

CME 307 / MS&E 311: Optimization

Gradient descent

Professor Udell

Management Science and Engineering
Stanford

April 24, 2023

Outline

Unconstrained minimization

Gradient descent

What functions?

Analysis via Polyak-Lojasiewicz condition

Unconstrained minimization

$$\text{minimize } f(x)$$

- ▶ $f : \mathbf{R}^n \rightarrow \mathbf{R}$ differentiable
- ▶ assume optimal value $f^* = \inf_x f(x)$ is attained (and finite)
- ▶ assume a starting point $x^{(0)}$ is known

unconstrained minimization methods

- ▶ produce sequence of points $x^{(k)}$, $k = 0, 1, \dots$ with

$$f(x^{(k)}) \rightarrow f^*$$

(we hope)

Solution of an optimization problem

$$\text{minimize } f(x)$$

for $f : \mathcal{D} \rightarrow \mathbf{R}$. x^* is a

- ▶ **global minimizer** if $f(x) \geq f(x^*)$ for all $x \in \mathcal{D}$.
- ▶ **local minimizer** if there is a neighborhood \mathcal{N} around x^* so that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N}$.
- ▶ **isolated local minimizer** if the neighborhood \mathcal{N} contains no other local minimizers.
- ▶ **unique minimizer** if it is the only global minimizer.

Solution of an optimization problem

$$\text{minimize } f(x)$$

for $f : \mathcal{D} \rightarrow \mathbf{R}$. x^* is a

- ▶ **global minimizer** if $f(x) \geq f(x^*)$ for all $x \in \mathcal{D}$.
- ▶ **local minimizer** if there is a neighborhood \mathcal{N} around x^* so that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N}$.
- ▶ **isolated local minimizer** if the neighborhood \mathcal{N} contains no other local minimizers.
- ▶ **unique minimizer** if it is the only global minimizer.

pictures!

First order optimality condition

Theorem

If $x^ \in \mathbf{R}^n$ is a local minimizer of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$, then $\nabla f(x^*) = 0$.*

First order optimality condition

Theorem

If $x^ \in \mathbf{R}^n$ is a local minimizer of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$, then $\nabla f(x^*) = 0$.*

proof: suppose by contradiction that $\nabla f(x^*) \neq 0$. consider points of the form $x_\alpha = x^* - \alpha \nabla f(x^*)$ for $\alpha > 0$. by definition of the gradient,

$$\lim_{\alpha \rightarrow 0} \frac{f(x_\alpha) - f(x^*)}{\alpha} = -\nabla f(x^*)^\top \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0$$

so for any sufficiently small $\alpha > 0$, we have $f(x_\alpha) < f(x^*)$, which contradicts the fact that x^* is a local minimizer.

Second order optimality condition

Theorem

If $x^ \in \mathbf{R}^n$ is a local minimizer of a twice differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$, then $\nabla^2 f(x^*) \succeq 0$.*

Second order optimality condition

Theorem

If $x^ \in \mathbf{R}^n$ is a local minimizer of a twice differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$, then $\nabla^2 f(x^*) \succeq 0$.*

proof: similar to the previous proof. use the fact that the second order approximation

$$f(x_\alpha) \approx f(x^*) + \nabla f(x^*)^\top (x_\alpha - x^*) + \frac{1}{2} (x_\alpha - x^*)^\top \nabla^2 f(x^*) (x_\alpha - x^*)$$

is accurate locally to show a contradiction unless $\nabla^2 f(x^*) \succeq 0$: if not, there is a direction v such that $v^\top \nabla^2 f(x^*) v < 0$. then $f(x + \alpha v) < f(x^*)$ for α arbitrarily small, which contradicts the fact that x^* is a local minimizer.

Outline

Unconstrained minimization

Gradient descent

What functions?

Analysis via Polyak-Lojasiewicz condition

Gradient descent

$$\text{minimize } f(x)$$

idea: go downhill

Algorithm Gradient descent

Given: $f : \mathbf{R}^d \rightarrow \mathbf{R}$, stepsize t , maxiters

Initialize: $x = 0$ (or anything you'd like)

For: $k = 1, \dots, \text{maxiters}$

▶ update x :

$$x \leftarrow x - t \nabla f(x)$$

Gradient descent: choosing a step-size

- ▶ **constant step-size.** $t^{(k)} = t$ (constant)
- ▶ **decreasing step-size.** $t^{(k)} = 1/k$
- ▶ **line search.** try different possibilities for $t^{(k)}$ until objective at new iterate

$$f(x^{(k)}) = f(x^{(k-1)} - t^{(k)} \nabla f(x^{(k-1)}))$$

decreases enough.

tradeoff: line search requires evaluating $f(x)$ (can be expensive)

Line search

define $x^+ = x - t\nabla f(x)$

- ▶ exact line search: find t to minimize $f(x^+)$
- ▶ the **Armijo rule** requires t to satisfy

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2$$

for some $c \in (0, 1)$, e.g., $c = .01$.

Line search

define $x^+ = x - t\nabla f(x)$

- ▶ exact line search: find t to minimize $f(x^+)$
- ▶ the **Armijo rule** requires t to satisfy

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2$$

for some $c \in (0, 1)$, e.g., $c = .01$.

a simple **backtracking line search** algorithm:

- ▶ set $t = 1$
- ▶ if step decreases objective value sufficiently, accept x^+ :

$$f(x^+) \leq f(x) - ct\|\nabla f(x)\|^2 \quad \implies \quad x \leftarrow x^+$$

otherwise, halve the stepsize $t \leftarrow t/2$ and try again

Demo: gradient descent

[https://github.com/stanford-cme-307/demos/blob/main/
gradient-descent.ipynb](https://github.com/stanford-cme-307/demos/blob/main/gradient-descent.ipynb)

Outline

Unconstrained minimization

Gradient descent

What functions?

Analysis via Polyak-Lojasiewicz condition

How well does GD work?

for $x \in \mathbf{R}^n$,

- ▶ $f(x) = x^T x$
- ▶ $f(x) = x^T A x$ for $A \succeq 0$
- ▶ $f(x) = \|x\|_1$ (nonsmooth but differentiable **almost** everywhere)
- ▶ $f(x) = 1/x$ on $x > 0$ (strictly convex but not strongly convex)

<https://github.com/stanford-cme-307/demos/blob/main/gradient-descent-contours.ipynb>

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

A: No.

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

A: No.

Q: Is a stationary point always a global minimum?

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

A: No.

Q: Is a stationary point always a global minimum?

A: No.

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

A: No.

Q: Is a stationary point always a global minimum?

A: No.

Q: ... for convex functions?

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

A: No.

Q: Is a stationary point always a global minimum?

A: No.

Q: ... for convex functions?

A: Yes.

First-order condition

Definition

$x^* \in \mathbf{R}^n$ is a **stationary point** of a differentiable function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ if $\nabla f(x^*) = 0$.

Q: Can a global minimum have a non-zero gradient?

A: No.

Q: Is a stationary point always a global minimum?

A: No.

Q: ... for convex functions?

A: Yes.

$\nabla f(x^*) = 0$ is the **first-order (necessary) condition** for optimality.

Invex function

Definition

A function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is **invex** if for some vector-valued function $\eta : \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}^n$,

$$f(x) - f(u) \geq \eta(x, u)^\top \nabla f(u) \quad \forall u \in \mathbf{R}^n, x \in \text{dom } f$$

Theorem (Craven and Glover, Ben-Israel and Mond)

A function is invex iff every stationary point is a global minimum.

Quadratic approximation

Suppose $f : \mathbf{R} \rightarrow \mathbf{R}$ is twice differentiable. For any $x \in \mathbf{R}$, approximate f about x :

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x).$$

If f is a quadratic function, $\nabla^2 f(x) = H$ is constant.

Quadratic approximation

Suppose $f : \mathbf{R} \rightarrow \mathbf{R}$ is twice differentiable. For any $x \in \mathbf{R}$, approximate f about x :

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x).$$

If f is a quadratic function, $\nabla^2 f(x) = H$ is constant.

Quadratic approximations are useful because quadratics are easy to minimize:

$$\begin{aligned} y^* &= \underset{y}{\operatorname{argmin}} f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T H(y - x) \\ &\implies \nabla f(x) + H(y^* - x) = 0 \\ y^* &= x - H^{-1}(\nabla f(x)). \end{aligned}$$

Quadratic approximation

Suppose $f : \mathbf{R} \rightarrow \mathbf{R}$ is twice differentiable. For any $x \in \mathbf{R}$, approximate f about x :

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x).$$

If f is a quadratic function, $\nabla^2 f(x) = H$ is constant.

Quadratic approximations are useful because quadratics are easy to minimize:

$$\begin{aligned} y^* &= \operatorname{argmin}_y f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T H (y - x) \\ &\implies \nabla f(x) + H(y^* - x) = 0 \\ y^* &= x - H^{-1}(\nabla f(x)). \end{aligned}$$

If we approximate the Hessian of f by $H = \frac{1}{t}I$ for some $t > 0$ and choose x^+ to minimize the quadratic approximation, we obtain the **gradient descent** update with step size t :

$$x^+ = x - t \nabla f(x)$$

Quadratic upper bound

Definition (Smooth)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is **L -smooth** if for all $x, y \in \mathbf{R}$,

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2.$$

Equivalently, assuming the derivatives exist,

- ▶ the operator $\frac{1}{L}\nabla f$ is **L -Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

- ▶ $\nabla^2 f(x) \preceq LI$ for all $x \in \text{dom } f$.

Quadratic upper bound

Definition (Smooth)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is **L -smooth** if for all $x, y \in \mathbf{R}$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2.$$

Equivalently, assuming the derivatives exist,

- ▶ the operator $\frac{1}{L}\nabla f$ is **L -Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

- ▶ $\nabla^2 f(x) \preceq LI$ for all $x \in \text{dom } f$.

Q: For $A \succeq 0$, the quadratic function $f(x) = \frac{1}{2}x^T Ax$ is
?-smooth

Quadratic upper bound

Definition (Smooth)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is **L -smooth** if for all $x, y \in \mathbf{R}$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2.$$

Equivalently, assuming the derivatives exist,

- ▶ the operator $\frac{1}{L} \nabla f$ is **L -Lipschitz continuous**:

$$\|\nabla f(y) - \nabla f(x)\| \leq L \|y - x\|$$

- ▶ $\nabla^2 f(x) \preceq LI$ for all $x \in \text{dom } f$.

Q: For $A \succeq 0$, the quadratic function $f(x) = \frac{1}{2} x^T A x$ is
?-smooth

A: $\lambda_{\max}(A)$ -smooth

Quadratic lower bound

Definition (Strongly convex)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is μ -**strongly convex** if for all $x, y \in \mathbf{R}$,

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Equivalently, assuming the derivatives exist,

- ▶ the operator $\frac{1}{\mu}\nabla f$ is μ -**coercive**:

$$\|\nabla f(y) - \nabla f(x)\| \geq \mu\|y - x\|$$

- ▶ $\nabla^2 f(x) \succeq \mu I$ for all $x \in \text{dom } f$.

Quadratic lower bound

Definition (Strongly convex)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is μ -**strongly convex** if for all $x, y \in \mathbf{R}$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2.$$

Equivalently, assuming the derivatives exist,

- ▶ the operator $\frac{1}{\mu} \nabla f$ is μ -**coercive**:

$$\|\nabla f(y) - \nabla f(x)\| \geq \mu \|y - x\|$$

- ▶ $\nabla^2 f(x) \succeq \mu I$ for all $x \in \text{dom } f$.

Q: For $A \succeq 0$, the quadratic function $f(x) = \frac{1}{2} x^T A x$ is
?-strongly convex

Quadratic lower bound

Definition (Strongly convex)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ is μ -**strongly convex** if for all $x, y \in \mathbf{R}$,

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Equivalently, assuming the derivatives exist,

- ▶ the operator $\frac{1}{\mu}\nabla f$ is μ -**coercive**:

$$\|\nabla f(y) - \nabla f(x)\| \geq \mu\|y - x\|$$

- ▶ $\nabla^2 f(x) \succeq \mu I$ for all $x \in \text{dom } f$.

Q: For $A \succeq 0$, the quadratic function $f(x) = \frac{1}{2}x^T A x$ is
?-strongly convex

A: $\lambda_{\min}(A)$ -strongly convex

Optimizing the upper bound

start at $x^{(0)}$. suppose f is L -smooth, so for all $y \in \mathbf{R}$,

$$f(y) \leq f(x^{(0)}) + \nabla f(x)^T (y - x^{(0)}) + \frac{L}{2} \|y - x^{(0)}\|^2$$

let's choose next iterate $x^{(1)}$ to minimize this upper bound:

$$\begin{aligned} x^{(1)} &= \underset{y}{\operatorname{argmin}} f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2 \\ &\implies \nabla f(x^{(0)}) + L(x^{(1)} - x^{(0)}) = 0 \\ x^{(1)} &= x^{(0)} - \frac{1}{L} \nabla f(x^{(0)}) \end{aligned}$$

Optimizing the upper bound

start at $x^{(0)}$. suppose f is L -smooth, so for all $y \in \mathbf{R}$,

$$f(y) \leq f(x^{(0)}) + \nabla f(x)^T (y - x^{(0)}) + \frac{L}{2} \|y - x^{(0)}\|^2$$

let's choose next iterate $x^{(1)}$ to minimize this upper bound:

$$\begin{aligned} x^{(1)} &= \operatorname{argmin}_y f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2 \\ &\implies \nabla f(x^{(0)}) + L(x^{(1)} - x^{(0)}) = 0 \\ x^{(1)} &= x^{(0)} - \frac{1}{L} \nabla f(x^{(0)}) \end{aligned}$$

- ▶ **gradient descent** update with step size $t = \frac{1}{L}$
- ▶ lower bound ensures true optimum can't be too far away...

Outline

Unconstrained minimization

Gradient descent

What functions?

Analysis via Polyak-Lojasiewicz condition

Some important functions

for $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $x \in \mathbf{R}^n$,

- ▶ **Quadratic loss.** $\|Ax - b\|^2$
- ▶ **Logistic loss.** $f(x) = \sum_{i=1}^m \log(1 + \exp(b_i a_i^T x))$
where a_i is i th row of A

Some important functions

for $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $x \in \mathbf{R}^n$,

- ▶ **Quadratic loss.** $\|Ax - b\|^2$
- ▶ **Logistic loss.** $f(x) = \sum_{i=1}^m \log(1 + \exp(b_i a_i^T x))$
where a_i is i th row of A

Q: Which of these are smooth? Under what conditions?

Some important functions

for $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $x \in \mathbf{R}^n$,

- ▶ **Quadratic loss.** $\|Ax - b\|^2$
- ▶ **Logistic loss.** $f(x) = \sum_{i=1}^m \log(1 + \exp(b_i a_i^T x))$
where a_i is i th row of A

Q: Which of these are smooth? Under what conditions?

A: Both.

Some important functions

for $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $x \in \mathbf{R}^n$,

- ▶ **Quadratic loss.** $\|Ax - b\|^2$
- ▶ **Logistic loss.** $f(x) = \sum_{i=1}^m \log(1 + \exp(b_i a_i^T x))$
where a_i is i th row of A

Q: Which of these are smooth? Under what conditions?

A: Both.

Q: Which of these are strongly convex? Under what conditions?

Some important functions

for $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $x \in \mathbf{R}^n$,

- ▶ **Quadratic loss.** $\|Ax - b\|^2$
- ▶ **Logistic loss.** $f(x) = \sum_{i=1}^m \log(1 + \exp(b_i a_i^T x))$
where a_i is i th row of A

Q: Which of these are smooth? Under what conditions?

A: Both.

Q: Which of these are strongly convex? Under what conditions?

A: Quadratic loss is strongly convex if A is rank n . Logistic loss is strongly convex on a compact domain if A is rank n .

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

Theorem

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \rightarrow \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

source: [Karimi, Nutini, and Schmidt (2016)]

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

Theorem

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \rightarrow \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

source: [Karimi, Nutini, and Schmidt (2016)]

so logistic loss (on a compact set) and quadratic loss are Polyak-Lojasiewicz even when $m < n$

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

Theorem

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \rightarrow \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

source: [Karimi, Nutini, and Schmidt (2016)]

so logistic loss (on a compact set) and quadratic loss are Polyak-Lojasiewicz even when $m < n$

Q: Are all Polyak-Lojasiewicz functions convex?

The Polyak-Lojasiewicz condition

Definition (Polyak-Lojasiewicz condition)

A function $f : \mathbf{R} \rightarrow \mathbf{R}$ satisfies the **Polyak-Lojasiewicz condition** if

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$$

Theorem

Suppose $f(x) = g(Ax)$ where $g : \mathbf{R}^m \rightarrow \mathbf{R}$ is strongly convex and $A : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is linear. Then f is Polyak-Lojasiewicz.

source: [Karimi, Nutini, and Schmidt (2016)]

so logistic loss (on a compact set) and quadratic loss are Polyak-Lojasiewicz even when $m < n$

Q: Are all Polyak-Lojasiewicz functions convex?

A: No. A river valley is Polyak-Lojasiewicz but not convex.

why use Polyak-Lojasiewicz? Polyak-Lojasiewicz is weaker than strong convexity and yields simpler proofs

PL and invexity

Theorem

Every Polyak-Lojasiewicz function is invex. (That is, any stationary point of a Polyak-Lojasiewicz function is globally optimal.)

PL and invexity

Theorem

Every Polyak-Lojasiewicz function is invex. (That is, any stationary point of a Polyak-Lojasiewicz function is globally optimal.)

proof: if $\nabla f(\bar{x}) = 0$, then

$$0 = \frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(\bar{x}) - f^*) \geq 0$$

$\implies f(\bar{x}) = f^*$ is the global optimum.

strong convexity \implies Polyak-Łojasiewicz

Theorem

If f is μ -strongly convex, then f is μ -Polyak-Łojasiewicz.

strong convexity \implies Polyak-Lojasiewicz

Theorem

If f is μ -strongly convex, then f is μ -Polyak-Lojasiewicz.

proof: minimize the strong convexity condition over y :

$$\begin{aligned}\min_y f(y) &\geq \min_y \left(f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2 \right) \\ f^* &\geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2\end{aligned}$$

Types of convergence

- ▶ objective converges

$$f(x^{(k)}) \rightarrow f^*$$

- ▶ iterates converge

$$x^{(k)} \rightarrow x^*$$

under

- ▶ strong convexity: objective converges \implies iterates converge

proof: use strong convexity with $x = x^*$ and $y = x^{(k)}$:

$$f(x^{(k)}) - f^* \geq \frac{\mu}{2} \|x^{(k)} - x^*\|^2$$

- ▶ Polyak-Lojasiewicz: not necessarily true (x^* may not be unique)

Rates of convergence

- ▶ linear convergence with rate c

$$f(x^{(k)}) - f^* \leq c^k (f(x^{(0)}) - f^*)$$

- ▶ looks like a line on a semi-log plot
 - ▶ example: gradient descent on smooth strongly convex function
- ▶ sublinear convergence
 - ▶ looks slower than a line (curves up) on a semi-log plot
 - ▶ example: $1/k$ convergence

$$f(x^{(k)}) - f^* \leq \mathcal{O}(1/k)$$

- ▶ example: gradient descent on smooth convex function
 - ▶ example: stochastic gradient descent

Gradient descent converges linearly

Theorem

If $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is μ -Polyak-Lojasiewicz, L -smooth, and $x^ = \operatorname{argmin}_x f(x)$ exists, then gradient descent with stepsize L*

$$x^{(k+1)} = x^{(k)} - \frac{1}{L} \nabla f(x^{(k)})$$

converges linearly to f^ with rate $(1 - \frac{\mu}{L})$.*

Gradient descent converges linearly: proof

proof: plug in update rule to L -smoothness condition

$$\begin{aligned}f(x^{(k+1)}) - f(x^{(k)}) &\leq \nabla f(x^{(k)})^T (x^{(k+1)} - x^{(k)}) + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2 \\&\leq \left(-\frac{1}{L} + \frac{1}{2L}\right) \|\nabla f(x^{(k)})\|^2 \\&\leq -\frac{1}{2L} \|\nabla f(x^{(k)})\|^2 \\&\leq -\frac{\mu}{L} (f(x^{(k)}) - f^*) \triangleright (\text{using PL})\end{aligned}$$

decrement proportional to error \implies linear convergence:

$$\begin{aligned}f(x^{(k)}) - f^* &\leq \left(1 - \frac{\mu}{L}\right) (f(x^{(k-1)}) - f^*) \\&\leq \left(1 - \frac{\mu}{L}\right)^k (f(x^{(0)}) - f^*)\end{aligned}$$

Practical convergence

- ▶ Gradient descent with optimal stepsize converges even faster.

$$f(x^{(k+1)}) = \inf_{\alpha} f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)} - \frac{1}{L} \nabla f(x^{(k)}))$$

Practical convergence

- ▶ Gradient descent with optimal stepsize converges even faster.

$$f(x^{(k+1)}) = \inf_{\alpha} f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)} - \frac{1}{L} \nabla f(x^{(k)}))$$

- ▶ Local vs global convergence

Quiz

- ▶ A strongly convex function always satisfies the Polyak-Lojasiewicz condition
 - A. true
 - B. false
- ▶ Suppose $f : \mathbf{R} \rightarrow \mathbf{R}$ is L -smooth and satisfies the Polyak-Lojasiewicz condition. Then any stationary point $\nabla f(x) = 0$ of f is a global optimum:
 $f(x) = \operatorname{argmin}_y f(y) =: f^*$.
 - A. true
 - B. false
- ▶ Suppose $f : \mathbf{R} \rightarrow \mathbf{R}$ is L -smooth and satisfies the Polyak-Lojasiewicz condition. Then gradient descent on f converges linearly from any starting point.
 - A. true
 - B. false