MIT 6.7220/15.084 — Nonlinear Optimization Tue, Mar 12–14th 2024

Lecture 9A-B
Projected gradient descent and mirror descent

Instructor: Prof. Gabriele Farina (✉ gfarina@mit.edu)*

We continue our exploration of first-order methods by considering the case of *constrained* optimization problems of the form

$$\begin{aligned} \min_{x} \;& f(x) \\ \text{s.t.} \;& x \in \Omega \subseteq \mathbb{R}^n, \end{aligned}$$

where f is differentiable and Ω is a closed and convex set (with only one exception that we will highlight later).

1 Projected gradient descent

When applied without modifications to a constrained optimization problem, the gradient descent algorithm might quickly produce iterates x_t that leave the feasible set. The idea behind *projected* gradient descent is very intuitive: to avoid the issue of infeasible iterates, project the iterates of gradient descent back into Ω at every iteration. This leads to the update rule

$$\boxed{x_{t+1} \coloneqq \Pi_\Omega\Big(x_t - \eta \nabla f(x_t)\Big)}, \tag{1}$$

where the operation Π_Ω denotes Euclidean projection onto Ω. (Remember that projection onto a closed convex set exists and is unique.)

As it turns out, the projected gradient descent algorithm behaves fundamentally like the gradient descent algorithm. In particular, the gradient descent lemma and the Euclidean mirror descent lemma can be generalized to projected gradient descent with little effort. As a corollary, the same convergence rate of $f(x_t) - f(x_\star) \le \frac{1}{\eta t}\|x_0 - x_\star\|_2^2$ for L-smooth functions when $0 < \eta \le \frac{1}{L}$ applies to projected gradient descent as well.

Instead of developing the results for projected gradient descent, we introduce a generalization of the algorithm that is more permissive to the projection notion used. The correctness of the projected gradient descent will then follow as a direct corollary of this generalization.

2 Generalized projection: proximal steps and mirror descent

Depending on the feasible set Ω, *Euclidean* projections onto Ω might be expensive to compute in practice. A generalization of projected gradient descent called *mirror descent* affords more flexibility in the notion of distance used in the projection.

*These notes are class material that has not undergone formal peer review. The TAs and I are grateful for any reports of typos.

2.1 Distance-generating functions (DGFs)

An interesting generalization of the distance between two points is that of a *Bregman divergence* (also called Bregman distance). A Bregman divergence is *not* a distance in the technical sense—for example, it is not necessarily symmetric, and it need not satisfy the triangle inequality.

A Bregman divergence is constructed starting from a *strongly convex* function φ, called the *distance-generating function (DGF)* for the divergence.

Definition 2.1. Let $\varphi : \Omega \to \mathbb{R}$ be a differentiable and μ-strongly convex ($\mu > 0$) function with respect to a norm $\|\cdot\|$, that is, satisfy

$$\varphi(x) \geq \varphi(y) + \langle \nabla\varphi(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \qquad \forall x, y \in \Omega.$$

The Bregman divergence *centered in* $y \in \Omega$ is the function $\mathrm{D}_\varphi(x \,\|\, y)$ defined as

$$\mathrm{D}_\varphi(x \,\|\, y) \coloneqq \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), x - y\rangle.$$

Note that from its very definition it is clear that

$$\mathrm{D}_\varphi(x \,\|\, x) = 0 \quad \forall x \in \Omega \qquad \text{and} \qquad \mathrm{D}_\varphi(x \,\|\, y) \geq \frac{\mu}{2}\|x - y\|^2 \quad \forall x, y \in \Omega, \tag{2}$$

which in particular implies that $\mathrm{D}_\varphi(x \,\|\, y) = 0$ if and only if $x = y$. We now mention two very important special cases of Bregman divergences.

- When Ω is arbitrary and the DGF φ is set to be the is the squared Euclidean norm

  $$\varphi(x) \coloneqq \frac{1}{2}\|x\|_2^2,$$

  which is 1-strongly convex with respect to $\|\cdot\|_2$, the corresponding Bregman divergence is the squared Euclidean distance

  $$\mathrm{D}_\varphi(x \,\|\, y) = \frac{1}{2}\|x - y\|_2^2.$$

  Indeed, from the definition,

  $$\mathrm{D}_\varphi(x \,\|\, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y\rangle = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x\rangle + \|y\|_2^2 = \frac{1}{2}\|x - y\|_2^2.$$

- When $\Omega = \mathring{\Delta}^n$ is the set of full-support distributions over n objects,[1] and the distance-generating function φ is set to the negative entropy function

  $$\varphi(x) \coloneqq \sum_{i=1}^n x_i \log x_i,$$

  which is 1-strongly convex with respect to the ℓ_1 norm $\|\cdot\|_1$, [▷ You should check this!] the corresponding Bregman divergence is the *Kullback-Leibler (KL) divergence* [▷ And this too!]

  $$\mathrm{D}_\varphi(x \,\|\, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i},$$

  a commonly used notion of distance between distributions in machine learning and statistics.

[1] In this case, the set Ω is not closed, so the existence of the proximal step does not follow quite as directly. However, we can still show it using elementary arguments; we already encountered this in Lecture 1 (see Piazza for a hint).

A useful fact about Bregman divergences is that for any center y, they are as strongly convex in x as the original distance-generating function φ. More precisely, we have the following.

Theorem 2.1. Let $\varphi : \Omega \to \mathbb{R}$ be differentiable and μ-strongly convex with respect to a norm $\|\cdot\|$. For any $y \in \Omega$, the function $x \mapsto \mathrm{D}_\varphi(x \,\|\, y)$ is μ-strongly convex with respect to $\|\cdot\|$, that is,

$$\mathrm{D}_\varphi(x' \,\|\, y) \geq \mathrm{D}_\varphi(x \,\|\, y) + \left\langle \nabla_x \mathrm{D}_\varphi(x \,\|\, y), x' - x \right\rangle + \frac{\mu}{2}\|x' - x\|^2 \qquad \forall x, x' \in \Omega.$$

Proof. Using the definition of the Bregman divergence $\mathrm{D}_\varphi(\cdot \,\|\, \cdot)$, we have

$$\nabla_x \mathrm{D}_\varphi(x \,\|\, y) = \nabla\varphi(x) - \nabla\varphi(y),$$

so after expanding the definition of $\mathrm{D}_\varphi(\cdot \,\|\, \cdot)$ in the inequality of the statement, the statement is

$$\varphi(x') \geq \varphi(x) + \langle \nabla\varphi(y), x' - x\rangle + \langle \nabla\varphi(x) - \nabla\varphi(y), x' - x\rangle + \frac{\mu}{2}\|x' - x\|^2 \qquad \forall x, x' \in \Omega,$$

that is, $\varphi(x') \geq \varphi(x) + \langle \nabla\varphi(x), x' - x\rangle + \frac{\mu}{2}\|x' - x\|^2$ for all $x, x' \in \Omega$, which follows by the assumption of μ-strong convexity with respect to $\|\cdot\|$ for φ. □

2.2 Proximal steps

Proximal steps generalize the steps followed by the projected gradient descent algorithm (1). The key insight is the following: instead of interpreting (1) as the projection of the point $x_t - \eta\nabla f(x_t)$, we can interpret x_{t+1} as just another manifestation of the key principle of gradient descent-type algorithms: hoping that the objective can be approximated well with its first-order Taylor expansion in a neighborhood of each point. It then follows naturally that each updated point x_{t+1} produced by a gradient descent-type algorithm should trade off two competing objectives:

- moving as much as possible in the direction $-\nabla f(x_t)$; and
- staying in a neighborhood of Ω centered around point x_t, so as to not move too far.

The stepsize parameter $\eta > 0$ controls the tradeoff between the competing objectives.

When using a generic Bregman divergence $\mathrm{D}_\varphi(\cdot \,\|\, \cdot)$ as the notion of distance, the tradeoff between these two competing objectives can be formalized as the *proximal step* problem

$$\mathrm{Prox}_\varphi(\eta\nabla f(x_t), x_t) \coloneqq \arg\min_{\substack{x \\ \text{s.t. } x\in\Omega.}} \eta\langle \nabla f(x_t), x\rangle + \mathrm{D}_\varphi(x \,\|\, x_t)$$

We show in the next subsection that proximal steps are well-defined—that is, the solution to the optimization problem above exists and is unique. This leads to the *mirror descent* algorithm, defined by the update

$$\boxed{x_{t+1} \coloneqq \mathrm{Prox}_\varphi(\eta\nabla f(x_t), x_t)}\,. \tag{3}$$

■ **The Euclidean DGF recovers Euclidean projection.** As a sanity check to convince ourselves that the abstraction of proximal step is reasonable, we can verify that it generalizes the steps of projected gradient descent (1)—and therefore also of gradient descent, which is just projected gradient descent in which $\Omega = \mathbb{R}^n$. We do so in the next theorem.

Theorem 2.2. Consider the squared Euclidean norm distance-generating function $\varphi(x) = \frac{1}{2}\|x\|_2^2$. Then, proximal steps and projected gradient steps (1) are equivalent:

$$\mathrm{Prox}_{\varphi}(\eta\nabla f(x), x) = \Pi_{\Omega}\Big(x - \eta\nabla f(x)\Big) \qquad \forall x \in \Omega.$$

Proof. The Euclidean projection problem is given by

$$\begin{aligned} \min_{y} \ & \tfrac{1}{2}\|y - x + \eta\nabla f(x)\|_2^2 \\ \text{s.t.} \ & y \in \Omega. \end{aligned}$$

Expanding the squared Euclidean norm in the objective and removing terms that do not depend on the optimization variable y we can rewrite the problem as

$$\begin{aligned} \min_{y} \ & \tfrac{1}{2}\|y - x\|_2^2 + \eta\langle\nabla f(x), y - x\rangle \\ \text{s.t.} \ & y \in \Omega, \end{aligned}$$

which is exactly the proximal step problem since $\mathrm{D}_{\varphi}(y \,\|\, x) = \frac{1}{2}\|y - x\|_2^2$ as observed in the previous section. □

■ **The negative entropy DGF recovers the softmax update.** Proximal steps are very useful when computing Euclidean projections is expensive. For example, in the case of the negative entropy distance-generating function for full-support distributions, we can use the result in Lecture 2 to show that the proximal step corresponds to the *softmax* update

$$x_{t+1} \propto x_t \odot \exp\{-\eta\nabla f(x_t)\},$$

where $\odot$ denotes elementwise product. [▷ Try to work out the details!] Such a generalized notion of projection is significantly more practical than the algorithm for Euclidean projection you developed in Homework 1.

2.3 Properties of proximal steps

We now mention a few important properties of proximal steps.

■ **Proximal steps exist.** The argument is analogous to the one we used in Lecture 1 to argue the existence of Euclidean projections. More formally, consider a generic proximal step $\mathrm{Prox}_{\varphi}(g, y)$, in which the objective function to be minimized over $x \in \Omega$ is accordingly

$$h(x) \coloneqq \langle g, x\rangle + \mathrm{D}_{\varphi}(x \,\|\, y).$$

The infimum of the function over Ω must be less than or equal to the value of the objective in the valid choice $x = y$.

To apply the Weierstrass theorem, we need to show that we can safely restrict the domain to a compact subset. To do so, we can use the knowledge, from (2), that $\mathrm{D}_{\varphi}(x \,\|\, y) \geq \frac{\mu}{2}\|x - y\|^2$ for all $x \in \Omega$, where $\mu > 0$ and $\|\cdot\|$ are the strong convexity parameter and strong convexity norm of the underlying DGF φ. The value of the increment $h(x) - h(y)$ can therefore be lower-bounded using the generalized Cauchy-Schwarz inequality as

$$\begin{aligned} h(x) - h(y) &\geq \langle g, x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \\ &\geq \frac{\mu}{2}\|x - y\| \cdot \Big(\|x - y\| - \|g\|_*\Big). \end{aligned}$$

Hence, any point x such that $\|x - y\| \geq \|g\|_*$, we must have that $h(x) \geq h(y)$. Therefore, we can restrict the minimization of $h(x)$ to the compact set defined by the intersection between Ω and the closed

ball of radius $\|g\|_*$ centered in y, and Weierstrass guarantees the existence of a minimizer of f on this compact restriction of the domain.

■ **Proximal steps are unique.** The objective function minimized in proximal steps is defined as the sum of a linear function plus a Bregman divergence with a fixed center. Since Bregman divergences are strongly convex by Theorem 2.1, and linear terms do not affect strong convexity, the proximal step problem minimizes a strongly convex objective on a convex set. The uniqueness of the solution is therefore guaranteed (see also Homework 1).

■ **The three-point equality for proximal steps.** The following property is key in many proofs involving proximal steps. For that reason, we give it a name.

Theorem 2.3 (Three-point inequality for proximal steps). Consider a generic proximal set

$$x' = \operatorname{Prox}_\varphi(g, x).$$

Then,

$$\langle -g, y - x' \rangle \leq -\mathrm{D}_\varphi(y \,\|\, x') + \mathrm{D}_\varphi(y \,\|\, x) - \mathrm{D}_\varphi(x' \,\|\, x) \qquad \forall y \in \Omega.$$

Proof. The objective function of the proximal step problem is given by

$$h(z) \coloneqq \langle g, z \rangle + \mathrm{D}_\varphi(z \,\|\, x), \quad z \in \Omega.$$

The first-order optimality conditions applied to the solution $z = x'$ are therefore

$$\begin{aligned} -\nabla h(x') \in \mathcal{N}_\Omega(x') \quad &\iff \quad -g - \nabla\varphi(x') + \nabla\varphi(x) \in \mathcal{N}_\Omega(x') \\ &\iff \quad \langle -g - \nabla\varphi(x') + \nabla\varphi(x), y - x' \rangle \leq 0 \qquad \forall y \in \Omega \\ &\iff \quad \langle -g, y - x' \rangle \leq \langle \nabla\varphi(x') - \nabla\varphi(x), y - x' \rangle \quad \forall y \in \Omega. \end{aligned}$$

The statement now follows from using the identity

$$\langle \nabla\varphi(x') - \nabla\varphi(x), y - x' \rangle = -\mathrm{D}_\varphi(y \,\|\, x') + \mathrm{D}_\varphi(y \,\|\, x) - \mathrm{D}_\varphi(x' \,\|\, x),$$

which can be checked directly from the definition of Bregman divergence. [▷ Verify this!] □

Corollary 2.1. An important corollary of the three-point inequality is obtained by setting $y = x$. In that case, the three-point inequality simplifies to

$$\langle -g, x - x' \rangle \leq -\mathrm{D}_\varphi(x \,\|\, x') - \mathrm{D}_\varphi(x' \,\|\, x).$$

Corollary 2.2. Continuing Corollary 2.1 by using the strong convexity bound (see Theorem 2.1) $\mathrm{D}_\varphi(x \,\|\, x') + \mathrm{D}_\varphi(x' \,\|\, x) \geq \mu\|x - x'\|^2$ to bound the right-hand size, and the generalized Cauchy-Schwarz inequality to bound the left-hand side, we find a bound on the norm of the proximal step:

$$\|x' - x\| \leq \frac{1}{\mu}\|g\|_*.$$

3 Analysis of mirror descent

As we have discussed in Lectures 7 and 8, the analysis of gradient descent (and its many variants and generalizations) typically goes through two fundamental—and conceptually complementary—lemmas:

- the gradient descent lemma, stating that

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2$$

  when f is L-smooth and $0 < \eta \leq \frac{1}{L}$; and
- the Euclidean mirror descent lemma, which states that

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\Big(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\Big) \qquad \forall y \in \mathbb{R}^n$$

  for convex f and arbitrary stepsize $\eta > 0$.

We now show that with little effort, we can generalize those results to the case of the mirror descent algorithm.

3.1 Generalizing the gradient descent lemma

We start by generalizing the gradient descent lemma.

Theorem 3.1. Let $f : \Omega \to \mathbb{R}$ be L-smooth with respect to the norm $\|\cdot\|$ for which φ is strongly convex, and $0 < \eta \leq \frac{\mu}{L}$. Each step of the mirror descent algorithm (3) satisfies

$$f(x_{t+1}) \leq f(x_t) - \frac{\mu}{2\eta}\|x_t - x_{t+1}\|^2.$$

Proof. From the quadratic upper bound, we have

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_t - x_{t+1}\|^2$$

Using Corollary 2.1 we therefore find

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{1}{\eta}\mathrm{D}_\varphi(x_{t+1} \,\|\, x_t) - \frac{1}{\eta}\mathrm{D}_\varphi(x_t \,\|\, x_{t+1}) + \frac{L}{2}\|x_t - x_{t+1}\|^2 \\ &\leq f(x_t) + \left(\frac{L}{2} - \frac{\mu}{\eta}\right)\|x_t - x_{t+1}\|^2 \\ &\leq f(x_t) - \frac{\mu}{2\eta}\|x_t - x_{t+1}\|^2, \end{aligned}$$

which is the statement. □

As expected, we find that the decrease in function value is monotonic, just like in the unconstrained case.

3.2 The "full" mirror descent lemma

We continue by generaling the Euclidean mirror descent lemma to its fully general version for arbitrary Bregman divergences. In particular, from Theorem 2.3 we have the following.

Theorem 3.2. Let $f : \Omega \to \mathbb{R}$ be convex. Each step of the mirror descent algorithm (3) satisfies

$$f(x_t) \leq f(y) + \langle \nabla f(x_t), x_t - x_{t+1} \rangle - \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_{t+1}) + \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_t) - \frac{1}{\eta}\mathrm{D}_\varphi(x_{t+1} \,\|\, x_t).$$

Proof. Using the linear lower bound property of convex functions (Lecture 3), we can write

$$\begin{aligned} f(x_t) &\leq f(y) - \langle \nabla f(x_t), y - x_t \rangle \\ &= f(y) + \langle \nabla f(x_t), x_t - x_{t+1} \rangle - \langle \nabla f(x_t), y - x_{t+1} \rangle. \end{aligned}$$

On the other hand, from Theorem 2.3 applied to the mirror descent step (that is, for the choices $g = \eta \nabla f(x_t), x' = x_{t+1}, x = x_t$), we have

$$-\eta \langle \nabla f(x_t), y - x_{t+1} \rangle \leq -\mathrm{D}_\varphi(y \,\|\, x_{t+1}) + \mathrm{D}_\varphi(y \,\|\, x_t) - \mathrm{D}_\varphi(x_{t+1} \,\|\, x_t).$$

Hence, dividing by η and plugging into the previous inequality, we obtain the statement. □

Observe that when φ is the square Euclidean norm, we recover exactly the Euclidean mirror descent lemma in the unconstrained case, upon substituting $\nabla f(x_t) = \frac{1}{\eta}(x_t - x_{t+1})$.

3.3 Convergence guarantees for L-smooth functions

If the function is convex and L-smooth with respect to the norm $\|\cdot\|$ for which φ is strongly convex, and $0 < \eta \leq \frac{\mu}{L}$, we can substitute the quadratic upper bound

$$\langle \nabla f(x_t), x_t - x_{t+1} \rangle \leq f(x_t) - f(x_{t+1}) + \frac{L}{2} \|x_t - x_{t+1}\|^2$$

into the mirror descent lemma (Theorem 3.2), obtaining

$$\begin{aligned} f(x_{t+1}) &\leq f(y) - \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_{t+1}) + \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_t) - \frac{1}{\eta}\mathrm{D}_\varphi(x_{t+1} \,\|\, x_t) + \frac{L}{2}\|x_t - x_{t+1}\|^2 \\ &\leq f(y) - \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_{t+1}) + \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_t) - \frac{\mu}{2\eta}\|x_{t+1} - x_t\|_2^2 + \frac{L}{2}\|x_t - x_{t+1}\|^2 \\ &\leq f(y) - \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_{t+1}) + \frac{1}{\eta}\mathrm{D}_\varphi(y \,\|\, x_t). \end{aligned}$$

Following the same steps as Lecture 7, telescoping and using the monotonicity of $f(x_t)$ proved in Theorem 3.2 we obtain the following guarantee.

Theorem 3.3. Let $f : \Omega \to \mathbb{R}$ be convex and L-smooth with respect to the norm $\|\cdot\|$ for which φ is strongly convex. Furthemore, let $0 < \eta \leq \frac{\mu}{L}$, and $x_\star \in \Omega$ be a minimizer of the function. Then, at any t, the iterate x_t produced by the mirror descent algorithm satisfies

$$f(x_t) - f(x_\star) \leq \frac{\mathrm{D}_\varphi(x_\star \,\|\, x_0)}{\eta t}.$$

4 Further readings

More in-detail treatments of the mirror descent algorithm can be found in several standard resources, including the nice monograph by Bubeck, S. [Bub15].

[Bub15] S. Bubeck, "Convex Optimization: Algorithms and Complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3–4, pp. 231–357, 2015, doi: 10.1561/2200000050.