1} \langle \nabla f(x_i), x_i - u \rangle \Big) \ ,$$

where $\alpha > 0$ is a trade-off parameter. Notice that an upper bound on $\widetilde{R}_k(u)$ can be simply

[3] For simplicity, we choose uniform weights here. For the purpose of proving convergence results, the weights of individual hyperplanes are typically uniform or only dependent on k.

[4] Several other update rules can be viewed as specializations or generalizations of the mentioned instantiations. For instance, the follow-the-regularized-leader (FTRL) step is a generalization of Nesterov's dual averaging step where the regularizers are allowed to be adaptively and incrementally selected (see [MS10]).

converted into one for $R_k(u)$ with an additive loss: $R_k(u) \leq \widetilde{R}_k(u) + \frac{w(u)}{\alpha k}$. Both Nemirovski's mirror descent and Nesterov's dual averaging attempt to minimize the maximum regularized regret at the next iteration (i.e., $\max_u \widetilde{R}_{k+1}(u)$), by choosing the next query point x_k to be the maximizer of the current regularized regret (i.e., $\arg\max_u \widetilde{R}_k(u)$). It turns out that this choice of query point successfully drives $\max_u \widetilde{R}_{k+1}(u)$ down. In fact, the smaller the queried gradient $\nabla f(x_k)$ is, the smaller the new maximum regularized regret $\max_u \widetilde{R}_{k+1}(u)$ will be. In general, one can show that:

$$\max_u \widetilde{R}_{k+1}(u) \leq \frac{k}{k+1} \max_u \widetilde{R}_k(u) + O\Big(\frac{\alpha}{k+1}\|\nabla f(x_k)\|_*^2\Big). \tag{1.5}$$

This bound can then be turned into a convergence proof requiring $T = O(\rho^2/\varepsilon^2)$ iterations.

We remark that the convergence argument sketched here crucially relies on the use of the regularized regret (instead of the original regret). In particular, Inequality (1.5) directly follows from a smoothness property of the maximum regularized regret with respect to the addition of new gradient hyperplanes, which only holds when the regularizer $w(u)$ is strongly convex. For more details of this view of dual averaging and the proof of (1.5), see Appendix A.4.

This paper. In this paper, we adopt *mirror descent* as our dual algorithm of choice, as it is more familiar to the Theoretical Computer Science audience. Indeed, the most common instantiation of mirror descent is perhaps the multiplicative-weight-update algorithm, which has become a standard tool in the design of algorithms [AHK12] (see Appendix A.2 for this relationship). We describe the mirror descent step for the constrained case and its analysis in Section 2.2. A great resource for an in-depth description of mirror descent is the textbook by Ben-Tal and Nemirovski [BN13].

1.1.3 Remark: A Few Exceptions

One may occasionally find analyses that do not immediately fall into the above two categories. To name a few, Dekel *et al.* [DGSX12] have applied dual averaging steps to a *smooth* objective, and shown that the convergence rate is the same as that of gradient descent. Shamir and Zhang [SZ13c] have studied non-smooth objectives and obtained an algorithm that converges slightly slower than dual averaging, but has an error guarantee on the last iterate, rather than the average history.

1.2 Our Conceptual Question

Following this high level description of gradient and mirror descent, it is useful to pause and observe the complementary nature of the two procedures. Gradient descent relies on primal progress, uses local steps and makes faster progress when the norms of the queried gradients $\nabla f(x_k)$ are large. In contrast, mirror descent works by ensuring dual progress, uses global steps and converges faster when the norms of the queried gradients are small.

This interpretation immediately leads to the question that inspires our work:

Can Gradient Descent and Mirror Descent be combined to obtain faster first-order algorithms?

In this paper, we initiate the formal study of this key conceptual question. We believe that the techniques and insights to answer this question have the potential to lead to faster and better motivated algorithms for many more computational problems.

1.3 Accelerated Gradient Method From Linear Coupling

In the seminal work [Nes83, Nes04], Nesterov has designed an accelerated gradient method for the class of L-smooth functions with respect to ℓ_2 norms, and this method performs quadratically faster than gradient descent —requiring $\Omega(L/\varepsilon)^{0.5}$ rather than $\Omega(L/\varepsilon)$ iterations. This is also shown to be asymptotically tight [Nes04]. Later in 2005, Nesterov himself generalizes this method to allow non-Euclidean norms in the definition of smoothness [Nes05]. All these versions of methods are referred to as *accelerated gradient methods*, or sometimes as Nesterov's accelerated methods.

Although accelerated gradient methods have been widely applied (to mention a few, see [SZ13a, SZ13b] for regularized optimizations, [Nes13, Lan11] for composite optimization, [Nes08] for cubic regularization, [Nes14] for universal method, and [LRS13] for an application on maxflow), little geometric explanation is known. For instance, Juditsky [Jud13] has mentioned that Nesterov's method "looks as an analytical trick."

In this paper, we provide a simple, alternative, but **complete** version of the accelerated gradient method. Here, by 'complete' we mean our method works for any norm, and for both the constrained and unconstrained case. This is in contrast with the (perhaps better-known) version of Nesterov [Nes04] that only works with the ℓ_2 Euclidean norm.[5]

Instead of using the *estimation sequence* technique provided in the original proof of Nesterov, we take a different path. Our key observation is to construct two sequences of updates: one sequence of gradient steps and one sequence of mirror steps. Recall that, according to the gradient-descent and mirror-descent analyses described above, the gradient steps perform well whenever the observed gradients are large; the mirror steps perform well whenever the observed gradients are small. Thus, intuitively, we hope to *couple* these two steps together, and choose the better method 'adaptively' according to the size of the gradient. We begin with a thought experiment.

Thought Experiment. Consider the case when the smooth property is with respect to the ℓ_2-norm, and the objective $f(x)$ is unconstrained. Suppose that $\|\nabla f(x)\|_2$, the size of the observed gradient, is *either* always $\geq K$, or always $\leq K$, where the cut-off value K is determined later. If $\|\nabla f(x)\|_2$ is always $\geq K$, we perform T gradient steps; otherwise we perform T mirror steps. Suppose in addition that we start with some $f(x_0)$ whose distance to $f(x^*)$ is at most 2ε, and we want to obtain some x so that $f(x) - f(x^*) \leq \varepsilon$.[6]

If T gradient steps are conducted, in each step the objective decreases by at least $\frac{\|\nabla f(\cdot)\|_2^2}{2L} \geq \frac{K^2}{2L}$ according to (1.2), and thus we only need to choose $T \geq \Omega(\frac{\varepsilon L}{K^2})$ steps in order to achieve an ε accuracy. On the other hand, if T mirror steps are conducted, we need $T \geq \Omega(\frac{K^2}{\varepsilon^2})$ steps according to the mirror-descent convergence. In sum, in this thought experiment, we need $T \geq \Omega\big(\max\big\{\frac{\varepsilon L}{K^2}, \frac{K^2}{\varepsilon^2}\big\}\big)$ steps to achieve a solution ε-close to the optimum.

Now, setting K to be the 'magic number' so that the two terms in the max function equal, we obtain $T \geq \Omega\big(\frac{L}{\varepsilon}\big)^{1/2}$. This is a quadratic improvement over $T \geq \Omega(\frac{L}{\varepsilon})$ from the gradient descent.

Towards the Actual Proof. To turn this thought experiment into an actual proof, we are facing the following obstacles. The gradient steps always decrease the objective, while the mirror step may very often increase the objective, cancelling the effect of the gradient steps. On the other hand, the mirror steps are *only* useful when a large number of iterations are performed in a row, and the performance guarantee is on the average of these iterations; if any primal step stands in the middle, this guarantee is destroyed.

Therefore, it is natural to design an algorithm that, in every single iteration k, performs *both* a gradient and a mirror step, and somehow ensure that the two steps are coupled together. However, the following additional difficulty arises: if from some starting point x_k, the gradient step instructs us to go to y_k, while the mirror step instructs us to go to z_k, then how do we continue? Do we look at the gradient at $\nabla f(y_k)$ or $\nabla f(z_k)$? In particular, if $\|\nabla f(y_k)\|_2$ is large, we can continue

[5]Some authors have regarded the result in [Nes04] as the 'momentum analysis' or 'momentum method' [OC13, SBC14]. To the best of our knowledge, all the momentum analysis only applies to Euclidean spaces. We point out the importance of allowing non-Euclidean norms in Appendix A.1. (Our proof also extends to the proximal version of first-order methods, but for simplicity, we choose to include only the constrained version.)

[6]It is worth noting that for first-order methods, the heaviest computation always happens in this 2ε to ε procedure.

performing gradient steps from y_k; or if $\|\nabla f(z_k)\|_2$ is small, we can continue performing mirror steps from z_k. However, what if $\|\nabla f(y_k)\|_2$ is small but $\|\nabla f(z_k)\|_2$ is large?

This problem is implicitly solved by Nesterov using the following simple idea[7]: in the k-th step, we can choose a linear combination $x_{k+1} \leftarrow \tau z_k + (1-\tau) y_k$, and use this same gradient $\nabla f(x_{k+1})$ to continue the gradient and mirror steps. Whenever τ is carefully chosen (just like the 'magic number' K being selected), the two descent sequences provide a coupled bound on the error guarantee, and we recover the method of [Nes05].

Finally, we point out that our method also recovers the strong convexity version of [Nes04], and therefore is a full proof to all existing versions of accelerated gradient methods for smooth convex optimization problems.

1.4 Conclusion

We provide a simple variant of the accelerated gradient method with a reinterpretation of its convergence analysis. Providing such an intuitive, yet formal interpretation has been a long-open question in Optimization [Jud13]. We believe that our interpretation is one important step towards this general goal, and may facilitate the study of accelerated gradient methods in a white-box manner, so as to apply them to problems outside its original scope.

In addition, we believe that our complementary view of gradient descent and mirror descent is a very fundamental (and to the best of our knowledge, new!) conceptual message in the design of first-order methods. This has the potential to lead to faster and better motivated algorithms for many more computational problems. Indeed, we have already succeeded in this direction in our separate papers [AO15, AO14], where we have proposed faster nearly-linear-time algorithms for approximately solving positive linear programs, both in parallel and in sequential.[8]

2 Preliminaries

2.1 Review of Primal Descent

Consider a function $f(x)$ that is convex and differentiable on a closed convex set $Q \subseteq \mathbb{R}^n$,[9] and assume that f is *L-smooth* (or has L-Lipschitz continuous gradient) with respect to $\|\cdot\|$, that is

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in Q$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.[10]

Definition 2.1. *For any $x \in Q$, the* gradient (descent) step *(with step length $\frac{1}{L}$) is*

$$\widetilde{x} = \mathsf{Grad}(x) \stackrel{\text{def}}{=} \arg\min_{y\in Q} \left\{ \frac{L}{2}\|y - x\|^2 + \langle \nabla f(x), y - x\rangle \right\}$$

and we let $\mathsf{Prog}(x) \stackrel{\text{def}}{=} -\min_{y\in Q}\{\frac{L}{2}\|y - x\|^2 + \langle \nabla f(x), y - x\rangle\} \geq 0.$

---

[7] We wish to point out that Nesterov has phrased his method differently from ours, and little is known on why this linear combination is needed from his proof, except for being used as an algebraic trick to cancel specific terms.

[8] In our paper [AO15], we have designed an iterative algorithm whose update steps can be viewed both as gradient and as mirror steps, therefore allowing us to apply two complementary analyses to support each other; this breaks the $O(1/\varepsilon^4)$ barrier in the parallel packing/covering LP solver running time since [LN93].

In our paper [AO14], we have designed algorithms whose update steps can be viewed as linear couplings of (the coordinates version of) gradient and mirror steps; this breaks the $O(1/\varepsilon^2)$ barrier in the sequential packing/covering LP solver running time since [BBR97, You01, BBR04].

Neither of the two papers is any direct variant of accelerated gradient methods, and their objectives are not even smooth.

[9] In most of the applications, Q is simple enough so that the gradient steps (and mirror steps as well) can be computed explicitly and efficiently. For instance, one may use the positive orthant, $Q = \{x \in \mathbb{R}^n : x \geq 0\}$, the unit sphere, $Q = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, and many others.

[10] $\|\xi\|_* \stackrel{\text{def}}{=} \max\{\langle \xi, x\rangle : \|x\| \leq 1\}$. For instance, ℓ_p norm is dual to ℓ_q norm if $\frac{1}{p} + \frac{1}{q} = 1$.

In particular, when $\|\cdot\| = \|\cdot\|_2$ is the ℓ_2-norm and $Q = \mathbb{R}^n$ is unconstrained, the gradient step can be simplified as $\mathsf{Grad}(x) = x - \frac{1}{L}\nabla f(x)$. Or, slightly more generally, when $\|\cdot\| = \|\cdot\|_2$ is the ℓ_2-norm but Q may be constrained, we have $\mathsf{Grad}(x) = x - \frac{1}{L}g_Q(x)$ where $g_Q(x)$ is the gradient mapping of f at x (see [Nes04, Chapter 2.2.3]).

The classical theory on smooth convex programming gives rise to the following lower bound on the amount of objective decrease (whose proof is provided in Appendix B for completeness).

Gradient Descent Guarantee

$$f(\mathsf{Grad}(x)) \leq f(x) - \mathsf{Prog}(x) \tag{2.1}$$

or in the special case when $Q = \mathbb{R}^n$ $\quad f(\mathsf{Grad}(x)) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_*^2$.

From the above descent guarantee, one can deduce the convergence rate of the gradient descent steps. In particular, if $\|\cdot\| = \|\cdot\|_2$ is the Euclidean norm, and the gradient step $x_{k+1} = \mathsf{Grad}(x_k)$ is applied T times, we obtain the following convergence guarantee (see [Nes04, Chapter 2.1.5])

$$f(x_T) - f(x^*) \leq O\Big(\frac{L\|x_0 - x^*\|_2^2}{T}\Big) \quad \text{or equivalently} \quad T \geq \Omega\Big(\frac{L\|x_0 - x^*\|_2^2}{\varepsilon}\Big) \Rightarrow f(x_T) - f(x^*) \leq \varepsilon \ .$$

Here, x^* is any minimizer of $f(x)$. If $\|\cdot\|$ is a general norm, but $Q = \mathbb{R}^n$ is unconstrained, the above convergent rate becomes $f(x_T) - f(x^*) \leq O\big(\frac{LR^2}{T}\big)$, where $R = \max_{x: f(x) \leq f(x_0)} \|x - x^*\|$. We provide the proof of this later case in Appendix B because it is less known and we cannot find it in the optimization literature.

Note that, we are unaware of any universal convergence proof for both the general norm and the unconstrained case. As we shall see later in Section 4, this convergence rate can be improved by accelerated gradient methods, even for the general norm $\|\cdot\|$ and the constrained case.

2.2 Review of Mirror Descent

Consider some function $f(x)$ that is convex on a closed convex set $Q \subseteq \mathbb{R}^n$, and assume that f is *ρ-Lipschitz continuous* with respect to norm $\|\cdot\|$, that is

$$|f(x) - f(y)| \leq \rho\|x - y\|, \quad \forall x, y \in Q \ .$$

Notice that this is equivalent to saying that f admits a subgradient $\partial f(x)$ at every point $x \in Q$, and satisfies $\|\partial f(x)\|_* \leq \rho$ for all x. (Recall that $\partial f(x) = \nabla f(x)$ if f is differentiable.)

The mirror descent method requires one to choose a distance generating function.

Definition 2.2. *We say that $w(x)\colon Q \to \mathbb{R}$ is a* distance generating function (DGF)*, if w is 1-strongly convex with respect to $\|\cdot\|$, or in symbols*

$$w(y) \geq w(x) + \langle\nabla w(x), y - x\rangle + \frac{1}{2}\|x - y\|^2 \quad \forall x \in Q \setminus \partial Q, \forall y \in Q \ .$$ [11]

Accordingly, the Bregman divergence *(or* prox-term*) is given as*

$$V_x(y) \stackrel{\text{def}}{=} w(y) - \langle\nabla w(x), y - x\rangle - w(x) \quad \forall x \in Q \setminus \partial Q, \forall y \in Q \ .$$

The property of DGF ensures that $V_x(x) = 0$ and $V_x(y) \geq \frac{1}{2}\|x - y\|^2 \geq 0$.

Common examples of DGFs include `(i)` $w(y) = \frac{1}{2}\|y\|_2^2$, which is strongly convex with respect to the ℓ_2-norm over any convex set Q, and the corresponding $V_x(y) = \frac{1}{2}\|x - y\|_2^2$, and `(ii)` the entropy function $w(y) = \sum_i y_i \log y_i$, which is strongly convex with respect to the ℓ_1-norm over any $Q \subseteq \Delta \stackrel{\text{def}}{=} \{x \geq 0 \,:\, \mathbb{1}^T x = 1\}$, and the corresponding $V_x(y) = \sum_i y_i \log(y_i/x_i) \geq \frac{1}{2}\|x - y\|_1^2$.

[11] One can in fact only require w to have subgradients at all $x \in Q \setminus \partial Q$.

Definition 2.3. *The* mirror (descent) step *with step length α can be described as*

$$\widetilde{x} = \mathsf{Mirr}_x(\alpha \cdot \partial f(x)) \quad \textit{where} \quad \mathsf{Mirr}_x(\xi) \stackrel{\text{def}}{=} \arg\min_{y \in Q} \big\{ V_x(y) + \langle \xi, y - x \rangle \big\}$$

The core lemma of mirror descent is the following inequality. (Its proof can be found in Appendix B for completeness.)

Mirror Descent Guarantee

If $x_{k+1} = \mathsf{Mirr}_{x_k}\big(\alpha \cdot \partial f(x_k)\big)$, then

$$\forall u \in Q, \quad \alpha(f(x_k) - f(u)) \le \alpha \langle \partial f(x_k), x_k - u \rangle \le \frac{\alpha^2}{2} \|\partial f(x_k)\|_*^2 + V_{x_k}(u) - V_{x_{k+1}}(u) \ . \qquad (2.2)$$

The term $\langle \partial f(x_k), x_k - u \rangle$ features prominently in online optimization (see for instance the survey [Sha11]), where it is known as the *regret* at iteration k with respect to u.[12] It is not hard to see that, after telescoping (2.2) for $k = 0, \dots, T-1$, letting $\overline{x} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{k=0}^{T-1} x_k$ be the average of the x_k's, and letting x^* be the minimizer of $f(x)$, we have

$$\alpha T(f(\overline{x}) - f(x^*)) \le \sum_{k=0}^{T-1} \alpha \langle \partial f(x_k), x_k - x^* \rangle \le \frac{\alpha^2}{2} \sum_{k=0}^{T-1} \|\partial f(x_k)\|_*^2 + V_{x_0}(x^*) - V_{x_T}(x^*) \ . \qquad (2.3)$$

Finally, letting Θ be any upper bound on $V_{x_0}(x^*)$, and $\alpha = \frac{\sqrt{2\Theta}}{\rho \cdot \sqrt{T}}$ be the step length, inequality (2.2) ensures that

$$f(\overline{x}) - f(x^*) \le \frac{\sqrt{2\Theta} \cdot \rho}{\sqrt{T}} \quad \text{or equivalently} \quad T \ge \frac{2\Theta \cdot \rho^2}{\varepsilon^2} \Rightarrow f(\overline{x}) - f(x^*) \le \varepsilon \ . \qquad (2.4)$$

Notice that $\Theta = \frac{1}{2}\|x_0 - x^*\|_2^2$ when $\|\cdot\|$ is the Euclidean norm.

2.3 Remark

While their analyses share some similarities, mirror and gradient steps are often very different. This is particularly true when working with non-Euclidean norms. For example, if we consider an optimization problem over the simplex with underlying norm ℓ_1-norm, the gradient step gives $x' \leftarrow \arg\min_y \{\frac{1}{2}\|y - x\|_1^2 + \alpha \langle \nabla f(x), y - x \rangle\}$, while the mirror step with entropy regularizer gives $x' \leftarrow \arg\min_y \{\sum_i y_i \log(y_i/x_i) + \alpha \langle \nabla f(x), y - x \rangle\}$. We shall point out in Appendix A.1 that non-Euclidean norms are very important for certain applications.

In the special case of $w(x) = \frac{1}{2}\|x\|_2^2$ and $\|\cdot\| = \|\cdot\|_2$, gradient and mirror steps are indistinguishable from each other. However, as we have discussed earlier, these two update rules are often equipped with very different convergence analyses, even if they 'look the same'.

3 Warm-Up Accelerated Gradient Method with Fixed Step Length

We adopt the same setting as in Section 2.1: that is, $f(x)$ is convex and differentiable on its domain Q, and is L-smooth with respect to some norm $\|\cdot\|$. (Note that $f(x)$ may not have a good Lipschitz continuity parameter ρ, but we do not need such a property.)

In this section, we focus on the unconstrained case of $Q = \mathbb{R}^n$, and wish to combine gradient descent and mirror descent to produce a very simple accelerated method, which matches the running time of Nesterov's. We choose to explain this method first because it avoids the mysterious choice of the step lengths in the full accelerated gradient methods, and carries our conceptual message in a very clean way.

[12] The notion of *regret* is especially used in the language of multiplicative weight update methods, which can be viewed as mirror descent, see Appendix A.2.

As argued in Section 1.3, it is desirable to design an algorithm that, in every single step k, performs *both* a gradient and a mirror step, and ensures that the two steps are linearly coupled. In particular, we consider the following steps: starting from $x_0 = y_0 = z_0$, in each step $k = 0, 1, \dots, T-1$, we first compute $x_{k+1} \leftarrow \tau z_k + (1-\tau) y_k$ and then

- perform a gradient step $y_{k+1} \leftarrow \mathsf{Grad}(x_{k+1})$, and
- perform a mirror step $z_{k+1} \leftarrow \mathsf{Mirr}_{z_k}\big(\alpha \nabla f(x_{k+1})\big)$.[13]

Above, α is the (fixed) step length of the mirror step, while τ is the parameter controlling our coupling. The choices of α and τ will become clear at the end of this section, but from a high level,

- α will be determined from the mirror-descent analysis, similar to that in (2.3), and
- τ will be determined as the best parameter to balance the gradient and mirror steps, similar to the 'magic number' K in our thought experiment discussed in Section 1.3.

The classical gradient-descent and mirror-descent analyses immediately imply the following

Lemma 3.1. *For every $u \in Q = \mathbb{R}^n$,*

$$\begin{aligned} \alpha \langle \nabla f(x_{k+1}), z_k - u \rangle &\overset{①}{\leq} \frac{\alpha^2}{2} \|\nabla f(x_{k+1})\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u) \\ &\overset{②}{\leq} \alpha^2 L \big(f(x_{k+1}) - f(y_{k+1}) \big) + V_{z_k}(u) - V_{z_{k+1}}(u) \ . \end{aligned} \tag{3.1}$$

Proof. To deduce ①, we note that our mirror step $z_{k+1} = \mathsf{Mirr}_{z_k}(\alpha \nabla f(x_{k+1}))$ is essentially identical to that of $x_{k+1} = \mathsf{Mirr}_{x_k}(\alpha \nabla f(x_k))$ in (2.2), with only changes of variable names. Therefore, inequality ① is a simple copy-and-paste from (2.2) after changing the variable names (see the proof of (2.2) for details). The second inequality ② is from the gradient step guarantee $f(x_{k+1}) - f(y_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_{k+1})\|_*^2$ in (2.1). □

One can already see from the above Lemma 3.1 that, although the mirror step introduces an error of $\frac{\alpha^2}{2} \|\nabla f(x_{k+1})\|_*^2$, this error is proportional to the amount of the gradient step progress $f(x_{k+1}) - f(y_{k+1})$. To be clear, this captures the observation we have stated in the introduction: if $\|\nabla f(x_{k+1})\|_*$ is large, we can make a large gradient step, or if $\|\nabla f(x_{k+1})\|_*$ is small, the mirror step suffers from a small loss.

At this moment, if we choose $\tau = 1$ or equivalently $x_{k+1} = z_k$, the left hand side of inequality (3.1) gives us $\langle \nabla f(x_{k+1}), x_{k+1} - u \rangle$, the regret at iteration x_{k+1}. We therefore wish to telescope it for all choices of k in the spirit as mirror descent (see (2.3)); however, we face the problem that the terms $f(x_{k+1}) - f(y_{k+1})$ do not telescope.[14] On the other hand, if we choose $\tau = 0$ or equivalently $x_{k+1} = y_k$, then the terms $f(x_{k+1}) - f(y_{k+1}) = f(y_k) - f(y_{k+1})$ telescope, but the left hand side of (3.1) is no longer the regret.[15]

To overcome this issue, we need the linear coupling. We compute and upper bound the difference between the left hand side of (3.1) and the real 'regret':

$$\begin{aligned} &\alpha \langle \nabla f(x_{k+1}), x_{k+1} - u \rangle - \alpha \langle \nabla f(x_{k+1}), z_k - u \rangle \\ &= \alpha \langle \nabla f(x_{k+1}), x_{k+1} - z_k \rangle = \frac{(1-\tau)\alpha}{\tau} \langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle \leq \frac{(1-\tau)\alpha}{\tau} \big(f(y_k) - f(x_{k+1}) \big). \end{aligned} \tag{3.2}$$

[13] Here, the mirror step Mirr is defined by specifying any DGF $w(\cdot)$ that is 1-strongly convex over Q.

[14] In other words, although a gradient step may decrease the objective from $f(x_{k+1})$ to $f(y_{k+1})$, it may also get the objective increased from $f(y_k)$ to $f(x_{k+1})$.

[15] Indeed, our "thought experiment" in the introduction is conducted *as if* we both had $x_{k+1} = z_k$ and $x_{k+1} = y_k$, and therefore we could arrive at the desired (3.3) directly.

Above, we have used the choice of x_{k+1} that satisfies $\tau(x_{k+1} - z_k) = (1-\tau)(y_k - x_{k+1})$, as well as the convexity of $f(\cdot)$.

It is now clear that by choosing $\frac{1-\tau}{\tau} = \alpha L$ and combining (3.1) and (3.2), we immediately have

Lemma 3.2 (Coupling)**.** *Letting $\tau \in (0,1)$ satisfy that $\frac{1-\tau}{\tau} = \alpha L$, we have that*

$$\forall u \in Q = \mathbb{R}^n, \quad \alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle \le \alpha^2 L\big(f(y_k) - f(y_{k+1})\big) + \big(V_{z_k}(u) - V_{z_{k+1}}(u)\big) \ .$$

It is clear from the above proof that τ is introduced to precisely balance the objective decrease $f(x_{k+1}) - f(y_{k+1})$, and the (possible) objective increase $f(y_k) - f(x_{k+1})$. This is similar to the 'magic number' K discussed in the introduction.

Convergence Rate. Finally, we only need to telescope the inequality in Lemma 3.2 for $k = 0, 1, \dots, T-1$. Letting $\overline{x} \stackrel{\text{def}}{=} \frac{1}{T}\sum_{k=0}^{T-1} x_k$ and $u = x^*$, we have

$$\alpha T(f(\overline{x}) - f(x^*)) \le \sum_{k=0}^{T-1} \alpha\langle \partial f(x_k), x_k - x^*\rangle \le \alpha^2 L\big(f(y_0) - f(y_T)\big) + V_{x_0}(x^*) - V_{x_T}(x^*) \ . \quad (3.3)$$

Suppose that our initial point y_0 is of error at most d (i.e., $f(y_0) - f(x^*) \le d$), and $V_{x_0}(x^*) \le \Theta$, then (3.3) gives that

$$f(\overline{x}) - f(x^*) \le \frac{1}{T}\big(\alpha L d + \Theta/\alpha\big) \ .$$

Choosing $\alpha = \sqrt{\Theta/Ld}$ to be the value that balances the above two terms,[16] we obtain that $f(\overline{x}) - f(x^*) \le \frac{2\sqrt{L\Theta d}}{T}$. In other words,

in $T = 4\sqrt{L\Theta/d}$ steps, we can obtain some $\overline{x}$ satisfying $f(\overline{x}) - f(x^*) \le d/2$,

halving the distance to the optimum. If we restart this entire procedure a few number of times, halving the distance for every run, then we obtain an ε-approximate solution in

$$T = O\big(\sqrt{L\Theta/\varepsilon} + \sqrt{L\Theta/2\varepsilon} + \sqrt{L\Theta/4\varepsilon} + \cdots\big) = O\big(\sqrt{L\Theta/\varepsilon}\big)$$

iterations, matching the same guarantee of Nesterov's accelerated methods [Nes83, Nes04, Nes05].

It is important to note here that $\alpha = \sqrt{\Theta/Ld}$ increases as time goes (i.e., as d goes down), and therefore $\tau = \frac{1}{\alpha L + 1}$ decreases as time goes. This lesson instructs us that gradient steps should be given more weights than mirror steps, when it is closer to the optimum.[17]

Conclusion. Equipped with the basic knowledge of gradient descent and mirror descent, the above proof is quite straightforward and also gives intuition to how the two 'magic numbers' α and τ are selected. We are unaware of any similar accelerated gradient method that uses fixed step length like ours (when the objective is not known to be strongly convex).

However, this simple algorithm has several caveats. First, the value α depends on the knowledge of Θ; second, a good initial distance bound d has to be specified; and third, the algorithm has to be restarted. In the next section, we choose α and τ differently between iterations, in order to extend the above analysis to allow Q to be constrained, as well as overcome the mentioned caveats.

4 Final Accelerated Gradient Method with Variable Step Lengths

In this section, we recover the main result of [Nes05] in the unconstrained case, that is

[16] We remark here that this is essentially the way to choose α in mirror descent, see (2.3).

[17] One may find this counter-intuitive because when it is closer to the optimum, the observed gradients will become smaller, and therefore mirror steps should perform well due to our conceptual message in the introduction. This understanding is incorrect for two reasons. First, when it is closer to the optimum, the threshold between 'large'

Algorithm 1 $\texttt{AGM}(f, w, x_0, T)$

Input: f a differentiable and convex function on Q that is L-smooth with respect to $\|\cdot\|$;
w the DGF function that is 1-strongly convex with respect to the same $\|\cdot\|$ over Q;
x_0 some initial point; and T the number of iterations.

Output: y_T such that $f(y_T) - f(x^*) \leq \frac{4\Theta L}{T^2}$.

$V_x(y) \stackrel{\text{def}}{=} w(y) - \langle \nabla w(x), y - x\rangle - w(x)$.

$y_0 \leftarrow x_0, \quad z_0 \leftarrow x_0$.

for $k \leftarrow 0$ **to** $T-1$ **do**

  $\alpha_{k+1} \leftarrow \frac{k+2}{2L}$, and $\tau_k \leftarrow \frac{1}{\alpha_{k+1}L} = \frac{2}{k+2}$.

  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$.

  $y_{k+1} \leftarrow \mathsf{Grad}(x_{k+1})$ ⋄ $= \arg\min_{y\in Q}\left\{\frac{L}{2}\|y - x_{k+1}\|^2 + \langle \nabla f(x_{k+1}), y - x_{k+1}\rangle\right\}$

  $z_{k+1} \leftarrow \mathsf{Mirr}_{z_k}\big(\alpha_{k+1}\nabla f(x_{k+1})\big)$ ⋄ $= \arg\min_{z\in Q}\left\{V_{z_k}(z) + \langle \alpha_{k+1}\nabla f(x_{k+1}), z - z_k\rangle\right\}$

end for

return y_T.

Theorem 4.1. *If $f(x)$ is L-smooth with respect to $\|\cdot\|$ on Q, and $w(x)$ is 1-strongly convex with respect to the same $\|\cdot\|$ on Q, the algorithm $\texttt{AGM}(f, w, x_0, T)$ in Algorithm 1 ensures*

$$f\big(y_T\big) - f(x^*) \leq \frac{4\Theta L}{T^2} \quad .$$

Here, recall from Section 2.2 that Θ is any upper bound on $V_{x_0}(x^*)$.

We remark here that it is very important to allow the norm $\|\cdot\|$ to be general, rather than focusing on the ℓ_2-norm as in [Nes04]. See our discussion in Appendix A.1.

This time, we start from $x_0 = y_0 = z_0$, and in each step $k = 0, 1, \dots, T-1$, we first compute $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$ and then (as illustrated in Algorithm 1)

- perform a gradient step $y_{k+1} \leftarrow \mathsf{Grad}(x_{k+1})$, and
- perform a mirror step $z_{k+1} \leftarrow \mathsf{Mirr}_{z_k}\big(\alpha_{k+1}\nabla f(x_{k+1})\big)$.

Here, α_{k+1} is the step length of the mirror descent and its choice will become clear at the end of this section (and indeed increasing as time goes, similar to the warm-up case). The value of τ_k is chosen as $\frac{1}{\alpha_{k+1}L}$ comparing to $\frac{1}{\alpha_{k+1}L+1}$ in the warm-up case, in order to capture the constrained case $Q \neq \mathbb{R}^n$. Our eventual choice of α_{k+1} will ensure that $\tau_k \in (0,1]$ for each k.

We state the counterpart of Lemma 3.1, whose proof can be found in Appendix C:

Lemma 4.2. *If $\tau_k = \frac{1}{\alpha_{k+1}L}$, then it satisfies that for every $u \in Q$,*

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle &\overset{①}{\leq} \alpha_{k+1}^2 L\mathsf{Prog}(x_{k+1}) + V_{z_k}(u) - V_{z_{k+1}}(u) \\ &\overset{②}{\leq} \alpha_{k+1}^2 L\big(f(x_{k+1}) - f(y_{k+1})\big) + V_{z_k}(u) - V_{z_{k+1}}(u) \quad .\end{aligned}$$

We state the counterpart of Lemma 3.2, whose proof is only slightly different from Lemma 3.2 because we are using $\tau_k = \frac{1}{\alpha_{k+1}L}$ rather than $\tau = \frac{1}{\alpha L+1}$, and can be found in Appendix C:

Lemma 4.3 (Coupling)**.** *For any $u \in Q$,*

$$\big(\alpha_{k+1}^2 L\big) f(y_{k+1}) - \big(\alpha_{k+1}^2 L - \alpha_{k+1}\big) f(y_k) + \big(V_{z_{k+1}}(u) - V_{z_k}(u)\big) \leq \alpha_{k+1} f(u) \quad .$$

and 'small' gradients also become smaller, so one cannot rely only on mirror steps. Second, when it is closer to the optimum, mirror steps are more 'unstable' and may increase the objective more (in comparison to the current distance to the optimum), and thus should be given less weight.

Finally, we only need to set the sequence of α_k so that $\alpha_k^2 L \approx \alpha_{k+1}^2 L - \alpha_{k+1}$ as well as $\tau_k = 1/\alpha_{k+1}L \in (0,1]$. For instance, we can let $\alpha_k = \frac{k+1}{2L}$ so that $\alpha_k^2 L = \alpha_{k+1}^2 L - \alpha_{k+1} + \frac{1}{4L}$.

Proof of Theorem 4.1. After telescoping Lemma 4.3 with $k = 0, 1, \dots, T-1$ we obtain that

$$\alpha_T^2 L f(y_T) + \sum_{k=1}^{T-1} \frac{1}{4L} f(y_k) + \big(V_{z_T}(u) - V_{z_0}(u)\big) \le \sum_{k=1}^{T} \alpha_k f(u) \ .$$

By choosing $u = x^*$, we notice that $\sum_{k=1}^T \alpha_k = \frac{T(T+3)}{4L}$, $f(y_k) \ge f(x^*)$, $V_{z_T}(u) \ge 0$ and $V_{z_0}(x^*) \le \Theta$. Therefore, we obtain

$$\frac{(T+1)^2}{4L^2} L f(y_T) \le \Big(\frac{T(T+3)}{4L} - \frac{T-1}{4L}\Big) f(x^*) + \Theta \ ,$$

which after simplification implies $f(y_T) \le f(x^*) + \frac{4\Theta L}{(T+1)^2}$. □

Let us make two remarks.

- First, our accelerated method AGM is almost the same to that of Nesterov [Nes05], with the following (minor) differences: (1) we use mirror steps instead of dual averaging steps,[18] (2) we allow arbitrary starting points x_0, and (3) we use $\tau_k = \frac{2}{k+2}$ rather than $\tau_k = \frac{2}{k+3}$.
- This method is very different from the (perhaps better-known) version of Nesterov [Nes04], which is only applicable to the ℓ_2 Euclidean case, and is known by some authors as the 'momentum analysis' or 'momentum method' [OC13, SBC14]. To the best of our knowledge, the momentum analysis does not apply to non-Euclidean spaces.

5 Strong Convexity Version of Accelerated Gradient Method

When the objective $f(\cdot)$ is both σ-strongly convex and L-smooth with respect to the same norm $\|\cdot\|_2$, another version of accelerated gradient method exists and achieves a $\log(1/\varepsilon)$ convergence rate [Nes04, Theorem 2.2.2]. We show in this section that, our method $\texttt{AGM}(f, w, x_0, T)$ can be used to recover that strong-convexity accelerated method in one of the two ways. Therefore, the gradient-mirror coupling interpretation behind our paper still applies to the strong-convexity accelerated method.

One way to recover the strong-convexity accelerated method is to replace the use of the mirror-descent analysis on the regret term by its strong-convexity counterpart (also known as logarithmic-regret analysis, see for instance [HAK07, SS07]). This would incur some different parameter choices on α_k and τ_k, and results in an algorithm similar to that of [Nes04].

Another, but simpler way is to recursively apply Theorem 4.1. In light of the definition of strong convexity and Theorem 4.1, we have

$$\frac{\sigma}{2}\|y_T - x^*\|_2^2 \le f(y_T) - f(x^*) \le \frac{4 \cdot \frac{1}{2}\|x_0 - x^*\|_2^2 \cdot L}{T^2} \ .$$

In particular, in every $T = T_0 \stackrel{\text{def}}{=} \sqrt{8L/\sigma}$ iterations, we can halve the distance $\|y_T - x^*\|_2^2 \le \frac{1}{2}\|x_0 - x^*\|_2^2$. If we repeatedly invoke $\texttt{AGM}(f, w, \cdot, T_0)$ a sequence of ℓ times, each time feeding the initial vector x_0 with the previous output y_{T_0}, then in the last run of the T_0 iterations, we have

$$f(y_{T_0}) - f(x^*) \le \frac{4 \cdot \frac{1}{2^\ell}\|x_0 - x^*\|_2^2 \cdot L}{T_0^2} = \frac{1}{2^{\ell+1}}\|x_0 - x^*\|_2^2 \cdot \sigma \ .$$

By choosing $\ell = \log\big(\frac{\|x_0 - x^*\|_2^2 \cdot \sigma}{\varepsilon}\big)$, we conclude that

Corollary 5.1. *If $f(\cdot)$ is both σ-strongly convex and L-smooth with respect to $\|\cdot\|_2$, in a total of $T = O\big(\sqrt{\frac{L}{\sigma}} \cdot \log\big(\frac{\|x_0 - x^*\|_2^2 \cdot \sigma}{\varepsilon}\big)\big)$ iterations, we can obtain some x such that $f(x) - f(x^*) \le \varepsilon$.*

[18] We are unaware of the existence of this mirror-descent version of Nesterov's accelerated method recorded anywhere.

This is slightly better than the result $O\left(\sqrt{\frac{L}{\sigma}} \cdot \log\left(\frac{\|x_0 - x^*\|_2^2 \cdot L}{\varepsilon}\right)\right)$ in [Nes04, Theorem 2.2.2].

We remark here that O'Donoghue and Candès [OC13] have studied some heuristic adaptive restarting techniques which suggest that the above (and other) restarting version of the accelerated method practically outperforms the original method of Nesterov.

Acknowledgements

We thank Jon Kelner and Yin Tat Lee for helpful conversations, and Aaditya Ramdas for pointing out a typo in the previous version of this paper.

This material is based upon work partly supported by the National Science Foundation under Grant CCF-1319460 and by a Simons Graduate Student Award under grant no. 284059.

Appendix

A Several Remarks on First-Order Methods

A.1 Importance of Non-Euclidean Norms

Let us use a simple example to illustrate the importance of allowing arbitrary norms in studying first-order methods.

Consider the saddle point problem of $\min_{x\in\Delta_n}\max_{y\in\Delta_m} y^T Ax$, where A is an $m\times n$ matrix, $\Delta_n = \{x\in\mathbb{R}^n : x\ge 0 \wedge \mathbb{1}^T x = 1\}$ is the unit simplex in $\mathbb{R}^n$, and $\Delta_m = \{y\in\mathbb{R}^m : y\ge 0\wedge \mathbb{1}^T y = 1\}$. This problem is important to study because it captures packing and covering linear programs that have wide applications in many areas of computer science, see the discussion in [AO15].

Letting $\mu = \frac{\varepsilon}{2\log m}$, Nesterov [Nes05] has shown that the following objective

$$f_\mu(x) \stackrel{\mathrm{def}}{=} \mu \log\big(\frac{1}{m}\sum_{j=1}^m \exp^{\frac{1}{\mu}(Ax)_j}\big) \enspace ,$$

when optimized over $x\in\Delta_n$, can yield an additive $\varepsilon/2$ solution to the original saddle point problem.

This $f_\mu(x)$ is proven to be $\frac{1}{\mu}$-smooth with respect to the ℓ_1-norm over Δ_n, if all the entries of A are between $[-1,1]$. Instead, $f_\mu(x)$ is $\frac{1}{\mu}$-smooth with respect to the ℓ_2-norm over Δ_n, *only if* the sum of squares of every row of A is at most 1. This ℓ_2 condition is certainly stronger and less natural than the ℓ_1 condition, and the ℓ_1 condition one leads to the fastest (approximate) width-dependent positive LP solver (see the discussion in [AO15]).

Different norm conditions also yield different gradient and mirror descent steps. For instance, in the ℓ_1-norm case, the gradient step is $x' \leftarrow \arg\min_{x'\in\Delta_n}\big\{\frac{1}{2}\|x'-x\|_1^2 + \alpha\langle\nabla f_\mu(x), x'-x\rangle\big\}$, and the mirror step is $x'\leftarrow \arg\min_{x'\in\Delta_n}\big\{\sum_{i\in[n]} x'_i \log\frac{x'_i}{x_i} + \alpha\langle\nabla f_\mu(x), x'-x\rangle\big\}$. In the ℓ_2-norm case, gradient and mirror steps are both of the form $x'\leftarrow \arg\min_{x'\in\Delta_n}\big\{\frac{1}{2}\|x'-x\|_2^2 + \alpha\langle\nabla f_\mu(x), x'-x\rangle\big\}$.

One can find other applications as well in [Nes05] for the use of non-Euclidean norms, and an interesting example of ℓ_∞-norm gradient descent for nearly-linear time maximum flow in [KLOS14].

It is now important to note that, the methods in [Nes83, Nes04] work only for the ℓ_2-norm case, and it is not clear how the proof can be generalized to other norms until [Nes05]. Some other proofs (such as Fercoq and Richtárik [FR13]) only work for the ℓ_2-norm because the mirror steps are described as (a scaled version of) gradient steps.

A.2 Multiplicative Weight Updates as Mirror Descent

The multiplicative weight update (MWU) method (see the survey of Arora, Hazan and Kale [AHK12]) is a simple method that has been repeatedly discovered in theory of computation, machine learning, optimization, and game theory. The setting of this method is the following.

Let $\Delta_n = \{x\in\mathbb{R}^n : x\ge 0\wedge \mathbb{1}^T x = 1\}$ be the unit simplex in $\mathbb{R}^n$, and we call any vector in Δ_n an *action*. A player is going to play T actions $x_0,\dots,x_{T-1}\in\Delta_n$ in a row; only after playing x_k, the player observes a loss vector $\ell_k\in\mathbb{R}^n$ that may depend on x_k, and suffers from a loss value $\langle \ell_k, x_k\rangle$. The MWU method ensures that, if $\|\ell_k\|_\infty \le \rho$ for all $k\in[T]$, then the player has an (adaptive) strategy to choose the actions such that the average *regret* is bounded:

$$\frac{1}{T}\Big(\sum_{i=0}^{T-1}\langle \ell_k, x_k\rangle - \min_{u\in\Delta_n}\sum_{i=0}^{T-1}\langle\ell_k, u\rangle\Big) \le O\Big(\frac{\rho\sqrt{\log n}}{\sqrt{T}}\Big) \enspace . \tag{A.1}$$

The left hand side is called the average regret because it is the (average) difference between the suffered loss $\sum_{i=0}^{T-1}\langle\ell_k, x_k\rangle$, and the loss $\sum_{i=0}^{T-1}\langle\ell_k,u\rangle$ of the best action $u\in\Delta_n$ in hindsight. Another way to interpret (A.1) is to state that we can obtain an average regret of ε using $T = O(\frac{\rho^2 \log n}{\varepsilon^2})$ rounds.

The above result can be proven directly using mirror descent. Letting $w(x) \stackrel{\text{def}}{=} \sum_i x_i \log x_i$ be the entropy DGF over the simplex $Q = \Delta_n$, and its corresponding Bregman divergence $V_x(x') \stackrel{\text{def}}{=} \sum_{i\in[n]} x'_i \log \frac{x'_i}{x_i}$, we consider the following update rule.

Start from $x_0 = (1/n, \dots, 1/n)$, and update $x_{k+1} = \mathsf{Mirr}_{x_k}\big(\alpha \ell_k\big)$, or equivalently, $x_{k+1,i} = x_{k,i} \cdot \exp^{-\alpha \ell_{k,i}} / Z_k$, where $Z_k > 0$ is the normalization factor that equals to $\sum_{i=1}^n x_{k,i} \cdot \exp^{-\alpha \ell_{k,i}}$.[19] Then, the mirror-descent guarantee (2.2) implies that[20]

$$\forall u \in \Delta_n, \quad \alpha \langle \ell_k, x_k - u \rangle \le \frac{\alpha^2}{2} \|\ell_k\|_\infty^2 + V_{x_k}(u) - V_{x_{k+1}}(u) \enspace .$$

After telescoping the above inequality for all $k = 0, 1, \dots, T-1$, and using the upper bounds $\|\ell(x_k)\|_\infty \le \rho$ and $V_{x_0}(u) \le \log n$, we obtain that for all $u \in \Delta_n$,

$$\frac{1}{T} \sum_{k=0}^{T-1} \langle \ell_k, x_k - u \rangle \le \frac{\alpha \rho^2}{2} + \frac{\log n}{\alpha T} \enspace .$$

Setting $\alpha = \frac{\sqrt{\log n}}{\rho \sqrt{T}}$ we arrive at the desired average regret bound (A.1).

In sum, we have re-deduced the MWU method from mirror descent, and the above proof is quite different from most of the classical analysis of MWU (e.g., [PST95, FS95, AHK05, AHK12]). It can be generalized to solve the matrix version of MWU [OSV12, AHK12], as well as to incorporate the width-reduction technique [PST95, AHK12]. We ignore such extensions here because they are outside the scope of this paper.

A.3 Partial Equivalence Between Mirror Descent and Dual Averaging

In this section, we show the (folklore) equivalence between mirror descent and dual averaging in two special cases: i) when $Q = \mathbb{R}^{\ltimes}$ and w is a general regularizer, and ii) when $Q = \{x \ge 0 \,:\, \mathbb{1}^T x = 1\}$ is the n-dimensional simplex and w is the entropy regularizer. In fact, this equivalence holds more generally for all regularizers $w(\cdot)$ that are convex function of Legendre type with domain Q (see for instance [BMDG05, Roc96]).

Letting $\xi_i = \alpha_i \nabla f(x_i)$ be the observed (scaled) gradient at step i, the dual averaging method can be described as

$$\forall k \in [T], \quad x_k = \operatorname*{arg\,min}_{y \in Q} \Big\{ w(y) + \sum_{i=0}^{k-1} \langle \xi_i, y - x_i \rangle \Big\} \enspace . \tag{A.2}$$

The mirror descent method (with starting point $\widetilde{x}_0 = \arg\min_{y \in Q}\{w(y)\}$) can be described as

$$\forall k \in [T], \quad \widetilde{x}_k = \operatorname*{arg\,min}_{y \in Q} \Big\{ V_{\widetilde{x}_{k-1}}(y) + \langle \xi_{k-1}, y - \widetilde{x}_{k-1} \rangle \Big\} \enspace , \tag{A.3}$$

where as before, $V_x(y) \stackrel{\text{def}}{=} w(y) - \langle \nabla w(x), y - x \rangle - w(x)$ is the Bregman divergence of $w(\cdot)$.

Unconstrained Case. If $Q = \mathbb{R}^n$, by taking the derivative from (A.2), we obtain that $\nabla w(x_k) = -\sum_{i=0}^{k-1} \xi_i$. On the other hand, by taking the derivative from (A.3), we obtain that

$$\nabla V_{\widetilde{x}_{k-1}}(\widetilde{x}_k) = -\xi_{k-1} \Longleftrightarrow \nabla w(\widetilde{x}_k) - \nabla w(\widetilde{x}_{k-1}) = -\xi_{k-1} \enspace .$$

[19] This version of the MWU is often known as the Hedge rule [FS95]. Another commonly used version is to choose $x_{k+1,i} = \frac{x_{k,i}(1 - \alpha \ell_{k,i})}{Z_k}$. Since $e^{-t} \approx 1 - t$ whenever $|t|$ is small and our choice of α will make sure that $|\alpha \ell_{k,i}| \ll 1$, this is essentially identical to the Hedge rule.

[20] To be precise, we have replaced $\partial f(x_k)$ with ℓ_k. It is easy to see from the proof of (2.2) that this loss vector ℓ_k does not need to come from the subgradient of some objective $f(\cdot)$.

Combining this with the fact that $\nabla w(\widetilde{x}_0) = 0$, we conclude that $\nabla w(\widetilde{x}_k) = -\sum_{i=0}^{k-1} \xi_i$. This finishes the proof of $\widetilde{x}_k = x_k$ in the unconstrained $Q = \mathbb{R}^n$ case, because the solution x to $\nabla w(x) = -\sum_{i=0}^{k-1} \xi_i$ must be unique for a strongly convex function $w(\cdot)$.

Simplex Case. If $Q = \{x \geq 0 : \mathbb{1}^T x = 1\}$ is the simplex, $\|\cdot\| = \|\cdot\|_1$ is the ℓ_1-norm, $w(x) = \sum_i x_i \log x_i$ is the entropy regularizer, we can precisely compute according to (A.2) and (A.3) that for every iteration k and coordinate $j \in [n]$,

$$x_{k,j} = \frac{\exp^{-\sum_{i=0}^{k-1} \ell_{i,j}}}{Z_k} \quad \text{and} \quad \widetilde{x}_{k,j} = \frac{\widetilde{x}_{k-1,j} \cdot \exp^{-\ell_{k,j}}}{\widetilde{Z}_k} \ ,$$

where Z_k and $\widetilde{Z}_k$ are normalization constants that ensure $\mathbb{1}^T x_k = \mathbb{1}^T \widetilde{x}_k = 1$. It is a simple exercise to verify that $x_k = \widetilde{x}_k$ for every k.

A.4 Deducing the Mirror-Descent Guarantee via Gradient Descent

In this section, we re-deduce the convergence rate of mirror descent from gradient descent. In particular, we show that the dual averaging steps are equivalent to gradient steps on the Fenchel dual of the regularized regret, and deduce the same convergence bound as (2.4). (Similar proof can also be obtained for mirror steps but is notationally more involved.)

Given a sequence of points $x_0, \dots, x_{T-1} \in Q$, the (scaled) regret with respect to any point $u \in Q$ is $R(x_0, \dots, x_{T-1}, u) \stackrel{\text{def}}{=} \sum_{i=0}^{T-1} \alpha \langle \partial f(x_i), x_i - u \rangle$. Since it satisfies that $\alpha T \cdot (f(\overline{x}) - f(u)) \leq R(x_0, \dots, x_{T-1}, u)$, the average regret (after scaling) upper bounds on the distance between any point $f(u)$ and the average $\overline{x} = \frac{1}{T}(x_0 + \cdots + x_{T-1})$. Consider now the regularized regret

$$\widehat{R}(x_0, \dots, x_{T-1}) \stackrel{\text{def}}{=} \max_{u \in Q} \Big\{ \sum_{i=0}^{T-1} \alpha \langle \partial f(x_i), x_i - u \rangle - w(u) \Big\} \ ,$$

and we can rewrite it using the Fenchel dual $w^*(\lambda) \stackrel{\text{def}}{=} \max_{u \in Q}\{\langle \lambda, u \rangle - w(u)\}$ of $w(\cdot)$:

$$\widehat{R}(x_0, \dots, x_{T-1}) = w^*\Big(-\alpha \sum_{i=0}^{T-1} \partial f(x_i) \Big) + \sum_{i=0}^{T-1} \alpha \langle \partial f(x_i), x_i \rangle \ .$$

The classical theory of Fenchel duality tells us that $w^*(\lambda)$ is 1-smooth with respect to the dual norm $\|\cdot\|_*$, because $w(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|$. We also have $\nabla w^*(\lambda) = \arg\max_{u \in Q}\{\langle \lambda, u \rangle - w(u)\}$. (See for instance [Sha11].)

With enough notations introduced, let us now minimize $\widehat{R}$ by intelligently selecting $x_0, \dots, x_{T-1}$. Perhaps a little counter-intuitively, we start from $x_0 = \cdots = x_{T-1} = x^*$ and accordingly $\partial f(x^*) = 0$ (if there are multiple subgradients at x^*, choose the zero one). This corresponds to a regret value of zero and a regularized regret $\widehat{R}(x^*, \dots, x^*) = w^*(0) = -\min_{u \in Q}\{w(u)\}$.

Next, we choose the values of $x_0, \dots, x_{T-1}$ one by one. We choose $x_0 = \arg\min_{u \in Q}\{w(u)\}$ as the starting point.[21] Suppose that the values of $x_0, \dots, x_{k-1}$ are already determined, and we are ready to pick $x_k \in Q$. Let us compute the changes in the regularized regret as a function of x_k:

$$\begin{aligned}
\Delta \widehat{R} &= \widehat{R}(x_0, \dots, x_k, x^*, \dots, x^*) - \widehat{R}(x_0, \dots, x_{k-1}, x^*, \dots, x^*) \\
&= w^*\Big(-\alpha \sum_{i=0}^{k} \partial f(x_i) \Big) - w^*\Big(-\alpha \sum_{i=0}^{k-1} \partial f(x_i) \Big) + \alpha \langle \partial f(x_k), x_k \rangle \\
&\leq \Big\langle \nabla w^*\Big(-\alpha \sum_{i=0}^{k-1} \partial f(x_i) \Big), -\alpha \partial f(x_k) \Big\rangle + \frac{1}{2} \big\| \alpha \partial f(x_k) \big\|_*^2 + \alpha \langle \partial f(x_k), x_k \rangle \ . \qquad \text{(A.4)}
\end{aligned}$$

[21] Dual averaging steps typically demand the first point x_0 to be at the minimum of the regularizer $w(\cdot)$, because that leads to the cleanest analysis. This can be relaxed to allow an arbitrary starting point.

Here, the last inequality is because $w^*(a) - w^*(b) \le \langle \nabla w^*(b), a - b\rangle + \frac{1}{2}\|a-b\|_*^2$, owing to the smoothness of $w^*(\cdot)$. At this moment, it is clear to see that if one chooses

$$x_k = \nabla w^*\Big(-\alpha \sum_{i=0}^{k-1} \partial f(x_i) \Big) = \arg\min_{u \in Q} \Big\{ w(u) + \sum_{i=0}^{k-1} \alpha\langle \partial f(x_i), u\rangle \Big\} \enspace,$$

the first and third terms in (A.4) cancel out, and we obtain $\Delta \widehat{R} \le \frac{1}{2}\big\|\alpha \partial f(x_k)\big\|_*^2$. [22] In other words, the regularized regret increases by no more than $\frac{1}{2}\big\|\alpha\partial f(x_k)\big\|_*^2 \le \alpha^2\rho^2/2$ in each step, so in the end we have $\widehat{R}(x_0,\dots,x_{T-1}) \le -w(x_0) + \alpha^2\rho^2 T/2$.

In sum, by the definition of the regularized regret, we have

$$\alpha T \cdot (f(\overline{x}) - f(x^*)) - w(x^*) \le \sum_{i=0}^{T-1} \alpha\langle \partial f(x_i), x_i - x^*\rangle - w(x^*) \le \widehat{R}(x_0,\dots,x_{T-1}) \le -w(x_0) + \frac{\alpha^2\rho^2 T}{2} \enspace.$$

This implies the following upper bound on the optimality of $f(\overline{x})$

$$f(\overline{x}) - f(x^*) \le \frac{\alpha\rho^2}{2} + \frac{w(x^*) - w(x_0)}{\alpha T} = \frac{\alpha\rho^2}{2} + \frac{V_{x_0}(x^*)}{\alpha T} \le \frac{\alpha\rho^2}{2} + \frac{\Theta}{\alpha T} \enspace.$$

Finally, choosing $\alpha = \frac{\sqrt{2\Theta}}{\rho\cdot\sqrt{T}}$ to be the step length, we arrive at $f(\overline{x}) - f(x^*) \le \frac{\sqrt{2\Theta}\cdot\rho}{\sqrt{T}}$, which is the same convergence rate as (2.4).

B Missing Proof of Section 2

For the sake of completeness, we provide self-contained proofs of the mirror descent and mirror descent guarantees in this section.

B.1 Missing Proof for Gradient Descent

Gradient Descent Guarantee

$$f(\mathsf{Grad}(x)) \le f(x) - \mathsf{Prog}(x) \tag{2.1}$$

or in the special case when $Q = \mathbb{R}^n$ $\quad f(\mathsf{Grad}(x)) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_*^2$.

Proof. [23] Letting $\widetilde{x} = \mathsf{Grad}(x)$, we prove the first inequality by

$$\begin{aligned}\mathsf{Prog}(x) &= -\min_{y\in Q}\Big\{ \frac{L}{2}\|y-x\|^2 + \langle \nabla f(x), y - x\rangle \Big\} = -\Big(\frac{L}{2}\|\widetilde{x}-x\|^2 + \langle \nabla f(x), \widetilde{x}-x\rangle\Big) \\ &= f(x) - \Big(\frac{L}{2}\|\widetilde{x}-x\|^2 + \langle \nabla f(x), \widetilde{x}-x\rangle + f(x)\Big) \le f(x) - f(\widetilde{x}) \enspace.\end{aligned}$$

Here, the last inequality is a consequence of the smoothness assumption: for any $x, y \in Q$,

$$\begin{aligned} f(y) - f(x) &= \int_{\tau=0}^1 \langle \nabla f(x + \tau(y-x)), y - x\rangle d\tau \\ &= \langle \nabla f(x), y - x\rangle + \int_{\tau=0}^1 \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x\rangle d\tau \\ &\le \langle \nabla f(x), y - x\rangle + \int_{\tau=0}^1 \|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_* \cdot \|y-x\| d\tau \\ &\le \langle \nabla f(x), y - x\rangle + \int_{\tau=0}^1 \tau L\|y-x\|\cdot\|y-x\| d\tau = \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2 \end{aligned}$$

[22] This essentially proves (1.5) in the introduction after scaling: $\Delta\widehat{R} = \alpha(k+1)\max_u \widetilde{R}_{k+1}(u) - \alpha k \max_u \widetilde{R}_k(u)$.

[23] This proof can be found for instance in the textbook [Nes04].

The second inequality follows because in the special case of $Q = \mathbb{R}^n$, we have

$$\mathsf{Prog}(x) = -\min_{y \in Q} \Big\{ \frac{L}{2}\|y - x\|^2 + \langle \nabla f(x), y - x\rangle \Big\} = \frac{1}{2L}\|\nabla f(x)\|_*^2 \ . \qquad \square$$

Fact B.1 (Gradient Descent Convergence). *Let $f(x)$ be a convex, differentiable function that is L-smooth with respect to $\|\cdot\|$ on $Q = \mathbb{R}^n$, and x_0 any initial point in Q. Consider the sequence of T gradient steps $x_{k+1} \leftarrow \mathsf{Grad}(x_k)$, then the last point x_T satisfies that*

$$f(x_T) - f(x^*) \le O\big(\frac{LR^2}{T}\big) \ ,$$

where $R = \max_{x : f(x) \le f(x_0)} \|x - x^*\|$, and x^* is any minimizer of f.

Proof. [24] Recall that we have $f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_*^2$ from (2.1). Furthermore, by the convexity of f and Cauchy-Schwarz we have

$$f(x_k) - f(x^*) \le \langle \nabla f(x_k), x_k - x^* \rangle \le \|\nabla f(x_k)\|_* \cdot \|x_k - x^*\| \le R \cdot \|\nabla f(x_k)\|_* \ .$$

Letting $D_k = f(x_k) - f(x^*)$ denote the distance to the optimum at iteration k, we now obtain two relationships $D_k - D_{k+1} \ge \frac{1}{2L}\|\nabla f(x_k)\|_*^2$ as well as $D_k \le R \cdot \|\nabla f(x_k)\|_*$. Combining these two, we get

$$D_k^2 \le 2LR^2(D_k - D_{k+1}) \implies \frac{D_k}{D_{k+1}} \le 2LR^2\Big(\frac{1}{D_{k+1}} - \frac{1}{D_k}\Big) \ .$$

Noticing that $D_k \ge D_{k+1}$ because our objective only decreases at every round, we obtain that $\frac{1}{D_{k+1}} - \frac{1}{D_k} \ge \frac{1}{2LR^2}$. Finally, we conclude that at round T, we must have $\frac{1}{D_T} \ge \frac{T}{2LR^2}$, finishing the proof that $f(x_T) - f(x^*) \le \frac{2LR^2}{T}$. $\square$

B.2 Missing Proof for Mirror Descent

Mirror Descent Guarantee

If $x_{k+1} = \mathsf{Mirr}_{x_k}\big(\alpha \cdot \partial f(x_k)\big)$, then

$$\forall u \in Q, \quad \alpha(f(x_k) - f(u)) \le \alpha\langle \partial f(x_k), x_k - u\rangle \le \frac{\alpha^2}{2}\|\partial f(x_k)\|_*^2 + V_{x_k}(u) - V_{x_{k+1}}(u) \ . \qquad (2.2)$$

Proof. [25] we compute that

$$\begin{aligned}
\alpha\langle \partial f(x_k), x_k - u\rangle &= \langle \alpha\partial f(x_k), x_k - x_{k+1}\rangle + \langle \alpha\partial f(x_k), x_{k+1} - u\rangle \\
&\overset{①}{\le} \langle \alpha\partial f(x_k), x_k - x_{k+1}\rangle + \langle -\nabla V_{x_k}(x_{k+1}), x_{k+1} - u\rangle \\
&\overset{②}{=} \langle \alpha\partial f(x_k), x_k - x_{k+1}\rangle + V_{x_k}(u) - V_{x_{k+1}}(u) - V_{x_k}(x_{k+1}) \\
&\overset{③}{\le} \Big(\langle \alpha\partial f(x_k), x_k - x_{k+1}\rangle - \frac{1}{2}\|x_k - x_{k+1}\|^2\Big) + \big(V_{x_k}(u) - V_{x_{k+1}}(u)\big) \\
&\overset{④}{\le} \frac{\alpha^2}{2}\|\partial f(x_k)\|_*^2 + \big(V_{x_k}(u) - V_{x_{k+1}}(u)\big)
\end{aligned}$$

Here, ① is due to the minimality of $x_{k+1} = \arg\min_{x \in Q}\{V_{x_k}(x) + \langle \alpha\partial f(x_k), x\rangle\}$, which implies that $\langle \nabla V_{x_k}(x_{k+1}) + \alpha\partial f(x_k), u - x_{k+1}\rangle \ge 0$ for all $u \in Q$. ② is due to the triangle equality of Bregman

[24] Our proof follows almost directly from Nesterov [Nes04], but he only uses the Euclidean ℓ_2 norm.

[25] This proof can be found for instance in the textbook [BN13].

divergence.[26] ③ is because $V_x(y) \ge \frac{1}{2}\|x-y\|^2$ by the strongly convex of the DGF $w(\cdot)$. ④ is by Cauchy-Schwarz. □

C Missing Proofs of Section 4

Lemma 4.2. *If $\tau_k = \frac{1}{\alpha_{k+1}L}$, then it satisfies that for every $u \in Q$,*

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle &\overset{①}{\le} \alpha_{k+1}^2 L\mathsf{Prog}(x_{k+1}) + V_{z_k}(u) - V_{z_{k+1}}(u) \\ &\overset{②}{\le} \alpha_{k+1}^2 L\big(f(x_{k+1}) - f(y_{k+1})\big) + V_{z_k}(u) - V_{z_{k+1}}(u) \ .\end{aligned}$$

Proof. The second inequality ② is again from the gradient descent guarantee $f(x_{k+1}) - f(y_{k+1}) \ge \mathsf{Prog}(x_{k+1})$. To prove ①, we first write down the key inequality of mirror-descent analysis (whose proof is identical to that of (2.2))

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle &= \langle \alpha_{k+1}\nabla f(x_{k+1}), z_k - z_{k+1}\rangle + \langle \alpha_{k+1}\nabla f(x_{k+1}), z_{k+1} - u\rangle \\ &\overset{①}{\le} \langle \alpha_{k+1}\nabla f(x_{k+1}), z_k - z_{k+1}\rangle + \langle -\nabla V_{z_k}(z_{k+1}), z_{k+1} - u\rangle \\ &\overset{②}{=} \langle \alpha_{k+1}\nabla f(x_{k+1}), z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1}) \\ &\overset{③}{\le} \Big(\langle \alpha_{k+1}\nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}\|z_k - z_{k+1}\|^2\Big) + \big(V_{z_k}(u) - V_{z_{k+1}}(u)\big)\end{aligned}$$

Here, ① is due to the minimality of $z_{k+1} = \arg\min_{z\in Q}\{V_{z_k}(z) + \langle \alpha_{k+1}\nabla f(x_{k+1}), z\rangle\}$, which implies that $\langle \nabla V_{z_k}(z_{k+1}) + \alpha_{k+1}\nabla f(x_{k+1}), u - z_{k+1}\rangle \ge 0$ for all $u \in Q$. ② is due to the triangle equality of Bregman divergence (see Footnote 26 in Appendix B). ③ is because $V_x(y) \ge \frac{1}{2}\|x-y\|^2$ by the strongly convex of the $w(\cdot)$.

If one stops here and uses Cauchy-Shwartz $\langle \alpha_{k+1}\nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}\|z_k - z_{k+1}\|^2 \le \frac{\alpha_{k+1}^2}{2}\|\nabla f(x_{k+1})\|_*^2$, he will get the desired inequality in the special case of $Q = \mathbb{R}^n$, because $\mathsf{Prog}(x_{k+1}) = \frac{1}{2L}\|\nabla f(x_{k+1})\|_*^2$ from (2.1).

For the general unconstrained case, we need to use the special choice of $\tau_k = 1/\alpha_{k+1}L$ follows. Letting $v \overset{\text{def}}{=} \tau_k z_{k+1} + (1-\tau_k)y_k \in Q$ so that $x_{k+1} - v = (\tau_k z_k + (1-\tau_k)y_k) - v = \tau_k(z_k - z_{k+1})$, we have

$$\begin{aligned}&\langle \alpha_{k+1}\nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}\|z_k - z_{k+1}\|^2 \\ &= \langle \frac{\alpha_{k+1}}{\tau_k}\nabla f(x_{k+1}), x_{k+1} - v\rangle - \frac{1}{2\tau_k^2}\|x_{k+1} - v\|^2 \\ &= \alpha_{k+1}^2 L\left(\langle \nabla f(x_{k+1}), x_{k+1} - v\rangle - \frac{L}{2}\|x_{k+1} - v\|^2\right) \le \alpha_{k+1}^2 L\mathsf{Prog}(x_{k+1})\end{aligned}$$

where the last inequality is from the definition of $\mathsf{Prog}(x_{k+1})$. □

Lemma 4.3 (Coupling)**.** *For any $u \in Q$,*

$$\big(\alpha_{k+1}^2 L\big)f(y_{k+1}) - \big(\alpha_{k+1}^2 L - \alpha_{k+1}\big)f(y_k) + \big(V_{z_{k+1}}(u) - V_{z_k}(u)\big) \le \alpha_{k+1} f(u) \ .$$

[26] That is,

$$\begin{aligned}\forall x, y \ge 0, \quad \langle -\nabla V_x(y), y - u\rangle &= \langle \nabla w(x) - \nabla w(y), y - u\rangle \\ &= (w(u) - w(x) - \langle \nabla w(x), u - x\rangle) - (w(u) - w(y) - \langle w(y), u - y\rangle)) \\ &\quad - (w(y) - w(x) - \langle \nabla w(x), y - x\rangle) \\ &= V_x(u) - V_y(u) - V_x(y) \ .\end{aligned}$$

Proof. We deduce the following sequence of inequalities

$$\begin{aligned}
&\alpha_{k+1}\big(f(x_{k+1}) - f(u)\big)\\
&\leq \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle\\
&= \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle\\
&\overset{①}{=} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle\\
&\overset{②}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(f(y_k) - f(x_{k+1})) + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle\\
&\overset{③}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(f(y_k) - f(x_{k+1})) + \alpha_{k+1}^2 L\big(f(x_{k+1}) - f(y_{k+1})\big) + V_{z_k}(u) - V_{z_{k+1}}(u)\\
&\overset{④}{=} \big(\alpha_{k+1}^2 L - \alpha_{k+1}\big)f(y_k) - \big(\alpha_{k+1}^2 L\big)f(y_{k+1}) + \alpha_{k+1}f(x_{k+1}) + \big(V_{z_k}(u) - V_{z_{k+1}}(u)\big)
\end{aligned}$$

Here, ① uses the choice of x_{k+1} that satisfies $\tau_k(x_{k+1} - z_k) = (1-\tau_k)(y_k - x_{k+1})$; ② is by the convexity of $f(\cdot)$ and $1 - \tau_k \geq 0$; ③ uses Lemma 4.2; and ④ uses the choice of $\tau_k = 1/\alpha_{k+1}L$. □

References

[AHK05] Sanjeev Arora, Elad Hazan, and Satyen Kale. Fast Algorithms for Approximate Semidefinite Programming using the Multiplicative Weights Update Method. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 339–348. IEEE, 2005.

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing*, 8:121–164, 2012.

[AO14] Zeyuan Allen-Zhu and Lorenzo Orecchia. Nearly-linear time packing and covering LP solver with faster convergence rate than $O(1/\varepsilon^2)$. *ArXiv e-prints*, abs/1411.1124, November 2014.

[AO15] Zeyuan Allen-Zhu and Lorenzo Orecchia. Using optimization to break the epsilon barrier: A faster and simpler width-independent algorithm for solving positive linear programs in parallel. In *Proceedings of the 26th ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, 2015.

[BBR97] Yair Bartal, John W. Byers, and Danny Raz. Global optimization using local information with applications to flow control. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 303–312. IEEE Comput. Soc, 1997.

[BBR04] Yair Bartal, John W. Byers, and Danny Raz. Fast, Distributed Approximation Algorithms for Positive Linear Programming with Applications to Flow Control. *SIAM Journal on Computing*, 33(6):1261–1279, January 2004.

[BMDG05] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[BN13] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization.* Society for Industrial and Applied Mathematics, January 2013.

[CKM+11] Paul Christiano, Jonathan A. Kelner, Aleksander Madry, Daniel A. Spielman, and Shang-Hua Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proceedings of the 43rd annual ACM symposium on Theory of computing - STOC '11*, page 273, New York, New York, USA, October 2011. ACM Press.

[DGSX12] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *The Journal of Machine Learning Research*, 13(1):165–202, 2012.

[DSST10] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite Objective Mirror Descent. In *Proceedings of the 23rd Annual Conference on Learning Theory - COLT '10*, number 1, 2010.

[FR13] Olivier Fercoq and Peter Richtárik. Accelerated, Parallel and Proximal Coordinate Descent. *ArXiv e-prints*, abs/1312.5799:25, December 2013.

[FS95] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[HAK07] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, August 2007.

[Jud13] Anatoli Juditsky. Convex optimization ii: Algorithms. Lecture notes, November 2013.

[KLOS14] Jonathan A. Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An Almost-Linear-Time Algorithm for Approximate Max Flow in Undirected Graphs, and its Multicommodity Generalizations. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms - SODA '14*, number 1, April 2014.

[Lan11] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, January 2011.

[LN93] Michael Luby and Noam Nisan. A parallel approximation algorithm for positive linear programming. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing - STOC '93*, pages 448–457, New York, New York, USA, 1993. ACM Press.

[LRS13] Yin Tat Lee, Satish Rao, and Nikhil Srivastava. A new approach to computing maximum flows using electrical flows. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing - STOC '13*, page 755, New York, New York, USA, 2013. ACM Press.

[Mad13] Aleksander Madry. Navigating Central Path with Electrical Flows: From Flows to Matchings, and Back. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 253–262. IEEE, October 2013.

[McM11] H. Brendan McMahan. A Unified View of Regularized Dual Averaging and Mirror Descent with Implicit Updates. *arXiv preprint arXiv:1009.3240*, September 2011. Previously appeared in AISTATS 2011 as a conference paper entitled "Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization".

[MS10] H. Brendan McMahan and Matthew Streeter. Adaptive Bound Optimization for Online Convex Optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory - COLT '10*, February 2010.

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Doklady AN SSSR (translated as Soviet Mathematics Doklady)*, volume 269, pages 543–547, 1983.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Programming Volume: A Basic course*, volume I. Kluwer Academic Publishers, 2004.

[Nes05] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, December 2005.

[Nes07] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, June 2007.

[Nes08] Yurii Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[Nes13] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Nes14] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, May 2014.

[NY78] Arkadi Nemirovsky and David Yudin. *Problem complexity and method efficiency in optimization.* Nauka Publishers, Moscow (in Russian), 1978. John Wiley, New York (in English) 1983.

[OC13] Brendan O'Donoghue and Emmanuel Candès. Adaptive Restart for Accelerated Gradient Schemes. *Foundations of Computational Mathematics*, July 2013.

[OSV12] Lorenzo Orecchia, Sushant Sachdeva, and Nisheeth K. Vishnoi. Approximating the exponential, the lanczos method and an $\widetilde{O}(m)$-time spectral algorithm for balanced separator. In *STOC '12*. ACM Press, November 2012.

[PST95] Serge A. Plotkin, David B. Shmoys, and Éva Tardos. Fast Approximation Algorithms for Fractional Packing and Covering Problems. *Mathematics of Operations Research*, 20(2):257–301, May 1995.

[Roc96] R. Tyrrell Rockafellar. *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*. Princeton University Press, 1996.

[SBC14] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[Sha11] Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[She13] Jonah Sherman. Nearly Maximum Flows in Nearly Linear Time. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 263–269. IEEE, October 2013.

[SS07] Shai Shalev-Shwartz and Yoram Singer. Logarithmic regret algorithms for strongly convex repeated games. Technical report, 2007.

[SZ13a] Shai Shalev-Shwartz and Tong Zhang. Accelerated Mini-Batch Stochastic Dual Coordinate Ascent. In *NIPS*, pages 1–17, May 2013.

[SZ13b] Shai Shalev-Shwartz and Tong Zhang. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization. *arXiv preprint arXiv:1309.2375*, pages 1–38, September 2013.

[SZ13c] Ohad Shamir and Tong Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. In *Proceedings of the 30th International Conference on Machine Learning - ICML '13*, volume 28, 2013.

[Xia10] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. *The Journal of Machine Learning Research*, 11:2543–2596, 2010.

[You01] Neal E. Young. Sequential and parallel algorithms for mixed packing and covering. In *42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS'01)*, pages 538–546. IEEE Comput. Soc, 2001.