CSIE5410 Optimization algorithms

Lecture 3: Gradient descent

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

27.09.2018

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

Let $f : \mathbb{R}^p \to \mathbb{R}$ be differentiable on $\mathbb{R}^p$. Consider the optimization problem:

$$f^\star = \min_x \{\, f(x) \mid x \in \mathbb{R}^p \,\}.$$

This lecture introduces the gradient descent—arguably the simplest approach to solving such an optimization problem—and corresponding complexity analyses.

Recommended reading

- Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*. (Chapter 2).
- S. Bubeck. 2015. *Convex Optimization: Algorithms and Complexity*. (Chapter 3).
- D. P. Bertsekas. 2016. *Nonlinear Programming*. (Chapter 1).
- A. Mądry. 2017. Gradient descent: The mother of all algorithms? (Talk at the Simons Institute, video available on Youtube).

Table of contents

Prelude

Origin of gradient descent

Augustin-Louis Cauchy
(1789–1857)

How do one solve the equation

$$u(x) = 0, \quad x \in \mathbb{R}^p,$$

knowing that $u \geq 0$?

Notice that

$$u(x) = 0 \Leftrightarrow x \in \arg\min_x \{ u(x) \mid x \in \mathbb{R}^p \}.$$

It will suffice to let indefinitely decrease the function u, until it vanishes.

`https://www.flickr.com/photos/smithsonian/2550744825`

C. Lemaréchal. Cauchy and the gradient method. 2012.

Cauchy's argument

We have

$$u(x + \Delta) \approx u(x) + \langle \nabla u(x), \Delta \rangle \, .$$

Choose $\Delta = -\theta \nabla u(x)$ for some positive real number θ. Then

$$u(x - \theta \nabla u(x)) \approx u(x) - \theta \|\nabla u(x)\|_2^2$$

will become smaller than u if θ is small enough.

If the new value of u is not a minimum, one can deduce, again proceeding in the same way, a third value still smaller; and, so continuing, smaller and smaller values of u will be found, which will converge to a minimal value of u.

C. Lemaréchal. Cauchy and the gradient method. 2012.

Verification (1/2)

Proposition. Suppose that u is differentiable at $x \in \mathbb{R}^p$. If $\nabla u(x) \neq 0$, then $u(x - \theta \nabla u(x)) < u(x)$ for θ small enough.

Proof. Define $g(\theta) := u(x - \theta \nabla u(x))$. Then we have $g(\theta) = g(0) + g'(0)\theta + \varepsilon$ for some $\varepsilon = o(\theta)$; that is,

$$u(x - \theta \nabla u(x)) = u(x) - \theta \|\nabla u(x)\|_2^2 + \varepsilon.$$

Since $\lim_{\theta \downarrow 0} \frac{\varepsilon}{\theta} = 0$, we have, for θ small enough,

$$\frac{\varepsilon}{\theta} < \|\nabla u(x)\|_2^2.$$

A. A. Ahmadi. Computing and Optimization: Lecture 8. 2017.
http://aaa.princeton.edu/orf363

Verification (2/2)

Question. Is a sequence of decreasing non-negative numbers necessarily converging to zero?

Theorem. If a sequence is decreasing and bounded from below, then it converges to a number.

Answer. No. Consider the case $f(x) := x^2$. The sequence $(f(1 + (1/n)))_{n \in \mathbb{N}}$ is strictly decreasing, but is bounded from below by $f(1) = 1$, far away from the minimum $f(0) = 0$.

Prototype of gradient descent

Consider the problem of minimizing a differentiable function f on $\mathbb{R}^p$.

Algorithm Gradient descent

1: Set $x_0 \in \mathbb{R}^p$, $T_\varepsilon \in \mathbb{N}$.
2: **for** $t = 1, \ldots, T_\varepsilon$ **do**
3: Select $\eta_t \in]0, +\infty[$ ▷ *step size*
4: $x_t \leftarrow x_{t-1} - \eta_t \nabla f(x_{t-1})$
5: **end for**
6: $x_\varepsilon \leftarrow x_{T_\varepsilon}$

Issues

1. How do we choose T_ε?
2. How do we select the step size η_t?

Importance of choosing when to stop

One of the key features of *an algorithm*:

"An algorithm must always terminate after a finite number of steps."

How do you value the following result?

"We show that gradient descent converges to a local minimizer, almost surely with random initialization. This is proved by applying the Stable Manifold Theorem from dynamical systems theory."

D. Knuth. *The Art of Computer Programming Volume 1: Fundamental Algorithms. 1997.*

J. D. Lee *et al.* Gradient descent only converges to minimizers. 2016.

Importance of step size selection

Consider the problem of minimizing $f(x) := x^2$ on $\mathbb{R}$. Suppose that we set $x_0 := 2$.

Case 1. If we set $\eta_1 = 0.5$, then

$$x_1 = x_0 - \eta_1 f'(x_0) = 2 - 0.5 \times (2 \times 2) = 0,$$

which is the minimizer.

Case 2. If we set $\eta_t = \eta := 1$ for all t, then

$$\begin{aligned}
x_1 &= x_0 - \eta f'(x_0) = 2 - 1 \times (2 \times 2) = -2, \\
x_2 &= x_1 - \eta f'(x_1) = -2 - 1 \times (2 \times (-2)) = 2, \\
x_3 &= x_1 = -2, \\
x_4 &= x_2 = 2, \\
&\vdots
\end{aligned}$$

Smoothness & strong convexity

Necessary condition of optimality

Definition. Let f be a function differentiable on $\mathbb{R}^p$. We say that x is a *stationary point* of f on $\mathbb{R}^p$, if and only if $\nabla f(x) = 0$.

Proposition. If $x^\star \in \mathbb{R}^p$ is a minimizer of a function f differentiable on $\mathbb{R}^p$, then $x^\star$ is a stationary point.

Proof. Trivial. (Why?)

Question. Is a stationary point necessarily a minimizer? Why or why not?

Fixed-point iteration

Let $x^\star$ be a minimizer. Then we have $\nabla f(x^\star) = 0$. That is,

$$x^\star - \eta \nabla f(x^\star) = x^\star, \quad \forall \eta \in]0, +\infty[.$$

Define $T_\eta : x \mapsto x - \eta \nabla f(x)$. Then $x^\star$ is a *fixed point* of T_η, i.e., $x^\star$ is a solution to the equation

$$x = T_\eta(x).$$

Definition. We say that T_η is contractive for some $\eta \in]0, +\infty[$, if and only if there exists some $\rho \in [0, 1[$, such that

$$\|T_\eta(y) - T_\eta(x)\|_2 \leq \rho \|y - x\|_2, \quad \forall x, y \in \mathbb{R}^p.$$

Convergence of the fixed-point iteration

Proposition. If T_η is *contractive* for some $\eta \in]0, +\infty[$, then gradient descent with the constant step size η converges to an optimum.

Proof. Let $x^\star$ be a minimizer of f on $\mathbb{R}^p$. We have

$$\|T_\eta(x) - T_\eta(x^\star)\|_2 = \|T_\eta(x) - x^\star\|_2 \leq \rho\|x - x^\star\|_2, \quad \forall x \in \mathbb{R}^p,$$

for some $\rho \in]0, 1[$. Therefore,

$$\begin{aligned}\|x_t - x^\star\|_2 = \|T_\eta(x_{t-1}) - T_\eta(x^\star)\|_2 &\leq \rho\|x_{t-1} - x^\star\|_2 \\ &\leq \rho^t\|x_0 - x^\star\|_2, \quad \forall t \in \mathbb{N}.\end{aligned}$$

Emergence of smoothness & strong convexity (1/2)

Proposition. Let f be a function twice differentiable on $\mathbb{R}^p$. If T_η is contractive, then there exists some $\rho \in]0,1[$, such that

$$0 < \frac{1-\rho}{\eta} I \leq \nabla^2 f(x) \leq \frac{1+\rho}{\eta} I, \quad \forall x \in \mathbb{R}^p.$$

Quick check. Therefore, the function f must be *convex*.

Question. What is the meaning of the Hessian upper bound?

Emergence of smoothness & strong convexity (2/2)

Proof. By contractivity of T_η, there exists some $\rho \in]0,1[$ such that

$$\|(x + t\delta) - \eta\nabla f(x + t\delta) - x + \eta\nabla f(x)\|_2 \leq \rho\|(x + t\delta) - x\|_2,$$

for all $t \in \mathbb{R}$ and $\delta, x \in \mathbb{R}^p$. Then we have

$$\lim_{t\to 0} \left\| \delta - \frac{\eta}{t} \left[\nabla f(x + t\delta) - \nabla f(x)\right] \right\|_2 \leq \rho\|\delta\|_2,$$

i.e.,

$$\| \left[I - \eta\nabla^2 f(x)\right] \delta\|_2 \leq \rho\|\delta\|_2.$$

Therefore, it must hold that

$$-\rho I \leq I - \eta\nabla^2 f(x) \leq \rho I.$$

Strong convexity

Definition. Let f be a function twice differentiable on $\mathbb{R}^p$. We say that f is *μ-strongly convex* for some $\mu \in]0, +\infty[$, if and only if

$$\nabla^2 f(x) \geq \mu I, \quad \forall x \in \mathbb{R}^p.$$

Theorem. The notion of μ-strong convexity has two equivalent definitions.

1. For any $x, y \in \mathbb{R}^p$,

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y - x\|_2^2.$$

2. For any $x, y \in \mathbb{R}^p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2.$$

Proof of the first equivalent definitions (1/2)

Proof. ($\Leftarrow$) We have, for all $t \in \mathbb{R}$ and $x, \delta \in \mathbb{R}^p$,

$$\langle \nabla f(x + t\delta) - \nabla f(x), t\delta \rangle \geq \mu t^2 \|\delta\|_2^2.$$

Therefore, we write

$$\langle \delta, \nabla^2 f(x)\delta \rangle = \lim_{t \downarrow 0} \frac{\langle \nabla f(x + t\delta) - \nabla f(x), \delta \rangle}{t} \geq \mu \|\delta\|_2^2,$$

i.e., $\nabla^2 f(x) \geq \mu I$.

Proof of the first equivalent definitions (2/2)

Proof. ($\Rightarrow$) Define $\varphi(t) := \langle \nabla f(x + t(y - x)), y - x \rangle$. Then by the Newton-Leibniz formula, we write

$$\varphi(1) = \varphi(0) + \int_0^1 \varphi'(\tau)\,\mathrm{d}\tau$$

Then we obtain

$$\begin{aligned}
&\langle \nabla f(y), y - x \rangle - \langle \nabla f(x), y - x \rangle \\
&\quad = \int_0^1 \langle y - x, \nabla^2 f(x + \tau(y - x))(y - x) \rangle \,\mathrm{d}\tau \\
&\quad \geq \mu \|y - x\|_2^2.
\end{aligned}$$

Proof of the second equivalent definition (1/2)

Proof. ($\Leftarrow$) We have both

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2,$$
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|y - x\|_2^2.$$

Summing up the two inequalities, the desired inequality follows.

Proof of the second equivalent definition (2/2)

Proof. ($\Rightarrow$) Define $\psi(t) := f(x + t(y - x))$. Then we have, by the Newton-Leibniz formula again,

$$\psi(1) = \psi(0) + \int_0^1 \psi'(\tau)\,\mathrm{d}\tau.$$

Therefore, we write

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle \,\mathrm{d}\tau \\ &= \langle \nabla f(x), y - x \rangle + \int_0^1 \frac{1}{\tau} \langle \nabla f(x + \tau(y - x)) - \nabla f(x), \tau(y - x) \rangle \,\mathrm{d}\tau \\ &\geq \langle \nabla f(x), y - x \rangle + \int_0^1 \tau\mu \|y - x\|_2^2 \,\mathrm{d}\tau. \end{aligned}$$

Smoothness

Definition. We say that a twice differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is L-smooth for some $L \in]0, +\infty[$, if and only if

$$\nabla^2 f(x) \leq LI, \quad \forall x \in \mathbb{R}^p$$

Theorem. The notion of L-smoothness has two equivalent definitions.

1. For any $x, y \in \mathbb{R}^p$,

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L \|y - x\|_2^2.$$

2. For any $x, y \in \mathbb{R}^p$,

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

Proof. Similar to the previous slides.

Detour: Why is the smoothness parameter denoted by L?

Definition. We say that ∇f is L-Lipschitz continuous (w.r.t. the ℓ_2-norm), if and only if

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2, \quad \forall x, y \in \mathbb{R}^p.$$

Proposition. If ∇f is L-Lipschitz continuous, then f is L-smooth.

Proof. We write, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\langle \nabla f(y) - \nabla f(x), y - x\rangle &\leq \|\nabla f(y) - \nabla f(x)\|_2 \|y - x\|_2 \\ &\leq L\|y - x\|_2^2.\end{aligned}$$

Notice

It is simply for convenience of presentation that we start with the Hessian bounds and require the function to be twice differentiable.

When the function is differentiable, the following are definitions.

Definition. (μ-strong convexity)

1. $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y - x\|_2^2$.
2. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + (\mu/2)\|y - x\|_2^2$.

Definition. (L-smoothness)

1. $\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L \|y - x\|_2^2$.
2. $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + (L/2)\|y - x\|_2^2$.

Exercise

Let $A \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$. Define

$$f(x) := \frac{1}{2}\|Ax - b\|_2^2, \quad \forall x \in \mathbb{R}^p.$$

Question. Is f smooth? What is the corresponding parameter L?

Question. Is f strongly convex? What is the corresponding parameter μ?

Analysis of gradient descent: Smoothness

Problem formulation

Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex and L-smooth. Solve the optimization problem

$$x^\star \in \arg\min_x \{\, f(x) \mid x \in \mathbb{R}^p \,\}.$$

Example. Linear regression corresponds to

$$f(x) := \frac{1}{2}\|Ax - b\|_2^2,$$

which is L-smooth with $L = \|A^{\mathrm{T}}A\|_{2\to 2}$.

Definition. Let $M \in \mathbb{R}^{n\times p}$. The *induced operator norm* is defined as

$$\|M\|_{q_1\to q_2} := \max\left\{ \frac{\|Mx\|_{q_2}}{\|x\|_{q_1}} \;\middle|\; x \in \mathbb{R}^p \right\}.$$

Iteration complexity

Algorithm Gradient descent

1: Set $x_0 \in \mathbb{R}^p$, $T_\varepsilon \in \mathbb{N}$.
2: **for** $t = 1, \dots, T_\varepsilon$ **do**
3: Select $\eta_t \in]0, +\infty[$ ▷ *step size*
4: $x_t \leftarrow x_{t-1} - \eta_t \nabla f(x_{t-1})$
5: **end for**
6: $x_\varepsilon \leftarrow x_{T_\varepsilon}$

Theorem. If $\eta_t = \eta \in]0, 2L^{-1}[$ for all t, then

$$f(x_t) - f^\star \leq \frac{2(f(x_0) - f^\star)\|x_0 - x^\star\|_2^2}{2\|x_0 - x^\star\|_2^2 + t\eta(2 - L\eta)(f(x_0) - f^\star)}.$$

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

Monotonicity of function values

Proposition. It holds that

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

Proof. By smoothness, we have

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

Replace x_t with its explicit expression $x_{t-1} - \eta\nabla f(x_{t-1})$.

Key lemma

Lemma. It holds that

$$f(y) \geq f(x)+\langle \nabla f(x), y-x\rangle+\frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^p.$$

Proof. Define $\varphi(y) := f(y) - \langle \nabla f(x), y\rangle$. Then φ is also L-smooth, and achieves its minimum at x (Why?). We have

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2.$$

Lemma. It holds that

$$\langle \nabla f(y) - \nabla f(x), y-x\rangle \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^p.$$

Proof of the iteration complexity bound

The proof below follows exactly that of Nesterov.

Proof. Define $r_t := \|x_t - x^\star\|$. Then $(r_t)_{t\geq 0}$ is a non-increasing sequence, as

$$\begin{aligned} r_{t+1}^2 &= \|x_t - \eta \nabla f(x_t) - x^\star\|^2 \\ &= r_t^2 + \eta^2 \|\nabla f(x_t)\|^2 - 2\eta \left\langle \nabla f(x_t), x_t - x^\star \right\rangle \\ &\leq r_t^2 - \eta \left(\frac{2}{L} - \eta \right) \|\nabla f(x_t)\|^2, \end{aligned}$$

where we have used the previous lemma and the fact that $\nabla f(x^\star) = 0$.

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

Proof of the iteration complexity bound

Proof continued. Define $\Delta_t := f(x_t) - f^\star$. We write

$$\Delta_t \leq \langle \nabla f(x_t), x_t - x^\star \rangle \leq r_t \|\nabla f(x_t)\| \leq r_0 \|\nabla f(x_t)\|.$$

Recall that by monotonicity, we have

$$\Delta_{t+1} \leq \Delta_t - \omega \|\nabla f(x_t)\|^2 \leq \Delta_t - \frac{\omega}{r_0^2} \Delta_t^2,$$

where $\omega := \eta \left(1 - \frac{L}{2}\eta\right)$. Therefore, we obtain

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{\omega}{r_0^2} \frac{\Delta_t}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{\omega}{r_0^2}.$$

Summing up the inequalities, the theorem follows.

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

Most well-known version

Recall the monotonicity result is sharpest when $\eta = L^{-1}$.

Corollary. Set $\eta = L^{-1}$. Then it holds that

$$f(x_t) - f^\star \leq \frac{2L\|x_0 - x^\star\|_2^2}{t+4} = O(t^{-1}).$$

Proof. It remains to bound $f(x_0) - f^\star$ as

$$f(x_0) \leq f(x^\star) + \langle \nabla f(x^\star), x_0 - x^\star \rangle + \frac{L}{2}\|x_0 - x^\star\|_2^2.$$

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

Analysis of gradient descent: Smoothness + Strong convexity

Problem formulation

Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-smooth and μ-strongly convex. Solve the optimization problem

$$x^\star \in \arg\min_x \{\, f(x) \mid x \in \mathbb{R}^p \,\}.$$

Example. The ridge regression corresponds to

$$f(x) := \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_2^2,$$

which is L-smooth with $L = \|A^{\mathrm{T}} A\|_{2\to 2} + 2\lambda$ and μ-strongly convex with $\mu = 2\lambda$.

Implication of strong convexity

Proposition. The minimizer $x^\star$ is unique.

Proof. For any $x \in \mathbb{R}^p$, we have

$$f(x) \geq f(x^\star) + \langle \nabla f(x^\star), x - x^\star \rangle + \frac{\mu}{2}\|x - x^\star\|_2^2.$$

Recall that $\nabla f(x^\star)$ is necessarily zero.

Proposition. Let $f : \mathbb{R}^p \to]-\infty, +\infty]$ be convex. Let $y \in \mathbb{R}^p$. Then the $x^\star$ below is uniquely defined.

$$x^\star \in \operatorname*{arg\,min}_x \left\{ f(x) + \|x - y\|_2^2 \;\middle|\; x \in \mathbb{R}^p \right\}.$$

Proof. Exercise. Notice that f may not be differentiable.

Iteration complexity

Algorithm Gradient descent

1: Set $x_0 \in \mathbb{R}^p$, $T_\varepsilon \in \mathbb{N}$.
2: **for** $t = 1, \dots, T_\varepsilon$ **do**
3: Select $\eta_t \in]0, +\infty[$ ▷ *step size*
4: $x_t \leftarrow x_{t-1} - \eta_t \nabla f(x_{t-1})$
5: **end for**
6: $x_\varepsilon \leftarrow x_{T_\varepsilon}$

Theorem. If $\eta_t = \eta \in]0, 2(\mu + L)^{-1}]$, then it holds that

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

Remark. Is the bound always meaningful?

Improved key lemma

Lemma. It holds that

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(y) - \nabla f(x)\|_2^2.$$

Proof. Define

$$\varphi(x) := f(x) - \frac{\mu}{2} \|x\|_2^2.$$

Then φ is $(L - \mu)$-smooth.

- If $L = \mu$, then the lemma follows from an equivalent definition of strong convexity and the key lemma for smooth functions.
- Otherwise, applying the key lemma for smooth functions, we obtain

$$\langle \nabla \varphi(y) - \nabla \varphi(x), y - x \rangle \geq \frac{1}{L - \mu} \|\nabla \varphi(y) - \nabla \varphi(x)\|_2^2.$$

Proof of the iteration complexity bound

The proof below follows exactly that of Nesterov.

Proof. Define $r_t := \|x_t - x^\star\|$. Then we have

$$\begin{aligned}
r_{t+1}^2 &= r_t^2 + \eta^2 \|\nabla f(x_t)\|^2 - 2\eta \langle \nabla f(x_t), x_t - x^\star \rangle \\
&\leq r_t^2 + \eta^2 \|\nabla f(x_t)\|^2 - 2\eta \left(\frac{\mu L}{\mu + L} r_t^2 + \frac{1}{\mu + L} \|\nabla f(x_t)\|^2 \right) \\
&\leq \left(1 - \frac{2\eta\mu L}{\mu + L} \right) r_t^2 + \eta \left(\eta - \frac{2}{\mu + L} \right) \|\nabla f(x_t)\|^2 \\
&\leq \left(1 - \frac{2\eta\mu L}{\mu + L} \right) r_t^2.
\end{aligned}$$

Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*.

Most well-known version

Corollary. Set $\eta = \frac{2}{\mu+L}$. Then it holds that

$$\|x_t - x^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^t \|x_0 - x^\star\|_2,$$

$$f(x_t) - f^\star \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2t} \|x_0 - x^\star\|_2^2.$$

Proof. Direct calculation.

Remark. This is also called the *linear* convergence rate.

Linear convergence without strong convexity

Polyak-Łojasiewicz condition

Linear convergence rate can be achieved without strong convexity.

Definition. We say that the μ-Polyak-Łojasiewicz condition holds for some $\mu \in]0, +\infty[$, if and only if

$$f(x) - f^\star \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2, \quad \forall x \in \mathbb{R}^p.$$

Remark. The condition is weaker than strong convexity.

Theorem. If f is L-smooth, and the μ-Polyak-Łojasiewicz condition holds, then setting $\eta = L^{-1}$, it holds that

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

B. T. Polyak. 1963. Gradient methods for the minimisation of functionals.

Quick check (1/2)

Proposition. If f is μ-strongly convex, then it satisfies the μ-Polyak-Łojasiewitz condition.

Remark. The Polyak-Łojasiewitz condition can also be satisfied without strong convexity. For example, consider the function

$$f(x) := \begin{cases} (x-0.5)^2, & x \in [0.5, +\infty[, \\ 0, & x \in [-0.5, +0.5[, \\ (x+0.5)^2, & \text{otherwise.} \end{cases}$$

It is not strongly convex, but satisfies the 2-Polyak-Łojasiewitz condition.

Quick check (2/2)

Proof. (SC $\to$ PL) Define $\varphi(y) := f(y) - \langle \nabla f(x), y \rangle$. Then φ is μ-strongly convex and achieves its minimum at x. We write

$$\begin{aligned}
\varphi(x) &= \min_v \{\, \varphi(v) \mid v \in \mathbb{R}^p \,\} \\
&\geq \min_v \left\{ \varphi(y) + \langle \nabla\varphi(y), v - y \rangle + \frac{\mu}{2}\|v - y\|_2^2 \;\middle|\; v \in \mathbb{R}^p \right\} \\
&= \varphi(y) - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2.
\end{aligned}$$

That is,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

Set $x = x^\star$ in the inequality above.

Proof of the iteration complexity bound

This result was first proved by B. T. Polyak. The following proof is from H. Karimi *et al.*

Proof. By the smoothness condition, we have

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

Applying the Polyak-Łojasiewitz condition, we obtain

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

The theorem follows.

B. T. Polyak. 1963. Gradient methods for the minimisation of functionals.
H. Karimi *et al.* 2016. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition.

Conclusions

Summary (1/2)

Equivalent definitions of smoothness.

1. $\nabla^2 f(x) \leq LI$.
2. $\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L\|y - x\|_2^2$.
3. $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$.

Equivalent definitions of strong convexity.

1. $\nabla^2 f(x) \geq \mu I$.
2. $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu\|y - x\|_2^2$.
3. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$.

Summary (2/2)

Algorithm Gradient descent

1: Set $x_0 \in \mathbb{R}^p$, $T_\varepsilon \in \mathbb{N}$.
2: **for** $t = 1, \dots, T_\varepsilon$ **do**
3: Select $\eta_t \in]0, +\infty[$ ▷ *step size*
4: $x_t \leftarrow x_{t-1} - \eta_t \nabla f(x_{t-1})$
5: **end for**
6: $x_\varepsilon \leftarrow x_{T_\varepsilon}$

Convergence rates in function value.

- Smoothness only: $O(t^{-1})$.
- Smoothness & strong convexity: $O(\rho^{2t})$ for $\rho := \frac{L-\mu}{L+\mu}$.

Next lecture

- Accelerated gradient descent.
- Mirror descent*.