MIT 6.7220/15.084 — Nonlinear Optimization Tue, Mar 5th 2024

Lecture 7
Gradient descent

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

With this lecture, we start exploring first-order optimization methods for nonlinear optimization. We begin our exploration in the case of minimization of *unconstrained* differentiable functions, that is, optimization problems of the form

$$\begin{aligned} \min_x \ & f(x) \\ \text{s.t.} \ & x \in \mathbb{R}^n. \end{aligned}$$

In this case, as we have mentioned several times (see, for example, Lecture 2), a point $x_\star \in \mathbb{R}^n$ solves the optimization problem only if $\nabla f(x) = 0$. Furthermore, when f is convex, the necessary condition also becomes sufficient (Lecture 3).

1 The fundamental idea of gradient descent

The function f might be complicated. A fundamental idea for constructing an optimization algorithm is to approximate f with tractable simpler models and take steps assuming the model is locally accurate at the current point x_t. For gradient descent and, more generally, first-order methods, the local model of the function is obtained from truncating the Taylor approximation of f around the current point x_t to its first order:

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t \rangle \qquad \text{for all } x \text{ "close to" } x_t.$$

The idea of gradient descent is then to move in the direction that minimizes the approximation of the objective above, that is, move a certain amount $\eta > 0$ in the direction $-\nabla f(x_t)$ of steepest descent of the function:

$$\boxed{x_{t+1} \coloneqq x_t - \eta \nabla f(x_t)}. \tag{1}$$

We will call x_0 the *initial point*, and the parameter $\eta > 0$ the *"stepsize"* or *"learning rate"*.

2 Analysis for L-smooth functions

In order to give concrete rates for the convergence of gradient descent, we will make some assumptions regarding the smoothness of the function. In particular, we will require a bound on how fast the function's gradient can change by moving slightly in any direction. This implies that as we start taking a movement in the direction $-\nabla f(x_t)$ from x_t, the new gradients we will encounter on the way will still be mostly aligned with $\nabla f(x_t)$, and so $-\nabla f(x_t)$ will continue to be a good descent direction for a while.

*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 L-smoothness

Specifically, we will require that the gradient $\nabla f(x)$ be a L-Lipschitz continuous for some constant $L \geq 0$. This condition is often called *L-smoothness* in the literature. We present it now for general functions with arbitrary convex domains Ω; today, we will only care about the case $\Omega = \mathbb{R}^n$.

Definition 2.1 (L-smoothness). A differentiable function $f : \Omega \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz continuous, that is,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \qquad \forall x, y \in \Omega.$$

An immediate consequence of L-smoothness is that the function admits a *quadratic upper bound*. This property will come in extremely handy in the analysis.

Theorem 2.1 (Quadratic upper bound). Let $f : \Omega \to \mathbb{R}$ be L-smooth on a convex domain Ω. Then, we can upper bound the function f as

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 \qquad \forall x, y \in \Omega. \tag{2}$$

Proof. The idea is simple: we express the growth $f(y) - f(x)$ as the integral of the gradient on the line connecting x to y, and we then use the Lipschitzness bound on the growth of the gradient ∇f to bound the growth:

$$\begin{aligned}
f(y) - f(x) &= \int_0^1 \Big\langle \nabla f(x + t \cdot (y - x)), y - x \Big\rangle \, \mathrm{d}t \\
&= \left(\int_0^1 \Big\langle \nabla f(x + t \cdot (y - x)) - \nabla f(x), y - x \Big\rangle \, \mathrm{d}t \right) + \langle \nabla f(x), y - x \rangle \\
&\leq \left(\int_0^1 \|\nabla f(x + t \cdot (y - x)) - \nabla f(x)\|_2 \cdot \|y - x\|_2 \, \mathrm{d}t \right) + \langle \nabla f(x), y - x \rangle \\
&\leq \left(\int_0^1 tL\|y - x\|_2^2 \, \mathrm{d}t \right) + \langle \nabla f(x), y - x \rangle \\
&= \frac{L}{2}\|y - x\|_2^2 + \langle \nabla f(x), y - x \rangle.
\end{aligned}$$

Rearranging, we obtain the statement. □

We also mention the following characterization.

Theorem 2.2. For twice differentiable functions $f : \Omega \to \mathbb{R}$ defined on an open set $\Omega \subseteq \mathbb{R}^n$, an equivalent condition of L-smoothness is

$$\left| v^\top \nabla^2 f(x) v \right| \leq L \qquad \forall x \in \Omega, v \in \mathbb{R}^n : \|v\|_2 = 1.$$

2.2 Convergence in gradient norm: The gradient descent lemma

We can use the quadratic upper bound (Theorem 2.1) to quantify the improvement of one step of the gradient descent algorithm. This bound on the improvement is often called the *gradient descent lemma*.

Theorem 2.3 (Gradient descent lemma). Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Then, for any $0 < \eta \leq \frac{1}{L}$, each step of gradient descent (1) guarantees

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

Proof. We start by writing (2) for the choice $x = x_t, y = x_{t+1}$:

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|_2^2.$$

Plugging in the gradient descent step $x_{t+1} = x_t - \eta \nabla f(x_t)$, we therefore obtain

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \eta\|\nabla f(x_t)\|_2^2 + \frac{L}{2}\eta^2\|\nabla f(x_t)\|_2^2 \\ &= f(x_t) - \eta\left(1 - \frac{L}{2}\eta\right)\|\nabla f(x_t)\|_2^2. \end{aligned}$$

If $\eta \leq \frac{1}{L}$, the parenthesis is $1 - \frac{L}{2}\eta \geq \frac{1}{2}$, and the result follows. □

The previous result shows that for L-smooth functions, there exists a good choice of learning rate (namely, $\eta = \frac{1}{L}$) such that each step of gradient descent guarantees to improve the function value if the current point does not have a zero gradient.

Assuming that f is lower bounded, the gradient descent lemma also implies that the gradients of the points produced by gradient descent must eventually become small. Indeed, imagine running the algorithm for T iterations. Then, by repeatedly using the gradient descent lemma, we would have

$$f(x_T) \leq f(x_0) - \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2.$$

Assume now that f is lower bounded by $f_\star := \inf f$. Since $f(x_T) \geq f_\star$, the above implies that

$$\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2 \leq \frac{2}{\eta}(f(x_0) - f_\star).$$

So, at least one among the $\left\{\|\nabla f(x_t)\|_2^2\right\}_{t=0}^{T-1}$ must be upper bounded by $\frac{2}{\eta T}(f(x_0) - f_\star)$. More formally, we have the following.

Theorem 2.4. Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, and let $0 < \eta \leq \frac{1}{L}$. Then, by running gradient descent for T iterations, at least one of the points $\{x_t\}$ encountered must satisfy

$$\|\nabla f(x_t)\|_2 \leq \sqrt{2 \cdot \frac{f(x_0) - f_\star}{\eta T}}.$$

We remark that the above result does not require convexity of f.

2.3 Convergence in convex function value: The Euclidean mirror descent lemma

The result in Theorem 2.4 only guarantees the decrease of the objective's *gradient norm*, not of the function value. However, while a small gradient indicates a condition of almost local optimality, it does not imply that the function value is small. To see this, consider the function $f(x) = \varepsilon \log(1 + e^x)$. The gradient satisfies $\nabla f(x) \leq \varepsilon$ for all $x \in \mathbb{R}$, and so x might be arbitrarily far from optimal while the gradient is small.

■ **Step I: The Euclidean mirror descent lemma.** To give guarantees in function value, we will now further assume that f is *convex*, and we will leverage what we will call the *Euclidean mirror descent lemma*. In a future lecture, we will see a more general version of this lemma, which we will call *mirror descent lemma* without further qualifications.

Theorem 2.5 (Euclidean mirror descent lemma). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function. Then, for any choice of stepsize η, any two consecutive points (x_t, x_{t+1}) produced by the gradient descent algorithm (1) satisfy

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

Proof. The proof rests on the following critical observation, often called the *three-point equality*, which can be checked by expanding the squared norms: for any $y \in \mathbb{R}^n$,

$$\langle x_t - x_{t+1}, y - x_t \rangle = -\frac{1}{2}\Big(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\Big).$$

Since f is convex by hypothesis, it satisfies the linear lower bound we saw in Lecture 3. In particular, we can lower bound the value of $f(y)$, for any $y \in \mathbb{R}^n$, with the linearization at the point x_t, obtaining

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle.$$

Plugging in the relationship (1), that is, $x_{t+1} = x_t - \eta \nabla f(x_t)$, we then obtain

$$f(y) \geq f(x_t) + \frac{1}{\eta}\langle x_t - x_{t+1}, y - x_t \rangle.$$

Substituting the critical observation above and rearranging, we obtain

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

which is the statement. □

■ **Step II: Telescoping.** We can use the Euclidean mirror descent lemma above to get a rate of decrease in terms of the objective function value $f(x_t)$.

Two steps are necessary: first, since $x_{t+1} - x_t = -\eta \nabla f(x_t)$, by definition of the gradient descent algorithm (1), we can replace the term $\|x_{t+1} - x_t\|_2^2$ in the statement of Theorem 2.5 with $\eta^2 \|\nabla f(x_t)\|_2^2$, obtaining

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

Assuming f is L-smooth and $\eta \leq \frac{1}{L}$, we can then use the gradient descent lemma (Theorem 2.3) to bound

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

obtaining the following corollary.

Corollary 2.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth, and $0 < \eta \le \frac{1}{L}$. Then, any two consecutive points (x_t, x_{t+1}) produced by gradient descent satisfy

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

In particular, if f has a minimizer $x_\star$, the above corollary implies that, whenever $\eta \le \frac{1}{L}$,

$$f(x_{t+1}) \le f(x_\star) + \frac{1}{2\eta}\Big(\|x_\star - x_t\|_2^2 - \|x_\star - x_{t+1}\|_2^2\Big).$$

Summing over $t = 0, ..., T-1$, and noticing that the right-hand side is telescopic, we have

$$\sum_{t=0}^{T-1} f(x_{t+1}) \le T f(x_\star) + \frac{1}{2\eta}\|x_\star - x_0\|_2^2.$$

Since $f(x_t)$ is nonincreasing in t by the gradient descent lemma, then $\sum_{t=0}^{T-1} f(x_{t+1}) \ge T f(x_T)$, and so we can write

$$T f(x_T) \le T f(x_\star) + \frac{1}{2\eta}\|x_\star - x_0\|_2^2 \quad \Longrightarrow \quad f(x_T) \le f(x_\star) + \frac{1}{2T\eta}\|x_\star - x_0\|_2^2.$$

So, we have proved the following.

Theorem 2.6. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth with minimizer $x_\star \in \mathbb{R}^n$, and $0 < \eta \le \frac{1}{L}$. The t-th point $x_t \in \mathbb{R}^n$ produced by gradient descent satisfies

$$f(x_t) - f(x_\star) \le \frac{\|x_\star - x_0\|_2^2}{2t\eta} \qquad \forall t = 1, 2, 3, ...$$

2.4 Convergence in iterates

One might wonder whether Theorem 2.6 also implies that the distance between x_t and $x_\star$ shrinks at the same rate, say $\|x_t - x_\star\| \le O(\frac{1}{t})$. Unfortunately, such a result is false without a stronger assumption than L-smoothness. Indeed, as shown by Nesterov, Y. [Nes18] (Theorem 2.1.7 in his book), the convergence to the optimal point might be *arbitrarily slow*.

3 Faster convergence under the Polyak-Łojasiewicz condition

We now show that under stronger assumptions on the function f, gradient descent can be shown to converge to the optimal point exponentially fast, and *in iterates*. This contrasts with what we saw in the previous section, where $\text{poly}(1/\varepsilon)$ iterates were needed to reach an approximation guarantee of ε.

3.1 The PŁ condition

While most course material considers strong convexity (with which you played in the first problem set), we instead focus on a much more general condition called the *Polyak-Łojasiewicz (PŁ) condition*, introduced independently by Polyak [Pol63] and Łojasiewicz [Łoj63] in 1963.

Fundamentally, the PŁ condition establishes a lower bound on the norm of the function, which increases as the point becomes more and more suboptimal.

Definition 3.1 (PŁ condition). A lower-bounded function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the *Polyak-Łojasiewicz (PŁ) condition* if there exists $\mu > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_\star) \qquad \forall x \in \mathbb{R}^n,$$

where $f_\star := \inf f$.

The benefits of this approach compared to the standard approach of considering strongly convex functions are many:

- strong convexity implies the PŁ condition but not *vice versa*.
- the proof of the exponential rate is simpler.
- the PŁ condition does not imply convexity (unlike strong convexity).

In particular, the following condition is sufficient (but not necessary) for the PŁ condition:

Theorem 3.1. If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable and $\nabla^2 f(x) \succcurlyeq \mu I$ for all $x \in \mathbb{R}^n$ (that is, f is μ-strongly convex on $\mathbb{R}^n$), then f satisfies the PŁ condition with PŁ constant μ.

[▷ You should try to prove this!]

3.2 Gradient descent's convergence rate with the PŁ condition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth lower-bounded function that satisfies the PŁ condition for some PŁ constant $\mu > 0$. Using the gradient descent lemma (Theorem 2.3) and the PŁ condition together, for $\eta = \frac{1}{L}$ we have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|_2^2 \\ &\leq f(x_t) - \frac{\mu}{L}(f(x_t) - f_\star) \\ &= \left(1 - \frac{\mu}{L}\right) f(x_t) + \frac{\mu}{L} f_\star. \end{aligned}$$

Subtracting $f_\star$ on both sides yields

$$f(x_{t+1}) - f_\star \leq \left(1 - \frac{\mu}{L}\right)(f(x_t) - f_\star).$$

Solving the recurrence, we have proved the following.

Theorem 3.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function lower-bounded by $f_\star$ that satisfies the PŁ condition for some PŁ constant $\mu > 0$. Let stepsize $\eta = \frac{1}{L}$. The t-th point $x_t \in \mathbb{R}^n$ produced by gradient descent satisfies

$$f(x_t) - f_\star \leq \left(1 - \frac{\mu}{L}\right)^t (f(x_0) - f_\star).$$

3.3 Convergence in iterates

We now establish that a function that satisfies the PŁ condition must attain a minimum and that the iterates produced by gradient descent must converge at a rate exponential in t.

We show this from first principles, following Polyak, B. T. [Pol63]'s original proof.

Theorem 3.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, lower bounded by $f_\star$, and satisfy the PŁ condition with PŁ constant $\mu > 0$. Then, the sequence $\{x_t\}$ of iterates produced by gradient descent instantiated with $0 < \eta \leq \frac{1}{L}$ *converges* to some point $x_* \in \mathbb{R}^n$, with

$$\|x_t - x_\star\|_2^2 \leq \frac{8\eta L^2}{\mu^2}\left(1 - \frac{\mu}{L}\right)^{t-1}(f(x_0) - f_\star).$$

Hence, by continuity and using Theorem 3.2, $\lim_{t\to\infty} f(x_t) = f(x_\star) = f_\star$.

Proof. At all times t, we have

$$\begin{aligned}
\left\|x_{t+1} - x_t\right\|_2^2 &= \eta^2 \|\nabla f(x_t)\|_2^2 \\
&\leq 2\eta(f(x_t) - f(x_{t+1})) && \text{(gradient descent lemma)} \\
&\leq 2\eta(f(x_t) - f_\star) \\
&\leq 2\eta\left(1 - \frac{\mu}{L}\right)^t (f(x_0) - f_\star) && \text{(from Theorem 3.2).}
\end{aligned}$$

To simplify notation, let $\alpha^2 \coloneqq \left(1 - \frac{\mu}{L}\right)$. Clearly, $0 \leq \alpha < 1$. To show convergence of the sequence $\{x_t\}$, we will show that it is a Cauchy sequence and invoke the completeness of $\mathbb{R}^n$. Specifically, for any $m < n$,

$$\begin{aligned}
\|x_m - x_n\|_2 &\leq \sum_{t=m}^{n-1} \left\|x_t - x_{t+1}\right\|_2 \\
&\leq \sqrt{2\eta(f(x_0) - f_\star)} \cdot \sum_{t=m}^{n-1} \alpha^t \\
&\leq \sqrt{2\eta(f(x_0) - f_\star)} \cdot \frac{\alpha^m}{1-\alpha}. && \text{(closed form for a geometric series of ratio } < 1)
\end{aligned}$$

This shows that $\|x_m - x_m\|_2 \to 0$ as $m, n \to \infty$; hence, $\{x_t\}$ is a Cauchy sequence (by definition). By completeness of $\mathbb{R}^n$, it follows that x_t converges to some $x_\star$. Letting $m = t$ and $n \to \infty$ above,

$$\|x_t - x_\star\|_2 \leq \sqrt{2\eta(f(x_0) - f_\star)} \cdot \frac{\alpha^t}{1-\alpha}.$$

Squaring, we obtain

$$\begin{aligned}
\|x_t - x_\star\|_2^2 &\leq 2\eta(f(x_0) - f_\star) \cdot \frac{(\alpha^2)^t}{(1-\alpha)^2} \\
&\leq 8\eta(f(x_0) - f_\star) \cdot \frac{(\alpha^2)^t}{\alpha^2(1-\alpha^2)^2}. && \left(\text{since } \frac{1}{(1-\alpha)^2} \leq \frac{4}{\alpha^2(1-\alpha^2)^2}\right)
\end{aligned}$$

Plugging in the definition of $\alpha^2 = \left(1 - \frac{\mu}{L}\right)$, we finally obtain

$$\|x_t - x_\star\|_2^2 = \frac{8\eta L^2}{\mu^2}\left(1 - \frac{\mu}{L}\right)^{t-1}(f(x_0) - f_\star),$$

which is the statement. □

Further readings

The book by Nesterov, Y. [Nes18] is an authoritative reference for a comprehensive treatment of optimization algorithms for convex and nonconvex functions.

For properties of functions that satisfy the PŁ condition, the paper by Karimi, H., Nutini, J., & Schmidt, M. [KNS16] is an approachable source.

[Nes18] Y. Nesterov, *Lectures on Convex Optimization*. Springer International Publishing, 2018. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-91578-4

[Pol63] B. T. Polyak, "Gradient methods for the minimisation of functionals," *Ussr Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 864–878, Dec. 1963, doi: 10.1016/0041-5553(63)90382-3.

[Łoj63] S. Łojasiewicz, "A topological property of real analytic subsets," *Coll. du CNRS, Les équations aux dérivées partielles*, vol. 117, no. 87–89, p. 2–3, 1963.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt, "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition," in *Machine Learning and Knowledge Discovery in Databases*, Springer, Sep. 2016, pp. 795–811. doi: 10.1007/978-3-319-46128-1_50.