THEORY OF COMPUTING, Volume 15 (4), 2019, pp. 1–32
www.theoryofcomputing.org

RESEARCH SURVEY

Potential-Function Proofs for Gradient Methods

Nikhil Bansal* Anupam Gupta†

Received March 3, 2019; Revised August 6, 2019; Published September 12, 2019

Abstract: This note discusses proofs of convergence for gradient methods (also called "first-order methods") based on simple potential-function arguments. We cover methods like gradient descent (for both smooth and non-smooth settings), mirror descent, and some accelerated variants. We hope the structure and presentation of these amortized-analysis proofs will be useful as a guiding principle in learning and using these proofs.

1 Introduction

The "gradient descent" framework is a class of iterative methods for solving convex minimization problems—indeed, since the gradient gives the direction of steepest increase in function value, a natural approach to minimize the convex function is to move in the direction opposite to the gradient. Variants of this general versatile approach have been central to convex optimization for many years. In recent years, with the increased use of continuous methods in discrete optimization, this technique has also become central for algorithm design in general.

*Supported in part by NWO Vici grant 639.023.812 and an ERC consolidator grant 617951.
†Supported in part by NSF awards CCF-1536002, CCF-1540541, and CCF-1617790.

ACM Classification: G.1.6

AMS Classification: 68Q25

Key words and phrases: convex optimization, potential functions, amortized analysis

 DOI: 10.4086/toc.2019.v015a004

In this note we give convergence arguments for many commonly studied versions of gradient methods (also referred to as "first-order methods") using simple *potential-function* arguments. We find that presenting the proofs in the amortized-analysis framework is useful as a guiding principle, since it imparts a clear structure and direction to proofs. We hope others will also find this perspective useful, both in learning and teaching these techniques and proofs, and also in extending them to other domains.

A disclaimer: previously existing proofs for gradient methods are usually not difficult, and their individual components are not substantially different from the ones in this note. However, using an explicit potential to guide our proofs makes them arguably more intuitive. In fact, the intuition of viewing these gradient methods as trying to control a potential function is also known to the specialists; e. g., see the text of Nemirovski and Yudin [18, pp. 85–88] for a continuous perspective via Lyapunov functions. This is more explicit in recent papers [24, 29, 17, 30, 8] relating continuous and discrete updates to understand the acceleration phenomenon, e. g., Krichene et al. [17] give the potential function we use in §5.2. However, these potential-function proofs and intuitions have not yet permeated into the commonly presented expositions. The current note is an attempt to make such ideas more widely known.

Basic definitions. Recall that a set $K \subseteq \mathbb{R}^d$ is *convex* if for all $x, y \in K$, the *convex combination* $\lambda x + (1-\lambda)y \in K$ for all $\lambda \in [0,1]$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* over a convex set K if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y \in K, \forall \lambda \in [0,1].$$

This is called the *zeroth-order* definition. There are other equivalent notions: if the function is differentiable, the *first-order* definition is that f is convex over K if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in K. \tag{1.1}$$

(The *second-order* definition says that a twice-differentiable f is convex if its Hessian matrix $\nabla^2 f$ is positive-semidefinite.) For this note, we assume our convex sets K are closed, and the convex functions f are differentiable. However, the proofs extend to non-differentiable functions in the natural way, using subgradients. (See, e. g., [13] for more definitions and background on convexity and subgradients.)

The problems. Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$, and an error parameter ε, the *(unconstrained) convex minimization* problem is to find a point $\widehat{x}$ such that $f(\widehat{x}) - \min_{x \in \mathbb{R}^d} f(x) \leq \varepsilon$. In the *constrained* version of the problem, we are also given a convex set K, and the goal is to find a point $\widehat{x} \in K$ which has *error* $f(\widehat{x}) - \min_{x \in K} f(x) \leq \varepsilon$. In either case, let x^* denote the minimizer for $f(\cdot)$. We will be interested in bounding the number of gradient queries required to converge to the approximate minimizer $\widehat{x}$, as a function of the distance between x_0 and x^* and some parameters of the function f, depending on the particular variant of gradient descent.

In *online convex optimization* over a convex set K, at each timestep $t = 1, 2, \ldots$, the algorithm outputs a point $x_t \in K$ and an adversary produces a convex function f_t. The algorithm's *loss* at timestep t is defined to be $f_t(x_t)$. Now the *regret* of the algorithm is $\sum_{t=1}^{T} f_t(x_t) - \min_{x \in K} \sum_{t=1}^{T} f_t(x)$, and the goal is to determine the points x_t online (without the knowledge of the current or future functions $\{f_s\}_{s \geq t}$) to minimize the regret. Note that this generalizes the convex optimization setting above, which corresponds to setting $f_t = f$ at each time t, and that any algorithm with sublinear regret $o(T)$ can be used to find an approximate optimum $\widehat{x}$ up to any desired accuracy ε.

Assumptions. We assume that our convex functions are closed, convex, and differentiable, and that the convex sets K are also closed with non-empty interior. We assume access to a gradient oracle for each of the functions f_t we consider, i. e., given any point x, we can get the gradient $\nabla f_t(x)$ of function f_t at point x. We only work with the Euclidean norm $\|\cdot\|_2$ for the first few sections; general norms are discussed from §4 onwards.

References. In this survey, we focus only on the exposition of the proofs. We omit most citations, and also discussion of the "bigger picture." There are many excellent sources for other proofs of these results, with comprehensive bibliographies; e. g., see the classic text by Shor [23], the authoritative notes by Nesterov [20], and Ben-Tal and Nemirovski [3], the monographs of Bubeck [6] and Shalev-Shwartz [22], the textbooks by Cesa-Bianchi and Lugosi [7], Hazan [12], and lecture notes by Duchi [10] and Vishnoi [28].

There are several other perspectives on these proofs that the reader may find useful. One useful perspective is that of viewing these methods as discretizations of suitable continuous dynamics; this appears even in the classic work of Nemirovski and Yudin [18], and has been widely used recently (see, e. g., [24, 29, 17, 30]). Another useful perspective is exhibit a "dual" lower bound on the optimal value via the convex conjugate, and use the duality gap to bound the error (see, e. g., [8, 21]). We refer the interested reader to the respective papers for more details.

Finally, we discuss some concurrent and related work. Independently of our work, Karimi and Vavasis [15] give potential-based convergence proofs for conjugate gradient and accelerated methods; their potentials are similar to ours. And following up on a preprint of our results, Taylor and Bach [26] analyze stochastic gradient methods using potential functions.

1.1 Results and organization

All of the proofs use the same general potential: for some fixed point x^* (which can be thought of as the optimal or reference point) we have

$$\Phi_t = a_t \cdot (f(x_t) - f(x^*)) + b_t \cdot (\text{distance from } x_t \text{ to } x^*)\,. \tag{1.2}$$

Here a_t, b_t are non-negative, and naturally, different proofs use slightly different choice of these multipliers, and even the distance functions may vary. However, the general approach remains the same: we show that $\Phi_{t+1} - \Phi_t \le B_t$ (where B_t is often zero). Since the potential and distance terms remain non-negative, the telescoping sum gives

$$\Phi_T \le \Phi_0 + \sum_{t=0}^{T-1} B_t \qquad \Longrightarrow \qquad f(x_T) - f(x^*) \le \frac{\Phi_0 + \sum_t B_t}{a_T}\,.$$

We begin in §2 with proofs of the basic (projected) gradient descent, for general and strongly convex functions; these even work in the online regret-minimization setting where the function may change at each timestep. Here the analysis is more along the lines of *amortized analysis*: we show that the amortized cost, namely the cost of the algorithm plus the increase in potential is at most the optimal cost (plus B). I. e., $f_t(x_t) + (\Phi_{t+1} - \Phi_t) \leq f_t(x^*) + B$. This telescopes to imply that the average regret is

$$\frac{1}{T}\left(\sum_{t=1}^{T}(f_t(x_t) - f_t(x^*))\right) \leq B + \frac{\Phi_0}{T}.$$

The potential here is very simple: we set $a_t = 0$ and just use the distance of the current point x_t to the optimal point x^* (according to the "right" distance). For example, for basic gradient descent, the potential is just a scaled version of $\|x_t - x^*\|^2$.

Next, we give proofs of convergence for the case of *smooth* convex functions in §3. In the simplest case we just set $b_t = 0$ and use $a_t = t$ in (1.2) to prove $B_t \approx 1/t$. This gives an error of $\approx (\log T)/T$, which is in the right ballpark. (This can be optimized using better settings of the multipliers.) The proofs for projected smooth gradient descent, gradient descent for well-conditioned functions, and the Frank–Wolfe method, all follow this template. For these proofs, we now use the "value-based" terms in (1.2), i. e., the terms that depend on $f(x_t) - f(x^*)$.

We then extend our understanding to *mirror descent*. This is a substantial generalization of gradient descent to general norms. While the language necessarily becomes more technical (relying on dual norms and Bregman divergences), the ideas remain clean. Indeed, the structure of the potential-based proofs from §2 remains essentially the same as for basic gradient descent; the potential is now based on a Bregman divergence, a natural generalization of the squared distance. These proofs appear in §4.

An orthogonal extension is to potential-based proofs of Nesterov's accelerated gradient descent method for smooth and well-conditioned convex functions. The ideas in this section build on the simple calculations we would have seen in §2 and §3. We can now use the full power of both distance-based and value-based terms in the potential function (1.2), trading them off against each other. Moreover, in §5.1 we show how the basic analysis for smooth convex functions from §3 directly suggests how to obtain the accelerated algorithm by coupling together one cautious and one aggressive gradient descent step.

Organization. The paper follows the above outline. We start with proofs for general convex functions in §2 using simple distance-based potentials, then proceed to smooth and well-conditioned convex functions in §3 using more sophisticated potentials. We then discuss the generalization to mirror descent via Bregman divergences in §4. Finally, we give proofs for accelerated versions in §5. Note that §4 and §5 are independent, and may be read in either order.

2 Online analyses

2.1 Basic gradient descent

The basic analysis works even for the online convex optimization case: at each step we are given a function f_t, we play x_t, and want to minimize the regret. In this case the update rule is:

$$\boxed{x_{t+1} \leftarrow x_t - \eta_t \nabla f_t(x_t)}. \tag{2.1}$$

An equivalent form for this update, that is easily verified by taking derivatives with respect to x, is:

$$\boxed{x_{t+1} \leftarrow \arg\min_x \left\{\frac{1}{2}\|x - x_t\|^2 + \eta_t \langle x, \nabla f_t(x_t)\rangle\right\}}. \tag{2.2}$$

Intuitively, we want to move in the direction of the negative gradient, but do not want to move too far.

Theorem 2.1 (Basic gradient descent)**.** *Let $f_1,\dots,f_T:\mathbb{R}^n\to\mathbb{R}$ be G-Lipschitz functions, i. e. $\|\nabla f_t(x)\|\le G$ for all x,t. Then starting at point $x_0\in\mathbb{R}^n$ and using updates (2.1) with step size*

$$\eta_t=\eta=\frac{D}{G\sqrt{T}}$$

for T steps guarantees an average regret of

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(f_t(x_t)-f_t(x^*)\big)\le\eta\frac{G^2}{2}+\frac{D^2}{2\eta T}\le\frac{DG}{\sqrt{T}},$$

for all x^* with $\|x_0-x^*\|\le D$.

Proof. Consider the potential function

$$\boxed{\Phi_t=\frac{1}{2\eta}\|x_t-x^*\|^2} \tag{2.3}$$

which is positive for all t. We show that, for some upper bound B,

$$f_t(x_t)-f_t(x^*)+\Phi_{t+1}-\Phi_t\le B. \tag{2.4}$$

Summing over all times t, the average regret is

$$\frac{1}{T}\sum_{t=0}^{T-1}(f_t(x_t)-f_t(x^*))\le B+\frac{1}{T}(\Phi_0-\Phi_T)\le B+\frac{\Phi_0}{T}=B+\frac{D^2}{2\eta T}. \tag{2.5}$$

Now we can compute B, and then balance the two terms. While the potential uses differences of the form x_t-x^*, the key is to express as much as possible in terms of $x_{t+1}-x_t$, because the update rule (2.1) implies

$$x_{t+1}-x_t=-\eta\nabla f_t(x_t). \tag{2.6}$$

The change in potential. Using that $\|a+b\|^2-\|a\|^2=2\langle a,b\rangle+\|b\|^2$ for the Euclidean norm,

$$\begin{aligned}\frac{1}{2}(\|x_{t+1}-x^*\|^2-\|x_t-x^*\|^2)&=\langle x_{t+1}-x_t,x_t-x^*\rangle+\frac{1}{2}\|x_{t+1}-x_t\|^2\\&=\eta_t\langle\nabla f_t(x_t),x^*-x_t\rangle+\frac{\eta_t^2}{2}\|\nabla f_t(x_t)\|^2.\end{aligned} \tag{2.7}$$

The amortized cost. Setting $\eta_t=\eta$ for all steps,

$$\begin{aligned}&f_t(x_t)-f_t(x^*)+\Phi_{t+1}-\Phi_t\\&\quad=f_t(x_t)-f_t(x^*)+\langle\nabla f_t(x_t),x^*-x_t\rangle+\frac{\eta}{2}\|\nabla f_t(x_t)\|^2 &&\text{(by (2.7))}\\&\quad\le 0+\frac{\eta}{2}\|\nabla f_t(x_t)\|^2\le\frac{\eta G^2}{2} &&\text{(by convexity, and the bound on gradients.)}\end{aligned}$$

Substituting for B in (2.5) and simplifying with $\eta=D/(G\sqrt{T})$, we get the theorem. □

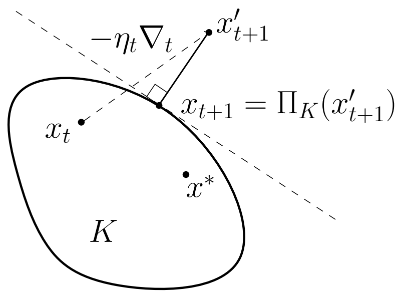

Figure 1: Projected Gradient Descent.

The regret bound implies a convergence result for the offline case, i. e., for the case where $f_t = f$ for all t. Here, setting $\widehat{x} := (1/T)\sum_{t=0}^{T-1} x_t$ shows

$$f(\widehat{x}) - f(x^*) = f\left(\frac{1}{T}\sum_t x_t\right) - f(x^*) \leq \frac{1}{T}\sum_t (f(x_t) - f(x^*)) \leq \frac{DG}{\sqrt{T}} \leq \varepsilon,$$

as long as $T \geq (DG/\varepsilon)^2$ and $\eta = \varepsilon/G^2$.

If the time horizon T is unknown, setting a time-dependent step size of $\eta_t = D/(G\sqrt{t})$ works, with an identical proof. It is also well-known that the convergence bound above is the best possible in general, modulo constant factors (see, e. g., [20, Thm 3.2.1] or [6, Thm 3.13]).

2.1.1 Projected gradient descent

If we want to solve the constrained minimization problem for a convex body K, we update as follows:

$$\boxed{\begin{aligned} x'_{t+1} &\leftarrow x_t - \eta\,\nabla f_t(x_t), \\ x_{t+1} &\leftarrow \Pi_K(x'_{t+1}). \end{aligned}} \qquad \begin{aligned}(2.8)\\(2.9)\end{aligned}$$

where $\Pi_K(x') := \arg\min_{x\in K} \|x - x'\|$ is the projection of x' onto the convex body K. See Figure 1.

Proposition 2.2 (Pythagorean property)**.** *Given a convex body $K \subseteq \mathbb{R}^n$, let $a \in K$ and $b' \in \mathbb{R}^n$. Let $b = \Pi_K(b')$. Then $\langle a-b, b'-b\rangle \leq 0$. Hence $\|a-b\|^2 \leq \|a-b'\|^2$.*

Proof. For the first part, the separating hyperplane at b has b' on one side, and all of K (and hence a) on the other side. Hence the angle between $b'-b$ and $a-b$ must be obtuse, giving the negative inner product. For the second part, $\|a-b'\|^2 = \|a-b\|^2 + \|b-b'\|^2 + 2\langle a-b, b-b'\rangle$. But the latter two terms are positive, which proves the lemma. □

Using this, we get that for any point $x^* \in K$,

$$\|x_{t+1} - x^*\|^2 \leq \|x'_{t+1} - x^*\|^2.$$

Using the same potential function (2.3), this inequality implies:

$$f_t(x_t) - f_t(x^*) + \Phi_{t+1} - \Phi_t \le f_t(x_t) - f_t(x^*) + \frac{1}{2\eta}\left(\|x'_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right),$$

or in other words, the projection only helps and we can follow the analysis from §2.1 starting at (2.7) to bound the amortized cost by $\eta G^2/2$. So this gives a regret bound identical to that of Theorem 2.1.

2.2 Strong convexity analysis

Let us a prove a better regret (and convergence) bound when the functions are "not too flat." A function f over a convex set K is α-*strongly convex*, where $\alpha \ge 0$, if for all $u, v \in K$

$$f(\lambda u + (1-\lambda)v) \le \lambda f(u) + (1-\lambda) f(v) - \frac{\alpha}{2}\lambda(1-\lambda)\|v-u\|^2 \tag{2.10}$$

for all $\lambda \in [0,1]$. For the case of differentiable functions, this is equivalent to saying that for all $x, y \in K$,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y-x\|^2. \tag{2.11}$$

For α-strongly convex functions f_t, we use the same update step but vary the step size η_t. Specifically,

$$\boxed{x_{t+1} \leftarrow \Pi_K(x_t - \eta_t \nabla f_t(x_t))}, \tag{2.12}$$

where $\eta_t = 1/(\alpha(t+1))$.

Theorem 2.3 (Strongly convex functions)**.** *If the functions f_t are α-strongly convex and G is an upper bound on $\|\nabla f_t(x)\|$ for all $x \in K$, the update rule (2.12) with $\eta_t = 1/(\alpha(t+1))$ guarantees an average regret of*

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(f_t(x_t) - f_t(x^*)\big) \le \frac{G^2 \log T}{2T\alpha}.$$

Proof. The potential function is now

$$\boxed{\Phi_t = \frac{1}{2\eta_{t-1}}\|x_t - x^*\|^2 = \frac{t\alpha}{2}\|x_t - x^*\|^2}. \tag{2.13}$$

The change in potential. Let $x'_{t+1} := x_t - \eta_t \nabla f_t(x_t)$ denote the intermediate point after the gradient step, but before performing the projection.

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \frac{\alpha(t+1)}{2}\|x_{t+1} - x^*\|^2 - \frac{\alpha t}{2}\|x_t - x^*\|^2 \\
&= \frac{\alpha}{2}\|x_t - x^*\|^2 + \frac{1}{2\eta_t}\left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\
&\le \frac{\alpha}{2}\|x_t - x^*\|^2 + \frac{1}{2\eta_t}\left(\|x'_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right) && \text{(by Prop. 2.2)} \\
&= \frac{\alpha}{2}\|x_t - x^*\|^2 + \langle \nabla f_t(x_t), x^* - x_t\rangle + \frac{\eta_t}{2}\|\nabla f_t(x_t)\|^2. && \text{(by (2.7))}
\end{aligned}$$

The amortized cost.

$$\begin{aligned} & f_t(x_t) - f_t(x^*) + \Phi_{t+1} - \Phi_t \\ & \quad \leq \underbrace{f_t(x_t) - f_t(x^*) + \frac{\alpha}{2}\|x_t - x^*\|^2 + \langle \nabla f_t(x_t), x^* - x_t \rangle}_{\leq 0 \text{ by } \alpha\text{-strong convexity}} + \frac{\eta_t}{2}\|\nabla f_t(x_t)\|^2 \\ & \quad \leq \frac{\eta_t}{2}\|\nabla f_t(x_t)\|^2 \leq \frac{\eta_t G^2}{2} \quad \text{(by bound on gradients)}. \end{aligned} \tag{2.14}$$

Now summing over all time steps t, the total regret is

$$\sum_t (f_t(x_t) - f_t(x^*)) \leq \Phi_0 + \sum_t \frac{\eta_t}{2} G^2 \leq 0 + \frac{G^2 \log T}{2\alpha}.$$

Hence total regret only increases logarithmically as $\log T$ with time if the f_t are strongly convex, as opposed to $\sqrt{T}$ in Theorem 2.1. □

This bound of $O(\log T)$ on the average regret is tight: Takimoto and Warmuth [25] show a matching lower bound. However, in the *offline* optimization setting where we have a fixed function $f_t = f$, using the same analysis but a better averaging shows a convergence rate of $O(1/T)$ with respect to a convex combination of the points x_t.

Theorem 2.4 (Strongly convex functions: Take II)**.** *Let f be α-strongly convex with gradients satisfying $\|\nabla f(x)\| \leq G$ for all x, and x_t be the iterates produced by applying the update rule (2.1) with $\eta_t = 1/(\alpha t)$. For any $T \geq 1$, let $\bar{x}_T := \sum_{t=1}^T \lambda_t x_t$ denote the convex combination of x_t with*

$$\lambda_t = \frac{2t}{T(T+1)}.$$

Then,

$$f(\bar{x}_T) - f(x^*) \leq \frac{G^2}{\alpha(T+1)}.$$

Proof. Instead of summing up (2.14) directly over t in the regret analysis above, we first multiply (2.14) by t, and then sum over t to obtain

$$\sum_{t=1}^T t(f_t(x_t) - f_t(x^*)) \leq \frac{1}{2\alpha} T G^2.$$

Using $f_t = f$ and dividing by $T(T+1)/2$ throughout, and by the convexity of f we obtain

$$f(\bar{x}_T) - f(x^*) \leq \frac{G^2}{\alpha(T+1)}.$$ □

Finally, we remark that in the constrained case the same analysis with the same potential function works, exactly for same reason as in 2.1.1.

3 Bounds for smooth functions

We now turn to the setting where the functions are Lipschitz smooth, i. e., when the gradient does not change too rapidly. We know that in the online case, the average regret of $O(1/\sqrt{T})$ is tight even for linear functions [7]. However we get better guarantees for the *offline* setting where the function $f_t = f$ for all time steps. The potential functions now look more like (1.2), and use the difference $(f(x_t) - f(x^*))$ in function value, not just in action space.

Define a function f to be β-*Lipschitz smooth* (or simply β-smooth) if for all u, v

$$f(\lambda u + (1-\lambda)v) \geq \lambda f(u) + (1-\lambda) f(v) - \frac{\beta}{2}\lambda(1-\lambda)\|v-u\|^2 \tag{3.1}$$

for all $\lambda \in [0,1]$. For the case of differentiable functions, this is equivalent to saying that for x, y, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y-x\|^2 . \tag{3.2}$$

Observe the inequalities here are in the opposite directions from the definitions of convexity (1.1) and strong-convexity (2.11). Indeed, smoothness implies that the function does not "grow too fast" anywhere. The smoothness condition is equivalent to requiring that the gradients are Lipschitz continuous, i. e., $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x-y\|_2$ for all x, y.

3.1 Smooth gradient descent

The update rule in this case has a time-invariant multiplier (where we use $\nabla_t := \nabla f(x_t)$ for brevity).

$$\boxed{x_{t+1} \leftarrow x_t - \frac{1}{\beta}\nabla_t} . \tag{3.3}$$

We first show an analysis based on a very natural potential, that gives a slightly sub-optimal bound with an additional $\log T$ factor. We improve this later by slightly modifying the potential.

Theorem 3.1 (Smooth functions)**.** *If f is β-smooth and $D := \max_x\{\|x-x^*\|_2 \mid f(x) \leq f(x_0)\}$, the update rule (3.3) guarantees*

$$f(x_T) - f(x^*) \leq \beta \, \frac{D^2(1+\ln T)}{2T} .$$

Proof. To show a convergence rate of $O(1/t)$, perhaps the most natural approach is to consider the potential

$$\boxed{\Phi_t = t \cdot (f(x_t) - f(x^*))}$$

and try to show that $\Phi_T = O(1)$. This works, but gives a weaker bound of $\Phi_T = O(\log T)$ (note that conveniently, $\Phi_0 = 0$) and hence $f(x_T) - f(x^*) = O(\log T)/T$. Later we get rid of the logarithmic term.

The potential change.

$$\begin{aligned}\Phi_{t+1}-\Phi_t &= (t+1)(f(x_{t+1})-f(x^*))-t(f(x_t)-f(x^*))\\ &= (t+1)(f(x_{t+1})-f(x_t))+(f(x_t)-f(x^*))\,.\end{aligned} \tag{3.4}$$

To bound the first term, we use the smoothness of f with $x=x_t$ and $y=x_{t+1}=x_t-\eta_t\nabla_t$:

$$f(x_{t+1})\le f(x_t)-\eta_t\cdot\|\nabla_t\|_2^2+\tfrac{\beta}{2}\cdot\eta_t^2\cdot\|\nabla_t\|_2^2\,.$$

The choice of $\eta_t=1/\beta$ minimizes the right hand side above to give

$$f(x_{t+1})\le f(x_t)-\tfrac{1}{2\beta}\|\nabla_t\|_2^2\,. \tag{3.5}$$

For the second term in (3.4), just use convexity and Cauchy-Schwarz:

$$\begin{aligned}f(x_t)-f(x^*)\le\langle\nabla_t,x_t-x^*\rangle &\le\|\nabla_t\|_2\cdot\|x_t-x^*\|_2\\ &\le 1/2\,(a\|\nabla_t\|^2+(1/a)\|x_t-x^*\|^2)\,,\end{aligned} \tag{3.6}$$

for any parameter $a>0$. Note that (3.5) ensures that $f(x_t)\le f(x_{t-1})\le\cdots\le f(x_0)$, so let us define $D:=\max\{\|x-x^*\|_2\mid f(x)\le f(x_0)\}$. So the potential change is

$$\Phi_{t+1}-\Phi_t\le(t+1)\cdot(-\tfrac{1}{2\beta})\|\nabla_t\|_2^2+\tfrac{1}{2}\,(a\|\nabla_t\|_2^2+D^2/a). \tag{3.7}$$

Choosing $a=(t+1)/\beta$ cancels the gradient terms. Hence the potential increase is at most $D^2\beta/(2(t+1))$, and

$$f(x_T)-f(x^*)=\frac{\Phi_T}{T}=\frac{1}{T}\sum_{t=0}^{T-1}(\Phi_{t+1}-\Phi_t)\le\frac{1}{T}\sum_{t=0}^{T-1}\frac{D^2}{2(t+1)}\beta\le\beta\,\frac{D^2(1+\ln T)}{2T}\,. \qquad\square$$

The intuition is evident from (3.5) and (3.6): we improve a lot by (3.5) when the gradients are large, or else we are close to the optimum by (3.6).

A tighter analysis. The logarithmic dependence in Theorem 3.1 can be removed by a simple trick of multiplying the potential by a linear term in t, which avoids the sum over $1/t$.

Theorem 3.2 (Smooth functions: Take II). *If f is β-smooth, and $D:=\max_x\{\|x-x^*\|_2\mid f(x)\le f(x_0)\}$, the update rule (3.3) guarantees*

$$f(x_T)-f(x^*)\le\beta\,\frac{2D^2}{T+1}\,.$$

Proof. Consider the following potential

$$\boxed{\Phi_t=t(t+1)\cdot(f(x_t)-f(x^*))}\,.$$

The potential change is

$$\Phi_{t+1}-\Phi_t=(t+1)(t+2)\cdot(f(x_{t+1})-f(x_t))+2(t+1)\cdot(f(x_t)-f(x^*))\,.$$

Plugging in (3.5) and (3.6) gives

$$\leq (t+1)(t+2)\cdot\left(-\tfrac{1}{2\beta}\|\nabla_t\|^2\right)+2(t+1)\cdot\|\nabla_t\|_2\cdot D \leq 2D^2\beta\cdot\frac{t+1}{t+2},$$

where the last inequality is the maximum value of the preceding expression obtained at $\|\nabla_t\| = 2\beta D/(t+2)$. Summing over the time steps, $\Phi_T \leq T\cdot 2D^2\beta$, so

$$f(x_T)-f(x^*) \leq \frac{2D^2\beta\cdot T}{T(T+1)} = \beta\,\frac{2D^2}{T+1}. \qquad \square$$

3.1.1 Yet another proof

Let's see yet another proof that gets rid of the logarithmic term. Interestingly, the potential function now combines both the difference in the function value, and the distance in the "action" space.

Theorem 3.3 (Smooth functions: Take III)**.** *If f is β-smooth, the update rule (3.3) guarantees*

$$f(x_T)-f(x^*) \leq \beta\,\frac{\|x_0-x^*\|^2}{2T}.$$

Proof. Consider the potential of the form

$$\boxed{\Phi_t = t\,(f(x_t)-f(x^*))+\frac{a}{2}\|x_t-x^*\|^2}$$

where a will be chosen based on the analysis below. As $\Phi_0 = (a/2)\|x_0-x^*\|^2$, if we show that Φ_t is non-increasing,

$$\frac{a}{2}\|x_0-x^*\|^2 = \Phi_0 \geq \Phi_T = T(f(x_t)-f(x^*))+\frac{a}{2}\|x_t-x^*\|^2$$

which gives

$$f(x_t)-f(x^*) \leq \frac{a}{2T}\,\|x_0-x^*\|^2$$

as desired.

The potential difference can be written as:

$$\Phi_{t+1}-\Phi_t = (t+1)\underbrace{(f(x_{t+1})-f(x_t))}_{(3.5)}+\underbrace{f(x_t)-f(x^*)}_{\text{(convexity)}}+\frac{a}{2}\underbrace{(\|x_{t+1}-x^*\|^2-\|x_t-x^*\|^2)}_{(2.7)}. \tag{3.8}$$

Using the bounds from the mentioned inequalities,

$$\leq (t+1)\cdot\overbrace{-\tfrac{1}{2\beta}\|\nabla_t\|_2^2}+\overbrace{\langle\nabla_t, x_t-x^*\rangle}+\frac{a}{2}\overbrace{\left(2\eta_t\langle\nabla_t, x^*-x_t\rangle+\eta_t^2\|\nabla_t\|_2^2\right)} \tag{3.9}$$

where $\eta_t = 1/\beta$ in this case. Now, we set $a = 1/\eta_t = \beta$ to cancel the inner-product terms, which gives

$$\Phi_{t+1}-\Phi_t \leq -\left(\frac{t}{2\beta}\right)\|\nabla_t\|^2 \leq 0. \qquad \square$$

This guarantee is almost the same as in Theorem 3.2, with a slightly better definition of the distance term ($\|x_0-x^*\|^2$ vs. D^2). But we will revisit and build on this proof when we talk about Nesterov acceleration in §5.

3.1.2 Projected smooth gradient descent

We now consider the constrained minimization problem for a convex body K. As previously, the update involves taking a step and then projecting back onto K:

$$\boxed{\begin{array}{c} x'_{t+1} \leftarrow x_t - (1/\beta)\,\nabla f(x_t)\,, \\ x_{t+1} \leftarrow \Pi_K(x'_{t+1})\,. \end{array}} \tag{3.10}$$

Theorem 3.4 (Constrained smooth optimization)**.** *If f is β-smooth, the update rule (3.10) guarantees*

$$f(x_T) - f(x^*) \le \frac{\frac{\beta}{2}\|x_0 - x^*\|^2}{T}\,.$$

Proof. We use the same potential as in Theorem 3.3:

$$\boxed{\Phi_t = t(f(x_t) - f(x^*)) + \frac{\beta}{2}\|x_t - x^*\|^2}.$$

The potential difference is now written as:

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= t(f(x_{t+1}) - f(x_t)) + f(x_{t+1}) - f(x^*) + \tfrac{\beta}{2}\underbrace{(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2)}_{\|a\|^2 - \|a+b\|^2 = -2\langle a,b\rangle - \|b\|^2} \\ &\le t\underbrace{(f(x_{t+1}) - f(x_t))}_{(\star)} + \underbrace{f(x_{t+1}) - f(x^*)}_{(\star\star)} - \tfrac{\beta}{2}\left(2\langle x_t - x_{t+1}, x_{t+1} - x^*\rangle + \|x_t - x_{t+1}\|^2\right).\end{aligned} \tag{3.11}$$

As x_{t+1} is the projected point, we cannot directly use (3.5) to bound the first and second terms, but we can show the following claim (which we prove later) that follows from smoothness:

Claim 3.5. *For any $y \in K$, $f(x_{t+1}) - f(y) \le \beta\langle x_t - x_{t+1}, x_t - y\rangle - (\beta/2)\|x_t - x_{t+1}\|^2$.*

Using Claim 3.5 with $y = x_t$ and $y = x^*$ to bound the first and second terms of (3.11), we get

$$\begin{aligned}&\le t\cdot \overbrace{0}^{(\star)} + \overbrace{\beta\langle x_t - x_{t+1}, x_t - x^*\rangle - \beta/2\|x_t - x_{t+1}\|^2}^{(\star\star)} - \beta/2\left(2\langle x_t - x_{t+1}, x_{t+1} - x^*\rangle + \|x_t - x_{t+1}\|^2\right) \\ &= \beta\langle x_t - x_{t+1}, x_t - x_{t+1}\rangle - \beta\|x_t - x_{t+1}\|^2 = 0\,.\end{aligned}$$

This completes the proof. □

Proof of Claim 3.5. We write $f(x_{t+1}) - f(y) = (f(x_{t+1}) - f(x_t)) + (f(x_t) - f(y))$. Now using smoothness and convexity for the first and second terms respectively, we have

$$\begin{aligned}f(x_{t+1}) - f(y) &\le \left(\langle \nabla_t, x_{t+1} - x_t\rangle + \beta/2\|x_{t+1} - x_t\|^2\right) + \langle \nabla_t, x_t - y\rangle \\ &= \langle \nabla_t, x_{t+1} - y\rangle + \beta/2\|x_{t+1} - x_t\|^2\,.\end{aligned} \tag{3.12}$$

Since $\beta\langle x_{t+1} - x'_{t+1}, x_{t+1} - y\rangle \le 0$ by the Pythagorean property Prop. 2.2,

$$\begin{aligned}\langle \nabla_t, x_{t+1} - y\rangle &= \langle \beta(x_t - x'_{t+1}), x_{t+1} - y\rangle \le \beta\langle x_t - x_{t+1}, x_{t+1} - y\rangle \\ &= \beta\langle x_t - x_{t+1}, x_t - y\rangle - \beta\|x_{t+1} - x_t\|^2\,.\end{aligned}$$

Substituting into (3.12) gives the result. □

3.1.3 The Frank–Wolfe method

One drawback of projected gradient descent is the projection step: given a point x' and a body K, finding the closest point $\Pi_K(x')$ might be computationally expensive. Instead, we can use a different rule, the *Frank–Wolfe* method (also called *conditional gradient descent*) [11], that implements each gradient step using linear optimization over the body K. Loosely, at each timestep we find the point in K that is furthest from the current point in the direction of the negative gradient, and move a small distance towards it.

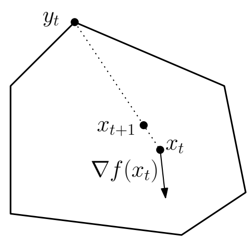

Figure 2: The Frank–Wolfe Update.

Formally, the update rule for Frank–Wolfe method is simple:

$$\boxed{\begin{aligned} y_t &\leftarrow \arg\min_{y\in K} \langle \nabla_t, y\rangle, \\ x_{t+1} &\leftarrow (1-\eta_t)x_t + \eta_t y_t. \end{aligned}} \tag{3.13}$$

Setting $\eta_t = 1/(t+1)$ in hindsight will give the following result.

Theorem 3.6 (Smooth functions: Frank–Wolfe)**.** *If f is β-smooth, K is a convex body with diameter $D := \max_{x,y\in K}\|x-y\|$, then the update rule (3.13) guarantees*

$$f(x_T) - f(x^*) \le \beta\, \frac{D^2\,(1+\ln T)}{2T}\,.$$

Proof. We use the simplest potential function from Theorem 3.1:

$$\boxed{\Phi_t = t\cdot(f(x_t) - f(x^*))},$$

and hence the change in potential is again:

$$\Phi_{t+1} - \Phi_t = (t+1)(f(x_{t+1}) - f(x_t)) + (f(x_t) - f(x^*))\,. \tag{3.14}$$

To bound the change in potential (3.14), we observe that $x_{t+1} - x_t = \eta_t(y_t - x_t)$.

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le \langle \nabla_t, x_{t+1} - x_t\rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 && \text{(by smoothness)} \\ &= \eta_t\langle \nabla_t, y_t - x_t\rangle + \frac{\beta\eta_t^2}{2}\|y_t - x_t\|^2 \end{aligned}$$

$$\leq \eta_t \langle \nabla_t, x^* - x_t \rangle + \frac{\beta \eta_t^2}{2} \|y_t - x_t\|^2 \qquad \text{(by optimality of } y_t\text{)}$$
$$f(x_t) - f(x^*) \leq \langle \nabla_t, x_t - x^* \rangle . \qquad \text{(by convexity)}$$

Setting $\eta_t := 1/(t+1)$ cancels the linear terms and hence the potential change (3.14) is at most $\beta \eta_t D^2/2$. Summing over t and using $\Phi_0 = 0$, the final potential $\Phi_T \leq D^2(1 + \ln T)$, and hence

$$f(x_T) - f(x^*) = \beta \, \frac{(1 + \ln T) \cdot D^2}{2T} . \qquad \square$$

We can remove the logarithmic dependence in the error by multiplying the potential by $(t+1)$ as in Theorem 3.2; this gives the following theorem, whose simple proof we omit.

Theorem 3.7 (Smooth functions: Frank–Wolfe, take II)**.** *If f is β-smooth, K is a convex body with $D := \max_{x,y \in K} \|x - y\|$, then the update rule (3.13) with $\eta_t = 2/(t+1)$ guarantees*

$$f(x_T) - f(x^*) \leq 2\beta \, \frac{D^2}{T+1} .$$

3.2 Well-conditioned functions

If a function is both α-strongly convex and β-smooth, it must be that $\alpha \leq \beta$. The ratio $\kappa := \beta/\alpha$ is called the *condition number* of the convex function. We now show a much stronger convergence guarantee for "well-conditioned" functions, i. e., functions with small κ values. The update rule is the same as for smooth functions:

$$\boxed{x_{t+1} \leftarrow x_t - \frac{1}{\beta} \nabla_t} . \tag{3.15}$$

Theorem 3.8 (GD: Well-conditioned)**.** *Given a function f that is both α-strongly convex and β-smooth, define $\kappa := \beta/\alpha$. The update rule (3.15) ensures*

$$f(x_T) - f(x^*) \leq \exp(-T/\kappa) \cdot (f(x_0) - f(x^*)) \qquad \text{for all } x^*.$$

Proof. We set $\gamma = 1/(\kappa - 1)$ for brevity,[1] and use the potential

$$\boxed{\Phi_t = (1+\gamma)^t \cdot (f(x_t) - f(x^*))} . \tag{3.16}$$

This is a natural potential to use, as we wish to show that $f(x_T) - f(x_0)$ falls exponentially with T.

[1]Note that $\kappa = 1$ iff $f(x) = ax^\intercal x + b^\intercal x + c$ for suitable scalars a, c and $b \in \mathbb{R}^n$; in this case it is easily checked that the optimum solution x^* is reached in a single step.

The potential change. A little rearrangement gives us

$$\Phi_{t+1} - \Phi_t = (1+\gamma)^t \cdot \Big((1+\gamma)\big(f(x_{t+1}) - f(x_t)\big) + \gamma\big(f(x_t) - f(x^*)\big) \Big). \tag{3.17}$$

We bound the two terms separately. Using the smoothness analysis from (3.5):

$$f(x_{t+1}) - f(x_t) \le -\frac{1}{2\beta}\|\nabla_t\|^2.$$

And by the definition of strong convexity,

$$f(x_t) - f(x^*) \le \langle \nabla_t, x_t - x^* \rangle - \frac{\alpha}{2}\|x_t - x^*\|^2 \le \frac{1}{2\alpha}\|\nabla_t\|^2$$

where the second inequality uses $\langle a,b\rangle - \|b\|^2/2 \le \|a\|^2/2$. Plugging this back into (3.17) gives

$$(1+\gamma)^t \left(-\frac{1+\gamma}{2\beta} + \frac{\gamma}{2\alpha} \right) \|\nabla_t\|^2$$

which is 0 by our choice of γ. Hence, after T steps,

$$\begin{aligned} f(x_T) - f(x^*) &\le (1+\gamma)^{-T}(f(x_0) - f(x^*)) = (1 - {}^1\!/\kappa)^T (f(x_0) - f(x^*)) \\ &\le \mathrm{e}^{-T/\kappa}(f(x_0) - f(x^*)). \end{aligned} \tag{3.18}$$

Hence the proof. □

Here we can show that the algorithm's point x_T also gets rapidly closer to x^*. If x^* is the optimal point, we know $\nabla f(x^*) = 0$. Now smoothness gives $f(x_0) - f(x^*) \le (\beta/2)\|x_0 - x^*\|^2$, and strong convexity gives $(\alpha/2)\|x_T - x^*\|^2 \le f(x_T) - f(x^*)$. Plugging into (3.18) gives us that

$$\|x_T - x^*\|^2 \le \kappa \mathrm{e}^{-T/\kappa} \cdot \|x_0 - x^*\|^2.$$

A few remarks. Firstly, Theorem 3.8 implies that reducing the error by a factor of $1/2$ can be achieved by increasing T *additively* by $\kappa \ln 2$. Hence if the condition number κ is constant, every constant number of rounds of gradient descent gives us one additional bit of accuracy! This behavior, where getting error bounded by ε requires $O(\log \varepsilon^{-1})$ steps, is called *linear convergence* in the numerical analysis literature.

One may ask if the convergence for smooth, and for well-conditioned functions is optimal as a function of T. The answer is no: a famed result of Nesterov gives faster (and optimal) convergence rates. We see this result and a potential-function-based proof in §5.

Finally, the proof of Theorem 3.8 can be extended to the constrained case using the same potential function and the update rule (3.10), but now using an analog of Claim 3.5 that shows that for any $y \in K$,

$$f(x_{t+1}) - f(y) \le \beta \langle x_t - x_{t+1}, x_t - y \rangle - {}^{\beta}\!/_2 \|x_t - x_{t+1}\|^2 - {}^{\alpha}\!/_2 \|x_t - y\|^2.$$

We omit the simple proof.

4 The mirror-descent framework

The gradient descent algorithms in the previous sections work by adding some multiple of the gradient to current point. However, this should strike the reader as somewhat strange, since the point x_t and the gradient $\nabla f(x_t)$ are objects that lie in different spaces and should be handled accordingly. In particular, if x_t lies in some vector space E, the gradient $\nabla f(x_t)$ lies in the dual vector space E^*. (This did not matter earlier since $\mathbb{R}^n$ equipped with the Euclidean norm is self-dual, but now we want to consider general norms and would like to be careful.)

A key insight of Nemirovski and Yudin [18] was that substantially more general and powerful results can be obtained, without much additional work, by considering these spaces separately. For example, it is well-known (and we will show) that the classic multiplicative-weights update method can be obtained as a special case of this general approach.

4.1 Basic mirror descent

The key idea in mirror descent is to define an injective mapping between E to E^*, which is called the *mirror map*. Given a point x_t, we first map it to E^*, make the gradient update there, and then use the inverse map back to obtain the point x_{t+1}.

We start some basic concepts and notation. Consider some vector space E with an inner product $\langle \cdot, \cdot \rangle$, and define a norm $\|\cdot\|$ on E. To measure distances in E^*, we use the *dual norm* defined as

$$\|y\|_* := \max_{x:\|x\|=1} \langle x, y \rangle . \tag{4.1}$$

By definition we have

$$\langle x, y \rangle \leq \|x\| \cdot \|y\|_* , \tag{4.2}$$

which is often referred to as the *generalized Cauchy-Schwarz* inequality.

A function h is α-*strongly convex* with respect to $\|\cdot\|$ if

$$h(y) \geq h(x) + \langle \nabla h(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2 . \tag{4.3}$$

Such a strongly convex function h defines a map from E to E^* via its gradient: indeed, the map $x \mapsto \nabla h(x)$ takes the point $x \in E$ into a point in the dual space E^*. The strong convexity ensures that the map is 1-1 (i. e., $\nabla h(x) \neq \nabla h(y)$ for $x \neq y$). Moreover, the map $\nabla h(\cdot)$ is also surjective, so for any $\theta \in E^*$ there is an inverse $x \in E$ such that $\nabla h(x) = \theta$. In fact, this inverse map is given by the gradient of the Fenchel dual for h, i. e., $\nabla h^*(\theta) = x \iff \nabla h(x) = \theta$. (For the reader not familiar with Fenchel duality, it suffices to interpret $\nabla h^*(\theta)$ merely as $(\nabla h)^{-1}(\theta)$.) Readers interested in the technical details can see, e. g., [2] or [4, Chapter 7].

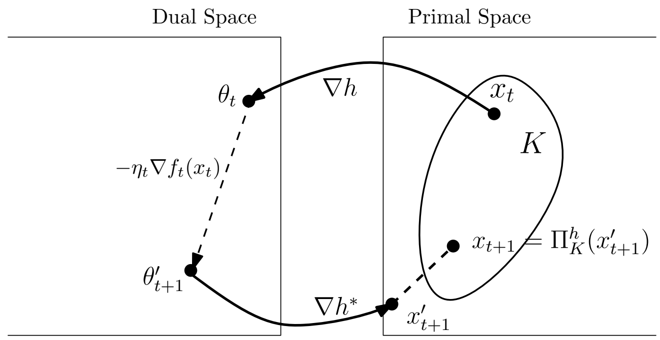

Figure 3: Mirror Descent.

4.1.1 The update rules

Since $\nabla h : E \to E^*$ gives us a map from the primal space E to the dual space E^*, we keep track of the image point $\theta_t = \nabla h(x_t)$ as well. Now, the updates are the natural ones, given by

$$\boxed{\begin{aligned} \theta'_{t+1} &= \theta_t - \eta_t \nabla f_t(x_t), \\ x'_{t+1} &= \nabla h^*(\theta'_{t+1}), \\ x_{t+1} &= \arg\min_{x \in K} D_h(x \,\|\, x'_{t+1}). \end{aligned}} \tag{4.4}$$

In other words, given $x_t \in E$, we add η_t times the negative gradient to its image $\theta_t = \nabla h(x_t)$ in the dual space to get θ'_{t+1}, pull the result back to $x'_{t+1} \in E$ (using the inverse mapping $x'_{t+1} = \nabla h^*(\theta'_{t+1})$), and project it back onto K to get x_t. Of course, we may not want to use the Euclidean distance for the projection; the "right" distance in this case is the *Bregman divergence* $D_h(y \,\|\, x)$ from x to y, which we discuss shortly.

An equivalent way to present the mirror descent update is the following:

$$\boxed{x_{t+1} = \arg\min_{x \in K} \left\{ \langle \eta_t \nabla f_t(x_t), x - x_t \rangle + D_h(x \,\|\, x_t) \right\}.} \tag{4.5}$$

This is a generalization of (2.2). The equivalence is easy to see in the unconstrained case (just take derivatives), for the constrained case one uses the KKT conditions.

4.1.2 Bregman divergences

Given a strictly convex function $h : \mathbb{R}^d \to \mathbb{R}$, define the *Bregman divergence* from x to y

$$D_h(y \,\|\, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle$$

to be the "error" at y between the actual function value and the value given by linearization at some point x. The convexity of h means this quantity is non-negative; if h is β-strongly convex with respect to the norm $\|\cdot\|$, then $D_h(y \parallel x) \geq (\beta/2)\|y-x\|^2$. Also, $D_h(y \parallel x)$ is a convex function of y (for a fixed x, this is a convex function minus a linear term), and the gradient of the divergence with respect to the first argument is $\nabla_y(D_h(y \parallel x)) = \nabla h(y) - \nabla h(x)$.

For example, the function $h(x) := (1/2)\|x\|_2^2$ is 1-strongly convex with respect to ℓ_2 (and hence strictly convex), and the associated Bregman divergence $D_h(y \parallel x) = (1/2)\|y-x\|_2^2$, half the squared ℓ_2 distance. This distance is not a metric, since it does not satisfy the triangle inequality. Or consider the *negative entropy function* $h(x) := \sum_i x_i \ln x_i$ defined on the probability simplex $\triangle_n := \{x \in [0,1]^n \mid \sum_i x_i = 1\}$. For $x, y \in \triangle_n$, the associated Bregman divergence $D_h(y \parallel x)$ is $\sum_i y_i \ln(y_i/x_i)$, the relative entropy or *Kullback-Leibler (KL) divergence* from x to y. This distance is not even symmetric in x and y.

Bregman projection. Given a convex body K and a strictly convex function h, we define the *Bregman projection* of a point x' on K as

$$\Pi_K^h(x') = \operatorname{argmin}_{x \in K} D_h(x \parallel x').$$

If $x' \in K$, then $\Pi_K^h(x') = x'$ because $D_h(x' \parallel x') = 0$. For $h(x) = (1/2)\|x\|^2$, this corresponds to the usual Euclidean projection. A very useful feature of Bregman projections is that they satisfy a "Pythagorean inequality" with respect to the divergence, analogous to Fact 2.2.

Proposition 4.1 (Generalized Pythagorean property)**.** *Given a convex body $K \subseteq \mathbb{R}^n$, let $a \in K$ and $b' \in \mathbb{R}^n$. Let $b = \Pi_K^h(b')$. Then*

$$\langle \nabla h(b') - \nabla h(b), a - b \rangle \leq 0.$$

In particular,

$$D_h(a \parallel b') \geq D_h(a \parallel b) + D_h(b \parallel b'), \tag{4.6}$$

and hence $D_h(a \parallel b) \leq D_h(a \parallel b')$.

Proof. Recall that for any convex function g and convex body K, if $x^* = \operatorname{argmin}_{x \in K} g(x)$ is the minimizer of g in K, then $\langle \nabla g(x^*), y - x^* \rangle \geq 0$ for all $y \in K$. Using $g(x) = D_h(x \parallel b')$, and noting that $g(x)$ is convex with $\nabla g(x) = \nabla h(x) - \nabla h(b')$ and that the minimizer $x^* = b$, we get $\langle \nabla h(b) - \nabla h(b'), a - b \rangle \geq 0$ for all $a \in K$.

For the second part, expand the terms using the definition of $D_h(a \parallel b)$ and cancel the common terms, the desired inequality turns out to be equivalent to $\langle \nabla h(b') - \nabla h(b), a - b \rangle \leq 0$. The last inequality uses that the divergences are non-negative. □

4.1.3 The analysis

We consider the more general *online* optimization setting, and prove the following regret bound.

Theorem 4.2. *Let K be a convex body, $f_1, \ldots, f_T$ be convex functions defined on K, $\|\cdot\|$ be a norm, and h be an α_h-strongly convex function with respect to $\|\cdot\|$. The mirror descent algorithm starting at x_0 and taking constant step size $\eta_t = \eta$ in every iteration, produces $x_1, \ldots, x_T$ such that*

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{n} f_t(x^*) \leq \frac{D_h(x^* \parallel x_0)}{\eta} + \frac{\eta \sum_{t=1}^{T} \|\nabla f_t(x_t)\|_*^2}{2\alpha_h}, \quad \text{for all } x^* \in K. \tag{4.7}$$

Proof. Define the potential

$$\boxed{\Phi_t = \frac{D_h(x^* \parallel x_t)}{\eta}}. \tag{4.8}$$

Observe that plugging in $h(x) = (1/2)\|x\|_2^2$ gives us the potential function (2.3) for the Euclidean norm.

The potential change. For brevity, use $\nabla_t := \nabla f_t(x_t)$.

$$\begin{aligned}
&D_h(x^* \parallel x_{t+1}) - D_h(x^* \parallel x_t)\\
&\qquad \le D_h(x^* \parallel x'_{t+1}) - D_h(x^* \parallel x_t) \qquad\qquad \text{(generalized Pythagorean property)}\\
&\qquad = h(x^*) - h(x'_{t+1}) - \langle \underbrace{\nabla h(x'_{t+1})}_{\theta'_{t+1}}, x^* - x'_{t+1}\rangle - h(x^*) + h(x_t) + \langle \underbrace{\nabla h(x_t)}_{\theta_t}, x^* - x_t\rangle\\
&\qquad = h(x_t) - h(x'_{t+1}) - \langle \theta'_{t+1}, x_t - x'_{t+1}\rangle - \langle \theta'_{t+1} - \theta_t, x^* - x_t\rangle\\
&\qquad = \underbrace{h(x_t) - h(x'_{t+1}) - \langle \theta_t, x_t - x'_{t+1}\rangle}_{\text{strong convexity}} + \langle \eta_t \nabla_t, x_t - x'_{t+1}\rangle + \langle \eta_t \nabla_t, x^* - x_t\rangle\\
&\qquad \le -\frac{\alpha_h}{2}\|x'_{t+1} - x_t\|^2 + \eta_t\langle \nabla_t, x_t - x'_{t+1}\rangle + \eta_t\langle \nabla_t, x^* - x_t\rangle\\
&\qquad \le \frac{\eta_t^2}{2\alpha_h}\|\nabla_t\|_*^2 + \eta_t\langle \nabla_t, x^* - x_t\rangle\,.
\end{aligned} \tag{4.9}$$

The last inequality uses generalized Cauchy-Schwarz to get $\langle a,b\rangle \le \|b\|\|a\|_* \le \|b\|^2/2 + \|a\|_*^2/2$. Observe that (4.9) precisely maps to (2.7) when we consider the Euclidean norm.

The amortized cost. Recall that we set $\eta_t = \eta$ for all steps. Hence, dividing (4.9) and substituting,

$$f_t(x_t) - f_t(x^*) + (\Phi_{t+1} - \Phi_t) \le \underbrace{f_t(x_t) - f_t(x^*) + \langle \nabla_t, x^* - x_t\rangle}_{\le 0 \text{ by convexity of } f_t} + \frac{\eta}{2\alpha_h}\|\nabla_t\|_*^2\,.$$

The total regret then becomes

$$\sum_t (f_t(x_t) - f_t(x^*)) \le \Phi_0 + \sum_t \frac{\eta}{2\alpha_h}\|\nabla_t\|_*^2 \le \frac{D_h(x^* \parallel x_0)}{\eta} + \frac{\eta \sum_{t=1}^T \|\nabla_t\|_*^2}{2\alpha_h}\,.$$

Hence the proof. □

4.1.4 Special cases

To get some intuition, let us look at some well-known special cases. If we use the ℓ_2 norm, and $h(x) := (1/2)\|x\|_2^2$ which is clearly 1-strongly convex with respect to ℓ_2, the associated Bregman divergence $D_h(x^* \parallel x) = (1/2)\|x^* - x\|_2^2$. Moreover, the Euclidean norm is self-dual, so if we bound $\|\nabla f_t\|_2$ by G, the total regret bound above is

$$\frac{1}{2\eta}\|x^* - x_0\|_2^2 + \eta T G^2/2\,.$$

This is the same result for projected gradient descent we derived in Theorem 2.1—and in fact the algorithm is also precisely the same.

Now consider the ℓ_1 norm, with K being the probability simplex $\triangle_n := \{x \in [0,1]^n \mid \sum_i x_i = 1\}$. If we choose the negative entropy function $h(x) := \sum_i x_i \ln x_i$, then $D_h(x^* \parallel x)$ is just the well-known Kullback-Liebler divergence. Moreover, Pinsker's inequality says that $KL(p \parallel q) \geq (1/2)\|p-q\|_1^2$, which implies that h is 1-strongly convex with respect to ℓ_1. Applying Theorem 4.2 now gives a regret bound of

$$\frac{KL(x^* \parallel x_0)}{\eta} + \frac{\eta}{2}\sum_t \|\nabla_t\|_\infty^2 .$$

Let's also see what the mirror descent algorithm does in this case. The mirror map takes the point x to $\nabla h(x) = (1+\log x_i)_i$, and the inverse map takes θ to $\nabla h^*(\theta) = (\mathrm{e}^{\theta_i - 1})_i$. This point may be outside the probability simplex, so we do a Bregman projection, which in this case corresponds to just a rescaling $x \mapsto x/\|x\|_1$. Unrolling the process, one can get a closed-form expression for the point x_T:

$$(x_T)_i = \frac{(x_0)_i \cdot \exp\{\sum_t (\nabla_t)_i\}}{\sum_j (x_0)_j \cdot \exp\{\sum_t (\nabla_t)_j\}} .$$

For example, if we specialize even further to online *linear* optimization, where each function $f_t(x) = \langle \ell_t, x\rangle$ for some $\ell_t \in [0,1]^n$, the gradient is ℓ_t and its ℓ_∞-norm is $\|\ell_t\|_\infty \leq 1$, giving us the familiar regret bound of

$$\frac{KL(x^* \parallel x_0)}{\eta} + \frac{\eta T}{2}$$

that we get from the multiplicative weights/Hedge algorithms. Which is not surprising, since this algorithm is precisely the Hedge algorithm!

4.2 An aside: Smooth functions and general norms

Let us consider minimizing a function that is smooth with respect to non-Euclidean norms, in the unconstrained case. When we consider an arbitrary norm $\|\cdot\|$, the definition of a smooth function (3.2) extends seamlessly. Now we can define an update rule by naturally extending (2.2):

$$\boxed{x_{t+1} \leftarrow \arg\min_x \left\{ \frac{1}{2}\|x - x_t\|^2 + \eta_t \langle \nabla_t, x - x_t\rangle \right\}}, \tag{4.10}$$

where the norm is no longer the Euclidean norm, but the norm in question. To evaluate the minimum on the right side, we can use basic Fenchel duality: given a function g, its Fenchel dual is defined as

$$g^\star(\theta) := \max_z \{\langle \theta, z\rangle - g(z)\} .$$

If we define $g(z) = (1/2)\|z\|^2$, it is known that $g^\star(\theta) = (1/2)\|z\|_\star^2$ (see [5, Example 3.27]). Hence

$$\min_x \left\{ \frac{1}{2}\|x - x_t\|^2 + \eta_t \langle \nabla_t, x - x_t\rangle \right\} = -\max_x \left\{ \eta_t \langle \nabla_t, x_t - x\rangle - \frac{1}{2}\|x_t - x\|^2 \right\}$$

$$= -\max_z \left\{ \langle \eta_t \nabla_t, z \rangle - \frac{1}{2}\|z\|^2 \right\} = -\frac{1}{2}\|\eta_t \nabla_t\|_\star^2. \tag{4.11}$$

If a function f is β-smooth with respect to the norm, then setting $\eta_t = 1/\beta$ gives:

$$\begin{aligned} f(x_{t+1}) &\overset{(3.2)}{\leq} f(x_t) + \langle \nabla_t, x_{t+1} - x_t \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t) + \beta\left(\langle \eta_t \nabla_t, x_{t+1} - x_t \rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2\right) = f(x_t) + \beta \cdot \left(-\frac{1}{2}\|\eta_t \nabla_t\|_\star^2\right), \end{aligned}$$

where the last equality uses that x_{t+1} is the minimizer of the expression in (4.11). Summarizing, we get

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla_t\|_\star^2. \tag{4.12}$$

This is analogous to the expression (3.3). Now we can continue the proof as in §3.1, again defining $D := \max\{\|x - x^*\| \mid f(x) \leq f(x_0)\}$, and using the generalized Cauchy-Schwarz inequality to get the general-norm analog of Theorem 3.2, due to Jaggi [14].

Theorem 4.3 (GD: Smooth functions for general norms)**.** *Given a function f that is β-smooth with respect to the norm $\|\cdot\|$, the update rule (4.10) ensures*

$$f(x_T) - f(x^*) \leq \beta\, \frac{2D^2}{T+1}.$$

5 Nesterov acceleration: A potential function proof

In §3, we proved a convergence rate of $O(1/T)$ for smooth functions, using both projected gradient descent and the Frank–Wolfe method. But the lower bound is only $\Omega(1/T^2)$. In this case, the algorithm can be improved: Yurii Nesterov showed how to do it using his *accelerated gradient descent* methods [19]. Recently there has been much interest in gaining a deeper understanding of this process, with proofs using "momentum" methods and continuous-time updates [29, 24, 17, 30, 8].

Let us now see potential-based proofs for his theorem, both for the smooth case, and for the well-conditioned case. We consider only the unconstrained case (i. e., when $K = \mathbb{R}^n$) and the Euclidean norm; the extension to general norms is sketched in §5.4.

5.1 An illustrative failed attempt

One way to motivate Nesterov's accelerated algorithm is to revisit the proof for smooth functions in §3.1.1. Let us recall the essential facts. The potential was

$$\boxed{\Phi_t = t\,(f(x_t) - f(x^*)) + \frac{a}{2}\|x_t - x^*\|^2}$$

for some $a > 0$. Hence the potential difference was:

$$\Phi_{t+1} - \Phi_t = (t+1)\underbrace{(f(x_{t+1}) - f(x_t))}_{(3.5)} + \underbrace{f(x_t) - f(x^*)}_{\text{(convexity)}} + \frac{a}{2}\underbrace{(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2)}_{(2.7)}$$

$$\leq (t+1)\cdot \overbrace{-\tfrac{1}{2\beta}\|\nabla_t\|_2^2} + \overbrace{\langle \nabla_t, x_t - x^*\rangle} + a\overbrace{\left(\eta_t\langle \nabla_t, x^* - x_t\rangle + \tfrac{\eta_t^2}{2}\|\nabla_t\|_2^2\right)} = -\tfrac{t}{2\beta}\|\nabla_t\|^2 \leq 0\,.$$

In that last expression we set $\eta_t = 1/\beta$ and $a = 1/\eta_t = \beta$ to cancel the inner-product terms.

Observe that the potential may be decreasing considerably, by $-t/(2\beta)\cdot\|\nabla_t\|^2$, but we are ignoring this large decrease. If we want to a show an $O(1/t^2)$ rate of convergence, a first (incorrect) attempt to get a better analysis would be to try to apply the analysis above with the potential changed to

$$\Phi_t = t(t+1)\,(f(x_t) - f(x^*)) + \frac{a}{2}\|x_t - x^*\|^2.$$

In particular note the factor $a_t = t(t+1)$ instead of $(t+1)$ above.

At first glance, the potential change $\Phi_{t+1} - \Phi_t$ would be

$$(t+2)(t+1)\cdot \underbrace{\left(-\tfrac{1}{2\beta}\|\nabla_t\|_2^2\right)}_{(3.5)} + 2(t+1)\cdot\langle \nabla_t, x_t - x^*\rangle + a\left(\eta_t\langle \nabla_t, x^* - x_t\rangle + \tfrac{\eta_t^2}{2}\|\nabla_t\|_2^2\right). \tag{5.1}$$

Now if we change the step length η_t from $1/\beta$ to something more aggressive, say $\eta_t = (t+1)/(2\beta)$, and choose $a = 4\beta$, the inner-product terms cancel, and the potential reduction seems to be at most

$$\|\nabla_t\|^2\left(-\frac{(t+1)(t+2)}{2\beta} + \frac{(t+1)^2}{2\beta}\right) \leq 0\,.$$

This would seem to give us an $O(1/T^2)$ convergence, with the standard update rule.

So where's the mistake? It's in our erroneous use of (3.5) in (5.1)—we used the cautious update with $\eta_t = 1/\beta$ to get the first term of $-1/(2\beta)\cdot\|\nabla_t\|_2^2$, but the aggressive update with $\eta_t = (t+1)(2\beta)$ elsewhere. To fix this, how about running *two* processes, one cautious and one aggressive, and then combining them together linearly (with decreasing weight on the aggressive term) to get the new point x_t? This is precisely what Nesterov's Accelerated Method does; let's see it now. (This is another way to arrive at the elegant linear-coupling view that Allen-Zhu and Orecchia present in [1].)

5.2 Getting the acceleration

As we just sketched, one way to view the accelerated gradient descent method is to run two iterative processes for y_t and z_t, and then combine them together to get the actual point x_t. The proof is almost the same as before.

The update steps. Start with $x_0 = y_0 = z_0$. At time t, play x_t. For brevity, define $\nabla_t := \nabla f(x_t)$. Now consider the update rules, where the color is used to emphasize the subtle differences (in particular, z is updated by the gradient at x in (5.3)):

$$y_{t+1} \leftarrow x_t - \tfrac{1}{\beta}\nabla f(x_t), \tag{5.2}$$

$$z_{t+1} \leftarrow z_t - \eta_t \nabla f(x_t), \tag{5.3}$$

$$x_{t+1} \leftarrow (1-\tau_{t+1})y_{t+1} + \tau_{t+1}z_{t+1}. \tag{5.4}$$

In (5.3), we will choose the "aggressive" step size $\eta_t = (t+1)/(2\beta)$ as we did in the above failed attempt. In (5.4) the mixing weight is $\tau_t = 2/(t+2)$, but this will arise organically below.

The potential. This is the same one from the failed attempt:

$$\boxed{\Phi(t) = t(t+1)\cdot(f(y_t) - f(x^*)) + 2\beta\cdot\|z_t - x^*\|^2}. \tag{5.5}$$

The potential change. Define $\Delta\Phi_t = \Phi(t+1) - \Phi(t)$. By the standard GD analysis in (2.7),

$$\frac{1}{2}(\|z_{t+1} - x^*\|^2 - \|z_t - x^*\|^2) = \frac{\eta_t^2}{2}\|\nabla_t\|^2 + \eta_t\langle\nabla_t, x^* - z_t\rangle \tag{5.6}$$

implies that

$$\Delta\Phi_t = t(t+1)\cdot(f(y_{t+1}) - f(y_t)) + 2(t+1)\cdot(f(y_{t+1}) - f(x^*)) + 4\beta\left(\frac{\eta_t^2}{2}\|\nabla_t\|^2 + \eta_t\langle\nabla_t, x^* - z_t\rangle\right).$$

By smoothness and the update rule for y_{t+1}, (3.5) implies

$$f(y_{t+1}) \le f(x_t) - \frac{1}{2\beta}\|\nabla_t\|^2.$$

Substituting, and noting $2\beta\eta_t^2 = (t+1)^2/2\beta \le (t+1)(t+2)/2\beta$ by the choice of η_t, the resulting (negative) squared-norm term can be dropped to give

$$\begin{aligned}\Delta\Phi_t &\le t(t+1)\cdot(f(x_t) - f(y_t)) + 2(t+1)\cdot(f(x_t) - f(x^*)) + 4\beta\eta_t\langle\nabla_t, x^* - z_t\rangle \\ &\le t(t+1)\cdot\langle\nabla_t, x_t - y_t\rangle + 2(t+1)\cdot\langle\nabla_t, x_t - x^*\rangle + 2(t+1)\cdot\langle\nabla_t, x^* - z_t\rangle\end{aligned}$$

using convexity for the first two terms, and $\eta_t = (t+1)/2\beta$ for the last one. Collecting like terms,

$$\Delta\Phi_t \le (t+1)\cdot\langle\nabla_t, (t+2)x_t - ty_t - 2z_t\rangle = 0, \tag{5.7}$$

by using (5.4) and $\tau_t = 2/(t+2)$. Hence $\Phi_t \le \Phi_0$ for all $t \ge 0$. This proves:

Theorem 5.1 (Accelerated GD)**.** *Given a β-smooth function f, the update rules (5.2)-(5.4) ensure*

$$f(y_t) - f(x^*) \le 2\beta\,\frac{\|z_0 - x^*\|^2}{t(t+1)}.$$

5.2.1 An aside: Optimizing parameters and making connections

Suppose we choose the generic potential

$$\Phi(t) = \lambda_{t-1}^2(f(y_t) - f(x^*)) + \frac{\beta}{2}\|z_t - x^*\|^2,$$

where $\lambda_t = \Theta(t^2)$, and try to optimize the calculation above. Having $\lambda_t^2 - \lambda_{t-1}^2 = \lambda_t$ and $\tau_t = 1/\lambda_t$ makes the calculations work out very cleanly. Solving this recurrence leads to the (somewhat exotic-looking) weights

$$\lambda_0 = 0, \qquad \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2} \tag{5.8}$$

used in the standard exposition of AGM2.

The update rules (5.2)–(5.4) have sometimes been called AGM2 (Nesterov's second accelerated method) in the literature. A different set of update rules (called AGM1) are the following: for the optimized choice of λ_t from (5.8), define:

$$\boxed{\begin{aligned} y_{t+1} &\leftarrow x_t - \frac{1}{\beta}\nabla f(x_t)\,, && (5.9)\\ x_{t+1} &\leftarrow \left(1 - \frac{1-\lambda_t}{\lambda_{t+1}}\right) y_{t+1} + \frac{1-\lambda_t}{\lambda_{t+1}} y_t\,. && (5.10)\end{aligned}}$$

Let us show the simple equivalence (also found in, e. g., [27, 9, 16]).

Lemma 5.2. *Using updates (5.9–5.10) and setting $z_t := \lambda_t x_t - (\lambda_t - 1)y_t = \lambda_t(x_t - y_t) + y_t$, and $\tau_t := 1/\lambda_t$ leads to the updates (5.2–5.4).*

Proof. Clearly y_t is the same as above, so it suffices to show that z_t and x_t behave identically. Indeed, rewriting the definition of z_t and substituting $\tau_t = 1/\lambda_t$ gives

$$x_t = (1-\tau_t)y_t + \tau_t z_t\,, \text{ and}$$
$$z_{t+1} - z_t = (\lambda_{t+1}(x_{t+1} - y_{t+1}) + y_{t+1}) - (\lambda_t(x_t - y_t) + y_t)\,. \tag{5.11}$$

Moreover, rewriting (5.10) gives

$$\lambda_{t+1}(x_{t+1} - y_{t+1}) - (1-\lambda_t)(y_t - y_{t+1}) = 0\,. \tag{5.12}$$

Subtracting (5.12) from (5.11) gives

$$z_{t+1} - z_t = \lambda_t y_{t+1} - \lambda_t x_t = -\frac{\lambda_t}{\beta}\nabla f(x_t)\,. \tag{5.13}$$

Recalling that $\lambda_t = 1/\tau_t = (\beta\eta_t)$, this is precisely the update rule $z_{t+1} \leftarrow z_t - \eta_t \nabla f(x_t)$. This shows the equivalence of the two update rules. $\square$

5.3 The constrained case with acceleration

The update steps. The update rule is very similar to the one above, it just involves projecting the points onto the body K. Formally, again we start with $x_0 = y_0 = z_0$. At time t, play x_t. For brevity, define $\nabla_t := \nabla f(x_t)$. Now consider the update rules, where again the color is used to emphasize the subtle differences:

$$\boxed{\begin{aligned} y_{t+1} &\leftarrow \Pi_K(x_t - \tfrac{1}{\beta}\nabla f(x_t))\,, && (5.14)\\ z_{t+1} &\leftarrow \Pi_K(z_t - \eta_t \nabla f(x_t))\,, && (5.15)\\ x_{t+1} &\leftarrow (1-\tau_{t+1})y_{t+1} + \tau_{t+1} z_{t+1}\,. && (5.16)\end{aligned}}$$

We now show that this update rule satisfies same guarantee as in Theorem 5.1 for the unconstrained case.

Theorem 5.3 (Accelerated GD)**.** *Given a β-smooth function f, the update rules (5.14)-(5.16) ensure*

$$f(y_t) - f(x^*) \le 2\beta\,\frac{\|z_0 - x^*\|^2}{t(t+1)}\,.$$

The potential.

$$\boxed{\Phi(t) = t(t+1)\cdot(f(y_t)-f(x^*)) + 2\beta\cdot\|z_t - x^*\|^2}. \tag{5.17}$$

Change in potential.

$$\Phi_{t+1}-\Phi_t = \underbrace{(t+1)(t+2)(f(y_{t+1})-f(x_t))}_{(\star)} \underbrace{-t(t+1)(f(y_t)-f(x_t))+2(t+1)(f(x_t)-f(x^*))}_{(\star\star)}$$
$$+\underbrace{2\beta\cdot\left(\|z_{t+1}-x^*\|^2-\|z_t-x^*\|^2\right)}_{(\star\star\star)}.$$

Using convexity on both differences in $(\star\star)$ gives

$$\begin{aligned}&\le -t(t+1)\langle\nabla_t, y_t-x_t\rangle + 2(t+1)\langle\nabla_t, x_t-x^*\rangle\\ &= (t+1)\cdot\langle\nabla_t, -t(y_t-x_t)+2(x_t-x^*)\rangle\,.\end{aligned}$$

Now $(1-\tau_t)(y_t-x_t)=\tau_t(x_t-z_t)$, and if $\tau_t = 2/(t+2)$, then we get $t(y_t-x_t)=2(x_t-z_t)$, and hence

$$(\star\star) = 2(t+1)\langle\nabla_t, z_t-x^*\rangle\,. \tag{5.18}$$

The expression $(\star\star\star)$ is

$$2\beta(\langle z_{t+1}-z_t, z_{t+1}-x^*\rangle - \|z_{t+1}-z_t\|^2\,.$$

By the Pythagorean property, $\langle z'_{t+1}-z_{t+1}, z_{t+1}-x^*\rangle\ge 0$ where $z'_{t+1}=z_t-\eta_t\nabla_t$. Setting $\eta_t=(t+1)/\beta$ and multiplying by 2β, adding to the term above, $(\star\star\star)$ is upper bounded by

$$-2(t+1)\langle\nabla_t, z_{t+1}-x^*\rangle.$$

Combine with (5.18) to cancel x^*, it suffices to show the following claim:

Lemma 5.4.

$$\overbrace{(t+1)(t+2)(f(y_{t+1})-f(x_t))}^{(\star)}+\overbrace{2(t+1)\langle\nabla_t, z_t-z_{t+1}\rangle-\|z_{t+1}-z_t\|^2}^{(\star\star)+(\star\star\star)}\le 0\,.$$

Proof. By smoothness,

$$f(y_{t+1})-f(x_t)\le\langle\nabla_t, y_{t+1}-x_t\rangle+\frac{\beta}{2}\|y_{t+1}-x_t\|^2\,.$$

As

$$\min_{y\in K}\left\{\langle\nabla_t, y-x_t\rangle+\frac{\beta}{2}\|y-x_t\|^2\right\}$$

in minimized by the definition of y_{t+1}, we get that for any $v\in K$, the RHS is

$$\le\langle\nabla_t, v-x_t\rangle+\frac{\beta}{2}\|v-x_t\|^2\,.$$

Define $v := (1-\tau_t)y_t+\tau_t z_{t+1}\in K$, so $v-x_t=\tau_t(z_{t+1}-z_t)$. Substituting, we get

$$f(y_{t+1})-f(x_t)\le\tau_t\langle\nabla_t, z_{t+1}-z_t\rangle+\frac{\tau_t^2\beta}{2}\|z_{t+1}-z_t\|^2\,.$$

Setting $\tau_t=2/(t+2)$, the claim follows. ☐

5.4 The extension to arbitrary norms

Given an arbitrary norm $\|\cdot\|$, the update rules now use the gradient descent update (4.10) for smooth functions for the y variables, and the mirror descent update rules (4.4) for the z variables:

$$\boxed{\begin{aligned} y_{t+1} &\leftarrow \arg\min_y \left\{ \frac{\beta}{2}\|y-x_t\|^2 + \langle \nabla f(x_t), y-x_t\rangle \right\}, \\ z_{t+1} &\leftarrow \arg\min_z \left\{ \langle \eta_t \nabla f(x_t), z\rangle + D_h(z \parallel z_t) \right\}, \\ x_{t+1} &\leftarrow (1-\tau_{t+1})y_{t+1} + \tau_{t+1} z_{t+1}\,. \end{aligned}} \tag{5.19}$$

Given the discussion in the preceding sections, the update rules are the natural ones: the first is the update (4.10) for smooth functions, and the second is the usual mirror descent update rule (4.5) for the strongly-convex function h. The step size is now set to

$$\eta_t = \frac{(t+1)\alpha_h}{2\beta},$$

where α_h is such that h is α_h-strongly convex with respect to $\|\cdot\|$. Then, the potential function becomes:

$$\boxed{\Phi(t) = t(t+1)\cdot(f(y_t)-f(x^*)) + \frac{4\beta}{\alpha_h}\cdot D_h(x^* \parallel z_t)}, \tag{5.20}$$

which on substituting $D_h(x^* \parallel z_t) = (1/2)\|x^*-z_t\|^2$ and $\alpha_h = 1$ gives (5.5).

We already have all the pieces to bound the change in potential. Use the mirror descent analysis (4.9) to get

$$D_h(x^* \parallel z_{t+1}) - D_h(x^* \parallel z_t) \le \frac{\eta_t^2}{2\alpha_h}\|\nabla_t\|_*^2 + \eta_t\langle \nabla_t, x^*-x_t\rangle.$$

which replaces (5.6). Infer

$$f(y_{t+1}) \le f(x_t) - \frac{1}{2\beta}\|\nabla_t\|_*^2$$

from (4.12) in the smooth case. Substitute these into the analysis from §5.2 (with minor changes for the α_h term) to get the following theorem:

Theorem 5.5 (Accelerated GD: General norms)**.** *Given a β-smooth function f with respect to norm $\|\cdot\|$, the update rules (5.19) ensure*

$$f(y_t) - f(x^*) \le \frac{4\beta}{\alpha_h}\cdot\frac{D_h(x^* \parallel z_0) - D_h(x^* \parallel z_t)}{t(t+1)}.$$

5.5 Acceleration for strongly convex functions

Now consider the case when the function f is well-conditioned with condition number $\kappa = \beta/\alpha$, and describe the algorithm of Nesterov with convergence rate $\exp(-t/\sqrt{\kappa})$ [20]. Again, this is the best possible for gradient methods.

Before describing the algorithm and its analysis, it is instructive to see how the result for the smoothed case from Section 5.2 already gives (in principle) such a result for the well-conditioned case.

Reduction from smooth case. Theorems 5.1 and 5.3 for the (constrained) smoothed case give that

$$f(y_t) - f(x^*) \leq 2\beta \frac{\|x_0 - x^*|^2}{t(t+1)} .$$

Together with strong convexity,

$$f(y_t) - f(x^*) \geq \frac{\alpha}{2}\|y_t - x^*\|^2 ,$$

this gives $\|y_t - x^*\|^2 \leq 4\kappa\|x_0 - x^*\|^2/t(t+1)$.

So in $t = 4\sqrt{\kappa}$ steps, the distance $\|y_t - x^*\|$ is at most half the initial distance $\|x_0 - x^*\|$ from the optimum. Starting the algorithm again with y_t as the initial point, and iterating this process thus gives an overall algorithm, that in t steps has error at most $2^{-t/4\sqrt{\kappa}}\|x - x_0\|$.

Restarting the algorithm after every few steps is not ideal, and we now describe Nesterov's algorithm with this improved convergence rate. For simplicity only consider the unconstrained case.

The update rules. We now use the following updates (which look very much like the `AGM1` updates):

$$y_{t+1} \leftarrow x_t - \frac{1}{\beta}\nabla f(x_t) , \tag{5.21}$$

$$x_{t+1} \leftarrow \left(1 + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) y_{t+1} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} y_t . \tag{5.22}$$

For the analysis, it will be convenient to define $\tau = 1/(\sqrt{\kappa}+1)$ and set

$$z_{t+1} := \tfrac{1}{\tau}\, x_{t+1} - \tfrac{1-\tau}{\tau}\, y_{t+1}. \tag{5.23}$$

We now show that that the error after t steps is

$$f(y_t) - f(x^*) \leq (1+\gamma)^{-t} \left(\frac{\alpha+\beta}{2}\, \|x_0 - x^*\|^2\right) , \tag{5.24}$$

where $\gamma = 1/(\sqrt{\kappa}-1)$ (as in §3.2, for $\kappa = 1$ the algorithm reaches optimum in a single step and $y_1 = x^*$, and hence we assume that $\kappa > 1$). This improves on the error of

$$(1+1/\kappa)^{-t}\frac{\beta}{2}\|x_0 - x^*\|^2$$

we get from §3.2.

The potential. Consider the potential

$$\boxed{\Phi(t) = (1+\gamma)^t \left(f(y_t) - f(x^*) + \frac{\alpha}{2}\|z_t - x^*\|^2\right) .}$$

Observe that

$$\Phi_0 = f(y_0) - f(x^*) + \frac{\alpha}{2}\|z_0 - x^*\|^2 .$$

As $x_0 = y_0 = z_0$, and by β-smoothness of f,

$$\Phi_0 \leq \frac{\alpha+\beta}{2}\, \|x_0 - x^*\|^2.$$

Change in potential. To show the error bound (5.24), it suffices to show that $\Delta\Phi(t) = \Phi(t+1) - \Phi(t) \leq 0$ for each t. This is equivalent to showing

$$(1+\gamma)(f(y_{t+1}) - f(x^*)) - (f(y_t) - f(x^*)) + \frac{\alpha}{2}\left((1+\gamma)\|z_{t+1} - x^*\|^2 - \|z_t - x^*\|^2\right) \leq 0\,.$$

We first bound the terms involving f in the most obvious way. As above, we use ∇_t as short-hand for $\nabla f(x_t)$. By β-smoothness and the update rule, again

$$f(y_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla_t\|^2\,.$$

So,

$$\begin{aligned}
&(1+\gamma)(f(y_{t+1}) - f(x^*)) - (f(y_t) - f(x^*)) \\
&\qquad \leq f(x_t) - f(y_t) + \gamma(f(x_t) - f(x^*)) - (1+\gamma)\frac{1}{2\beta}\|\nabla_t\|^2 \\
&\qquad \leq \langle \nabla_t, x_t - y_t\rangle + \gamma\left(\langle \nabla_t, x_t - x^*\rangle - \frac{\alpha}{2}\|x_t - x^*\|^2\right) - \frac{1+\gamma}{2\beta}\|\nabla_t\|^2\,,
\end{aligned} \tag{5.25}$$

where the last inequality used convexity and strong convexity respectively.

We now want to remove references to y_t. By definition (5.23),

$$z_t = \left(\frac{1}{\tau} - 1\right)(x_t - y_t) + x_t = \sqrt{\kappa}(x_t - y_t) + x_t\,,$$

so we infer $\gamma(z_t - x^*) = \sqrt{\kappa}\gamma(x_t - y_t) + \gamma(x_t - x^*)$. Using $\sqrt{\kappa}\gamma = 1+\gamma$, simple algebra gives

$$(x_t - y_t) + \gamma(x_t - x^*) = \frac{1}{1+\gamma}\left(\gamma(z_t - x^*) + \gamma^2(x_t - x^*)\right).$$

For brevity we use $X_t := x_t - x^*$, $Z_t = z_t - x^*$, and substitute the above expression into (5.25) to get

$$\frac{1}{1+\gamma}\langle \nabla_t, \gamma Z_t + \gamma^2 X_t\rangle - \frac{\alpha\gamma}{2}\|X_t\|^2 - \frac{1+\gamma}{2\beta}\|\nabla_t\|^2\,. \tag{5.26}$$

Now, let us upper bound the terms in $\Delta\Phi(t)$ involving z. Conveniently, we can relate z_{t+1} and z_t using a simple calculation that we defer for the moment.

Claim 5.6. *$z_{t+1} = (1 - \frac{1}{\sqrt{\kappa}})z_t + \frac{1}{\sqrt{\kappa}}x_t - \frac{1}{\alpha\sqrt{\kappa}}\nabla_t$ and so $z_{t+1} - x^* = \frac{1}{1+\gamma}Z_t + \frac{\gamma}{1+\gamma}X_t - \frac{\gamma}{\alpha(1+\gamma)}\nabla_t$.*

Now use Claim 5.6 and expand using $\|a+b+c\|^2 = \|a\|^2 + \|b\|^2 + \|c\|^2 + 2\langle a,b\rangle + 2\langle b,c\rangle + 2\langle a,c\rangle$:

$$\begin{aligned}
&(1+\gamma)\|z_{t+1} - x^*\|^2 - \|z_t - x^*\|^2 \\
&= \frac{1}{1+\gamma}\left(\|Z_t\|^2 + \gamma^2\|X_t\|^2 + \frac{\gamma^2}{\alpha^2}\|\nabla_t\|^2 + 2\gamma\langle Z_t, X_t\rangle - \frac{2\gamma}{\alpha}\langle \nabla_t, Z_t\rangle - \frac{2\gamma^2}{\alpha}\langle \nabla_t, X_t\rangle\right) - \|Z_t\|^2.
\end{aligned} \tag{5.27}$$

Now sum (5.26) and $\alpha/2$ times (5.27). The terms involving $\|\nabla_t\|^2$ cancel since

$$\frac{1+\gamma}{2\beta} = \frac{\alpha\gamma^2}{2\alpha^2(1+\gamma)}$$

(by the definition of γ). Moreover, the inner-product terms involving ∇_t also cancel. Hence the potential change is at most

$$\begin{aligned}\Delta\Phi(t) &\le \frac{\alpha\gamma}{2}\|X_t\|^2\left(-1+\frac{\gamma}{1+\gamma}\right)+\frac{\alpha}{2}\|Z_t\|^2\left(\frac{1}{1+\gamma}-1\right)+\frac{\alpha\gamma}{1+\gamma}\langle Z_t,X_t\rangle \\ &= -\frac{\alpha\gamma}{2(1+\gamma)}\left(\|X_t\|^2+\|Z_t\|^2-2\langle Z_t,X_t\rangle\right) && (5.28)\\ &= -\frac{\alpha\gamma}{2(1+\gamma)}\|Z_t-X_t\|^2 \le 0\,. && (5.29)\end{aligned}$$

Hence the potential does not increase, as claimed. It only remains to prove Claim 5.6.

Proof of Claim 5.6. The expression of x_{t+1} from (5.22) can be written as $(2-2\tau)y_{t+1}-(1-2\tau)y_t$. Plugging into the expression for z_{t+1} from (5.23) gives

$$\begin{aligned}z_{t+1} &= \frac{1}{\tau}\left((2-2\tau)y_{t+1}-(1-2\tau)y_t-(1-\tau)y_{t+1}\right)\\ &= \frac{1}{\tau}\left((1-\tau)y_{t+1}-(1-2\tau)y_t\right).\end{aligned}$$

Using the update rule (5.21) for y_{t+1}, and the relation $x_t=(1-\tau)y_t+\tau z_t$ to eliminate y_t

$$\begin{aligned}&= \frac{1}{\tau}\left((1-\tau)\left(x_t-\frac{1}{\beta}\nabla_t\right)-\frac{(1-2\tau)}{1-\tau}(x_t-\tau z_t)\right)\\ &= \frac{1-2\tau}{1-\tau}z_t+\frac{\tau}{1-\tau}x_t-\frac{1-\tau}{\tau\beta}\nabla_t\,.\end{aligned}$$

Using $\tau = 1/(\sqrt{\kappa}+1)$ and $\beta=\kappa\alpha$ now gives the claim. □

Acknowledgments

We thank the referees and Laci Babai for their detailed and thoughtful feedback that helped improve the presentation substantially. We also thank Sébastien Bubeck, Daniel Dadush, Jelena Diakonikolas, Nick Harvey, Elad Hazan, Greg Koumoutsos, Raghu Meka, Aryan Mokhtari, Marco Molinaro, Thomas Rothvoß, and Kunal Talwar for their helpful comments.

References

[1] ZEYUAN ALLEN-ZHU AND LORENZO ORECCHIA: Linear coupling: an ultimate unification of gradient and mirror descent. In *Proc. 8th Innovations in Theoret. Comp. Sci. Conf. (ITCS'17)*, pp. 3:1–3:22. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. [doi:10.4230/LIPIcs.ITCS.2017.3, arXiv:1407.1537] 22

[2] AMIR BECK AND MARC TEBOULLE: Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003. [doi:10.1016/S0167-6377(02)00231-6] 16

[3] AHARON BEN-TAL AND ARKADI NEMIROVSKI: *Lectures on Modern Convex Optimization*. Society for Industrial and Applied Mathematics, 2001. [doi:10.1137/1.9780898718829] 3

[4] DMITRI P. BERTSEKAS, ANGELIA NEDIĆ, AND ASUMAN E. OZDAGLAR: *Convex Analysis and Optimization*. Athena Scientific, 2003. 16

[5] STEPHEN BOYD AND LIEVEN VANDENBERGHE: *Convex Optimization*. Cambridge University Press, 2004. [doi:10.1017/CBO9780511804441] 20

[6] SÉBASTIEN BUBECK: Convex optimization: algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. [doi:10.1561/2200000050, arXiv:1405.4980] 3, 6

[7] NICOLÒ CESA-BIANCHI AND GÁBOR LUGOSI: *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006. 3, 9

[8] JELENA DIAKONIKOLAS AND LORENZO ORECCHIA: The approximate gap technique: a unified approach to optimal first-order methods. *SIAM J. Optim.*, 29(1):660–689, 2019. SIAM. [arXiv:1712.02485] 2, 3, 21

[9] YOEL DRORI AND MARC TEBOULLE: Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1-2):451–482, 2014. [doi:10.1007/s10107-013-0653-0, arXiv:1206.3209] 24

[10] JOHN C. DUCHI: Introductory Lectures on Stochastic Optimization, 2016. Available at author's webpage. 3

[11] MARGUERITE FRANK AND PHILIP WOLFE: An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. [doi:10.1002/nav.3800030109] 13

[12] ELAD HAZAN: Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016. [doi:10.1561/2400000013] 3

[13] JEAN-BAPTISTE HIRIART-URRUTY AND CLAUDE LEMARÉCHAL: *Fundamentals of Convex Analysis*. Springer, 2001. [doi:10.1007/978-3-642-56468-0] 2

[14] MARTIN JAGGI: Revisiting Frank–Wolfe: projection-free sparse convex optimization. In *Proc. 13th Internat. Conf. on Machine Learning (ICML'13)*, pp. 427–435, 2013. [PMLR]. 21

[15] SAHAR KARIMI AND STEPHEN A. VAVASIS: A single potential governing convergence of conjugate gradient, accelerated gradient and geometric descent, 2017. [arXiv:1712.09498] 3

[16] DONGHWAN KIM AND JEFFREY A. FESSLER: Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159(1-2):81–107, 2016. [doi:10.1007/s10107-015-0949-3, arXiv:1406.5468] 24

[17] WALID KRICHENE, ALEXANDRE M. BAYEN, AND PETER L. BARTLETT: Accelerated mirror descent in continuous and discrete time. In *Adv. in Neural Inform. Processing Systems 28 (NIPS'15)*, pp. 2845–2853, 2015. NIPS. 2, 3, 21

[18] ARKADI SEMËNOVICH NEMIROVSKI AND DAVID BERKOVICH YUDIN: *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, Inc., 1983. 2, 3, 16

[19] YURII NESTEROV: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983. LINK at Math-Net.ru. 21

[20] YURII NESTEROV: *Introductory Lectures on Convex Optimization – A Basic Course*. Kluwer Academic Publishers, 2004. [doi:10.1007/978-1-4419-8853-9] 3, 6, 26

[21] JAVIER PEÑA: Convergence of first-order methods via the convex conjugate. *Oper. Res. Lett.*, 45(6):561–564, 2017. [doi:10.1016/j.orl.2017.08.013, arXiv:1707.09084] 3

[22] SHAI SHALEV-SHWARTZ: Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. [doi:10.1561/2200000018] 3

[23] NAUM ZUSELEVICH SHOR: *Minimization Methods for Non-differentiable Functions*. Springer, 1985. [doi:10.1007/978-3-642-82118-9] 3

[24] WEIJIE SU, STEPHEN BOYD, AND EMMANUEL J. CANDÈS: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res. (JMLR)*, 17(153):1–43, 2016. 2, 3, 21

[25] EIJI TAKIMOTO AND MANFRED K. WARMUTH: The minimax strategy for Gaussian density estimation. In *Proc. 13th Ann. Conf. on Computational Learning Theory (COLT'00)*, pp. 100–106. Morgan Kaufmann Publ., 2000. ACM DL. 8

[26] ADRIEN TAYLOR AND FRANCIS BACH: Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Proc. 32th Ann. Conf. on Computational Learning Theory (COLT'19)*, pp. 2934–2992. MLR Press, 2019. [PMLR]. [arXiv:1902.00947] 3

[27] PAUL TSENG: On accelerated proximal gradient methods for convex-concave optimization, 2008. Available at author's webpage. 24

[28] NISHEETH K. VISHNOI: A mini-course on convex optimization, 2018. Available at author's webpage. 3

[29] ANDRE WIBISONO, ASHIA C. WILSON, AND MICHAEL I. JORDAN: A variational perspective on accelerated methods in optimization. *Proc. Natl. Acad. Sci. USA*, 113(47):E7351–E7358, 2016. [doi:10.1073/pnas.1614734113, arXiv:1603.04245] 2, 3, 21

[30] ASHIA C. WILSON, BEN RECHT, AND MICHAEL I. JORDAN: A Lyapunov analysis of momentum methods in optimization, 2016. [arXiv:1611.02635] 2, 3, 21

AUTHORS

Nikhil Bansal
Researcher
Centrum voor Wiskunde en Informatica
Amsterdam
The Netherlands
bansal@gmail.com
http://www.win.tue.nl/~nikhil

Anupam Gupta
Professor
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA
anupamg@cs.cmu.edu
http://www.cs.cmu.edu/~anupamg

ABOUT THE AUTHORS

NIKHIL BANSAL is a researcher at the Centrum voor Wiskunde en Informatica, Amsterdam. He attended the Indian Institute of Technology, Mumbai for his B. Tech. degree, and received his Ph. D. from Carnegie Mellon University, where he was advised by Avrim Blum. He got fascinated by algorithms after taking an undergraduate class by Prof. Ajit A. Diwan. Since then he has enjoyed thinking about various kinds of algorithmic questions.

ANUPAM GUPTA is a faculty member in the Computer Science Department at Carnegie Mellon University. He grew up in Calcutta, India, attended the Indian Institute of Technology, Kanpur for his B. Tech. degree, and received his Ph. D. from the University of California, Berkeley, under the guidance of Alistair Sinclair. Anupam is interested in most kinds of algorithmic problems, and his research focus is on approximation and online algorithms, and metric embeddings.