CSIE5410 Optimization algorithms

Lecture 2: Convexity

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

20.09.2018

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

This lecture aims to answer the following questions.

- What is convexity?
- Why is convexity important?
- What are the implications of convexity?

Caveat. This lecture is relatively dry...

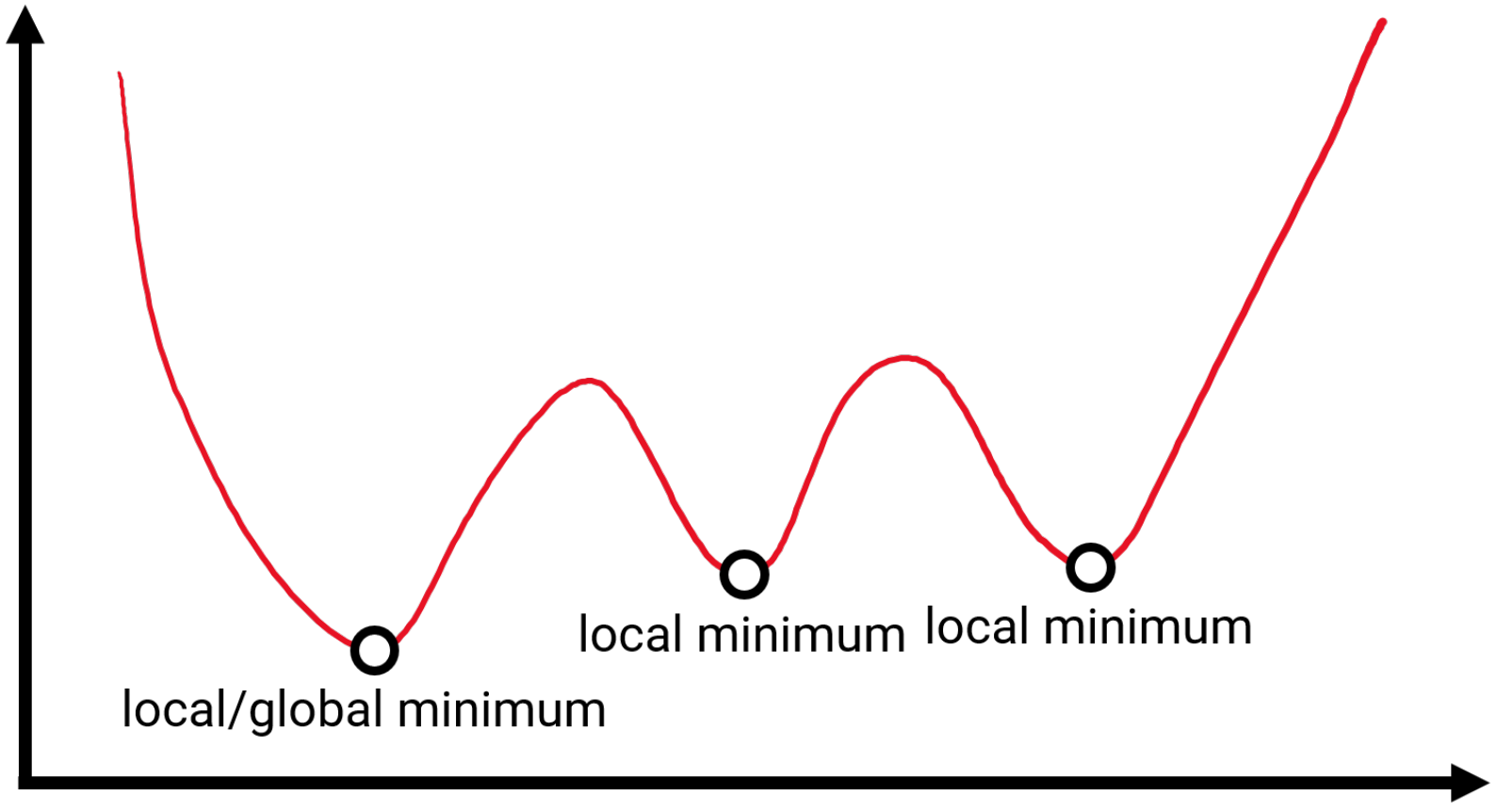
Recommended reading

- Yu. Nesterov. 2004. *Introductory Lectures on Convex Optimization*. (Chapter 1–2).
- R. T. Rockafellar. 1970. *Convex Analysis*. (Chapter 4, 23–25).
- B. T. Polyak. 1987. *Introduction to Optimization*. (Section 1.1–1.3).
- D. P. Bertsekas. 2016. *Nonlinear Programming*. (Appendix A–B).

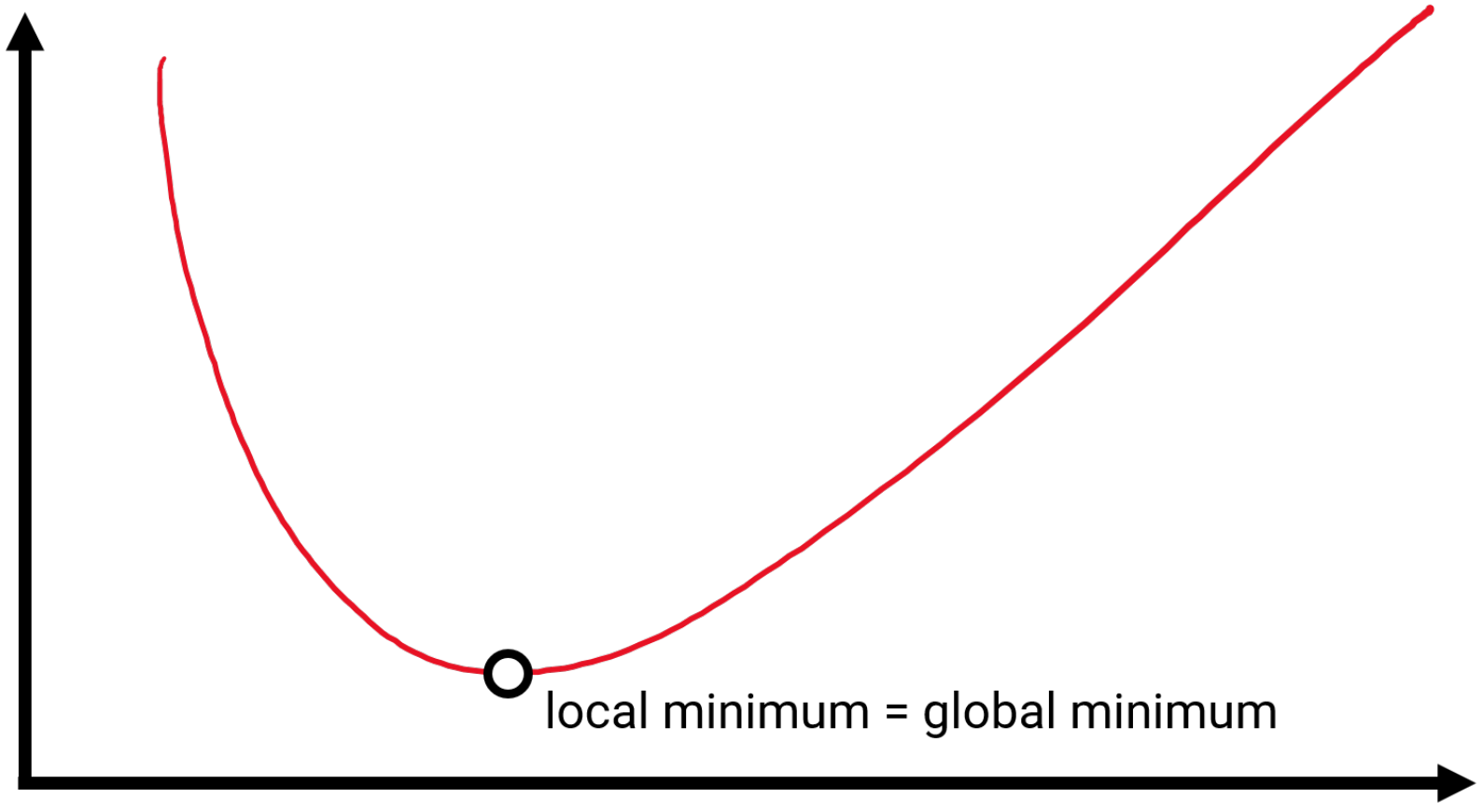
Table of contents

Why is convexity important?

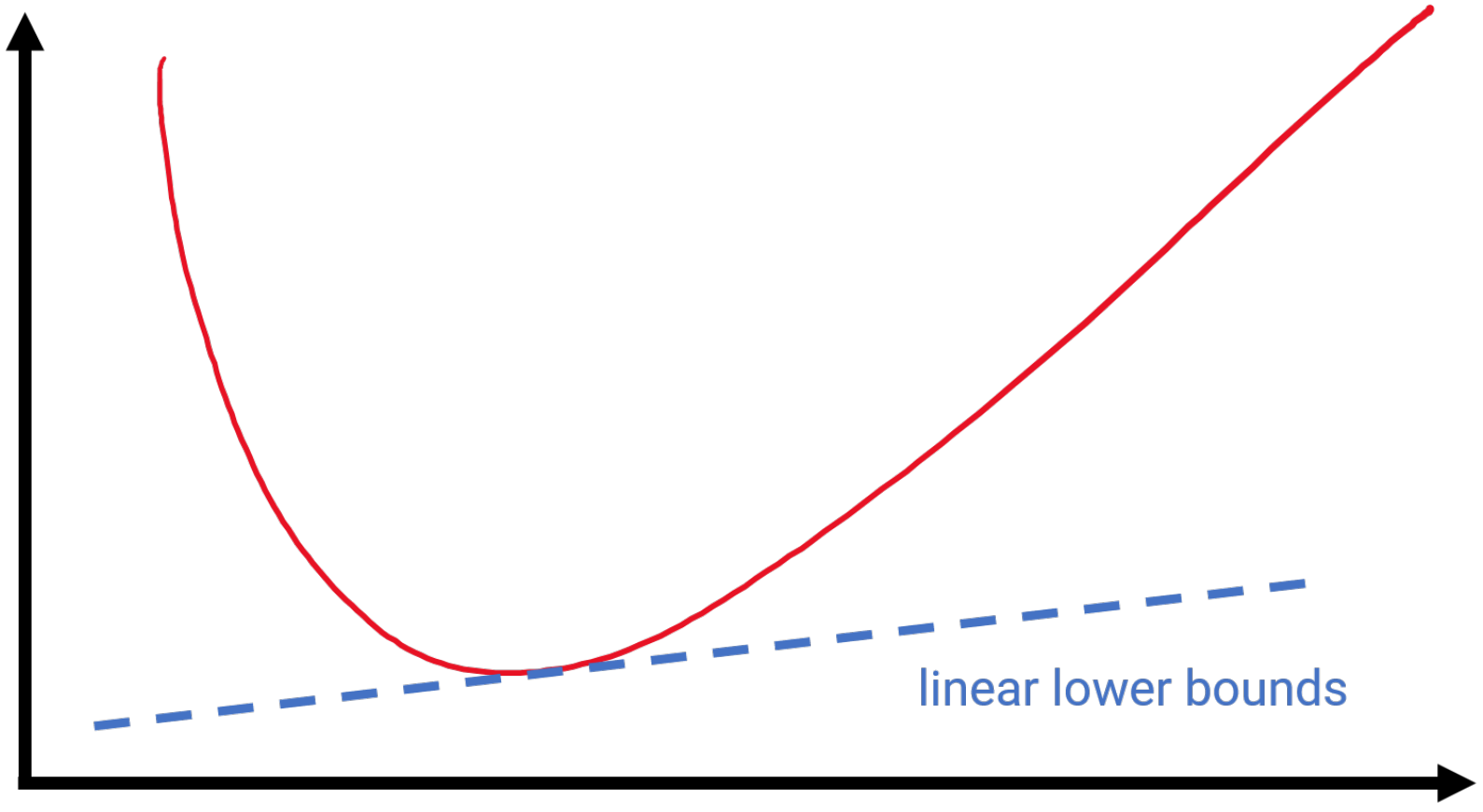
Local vs. global minima



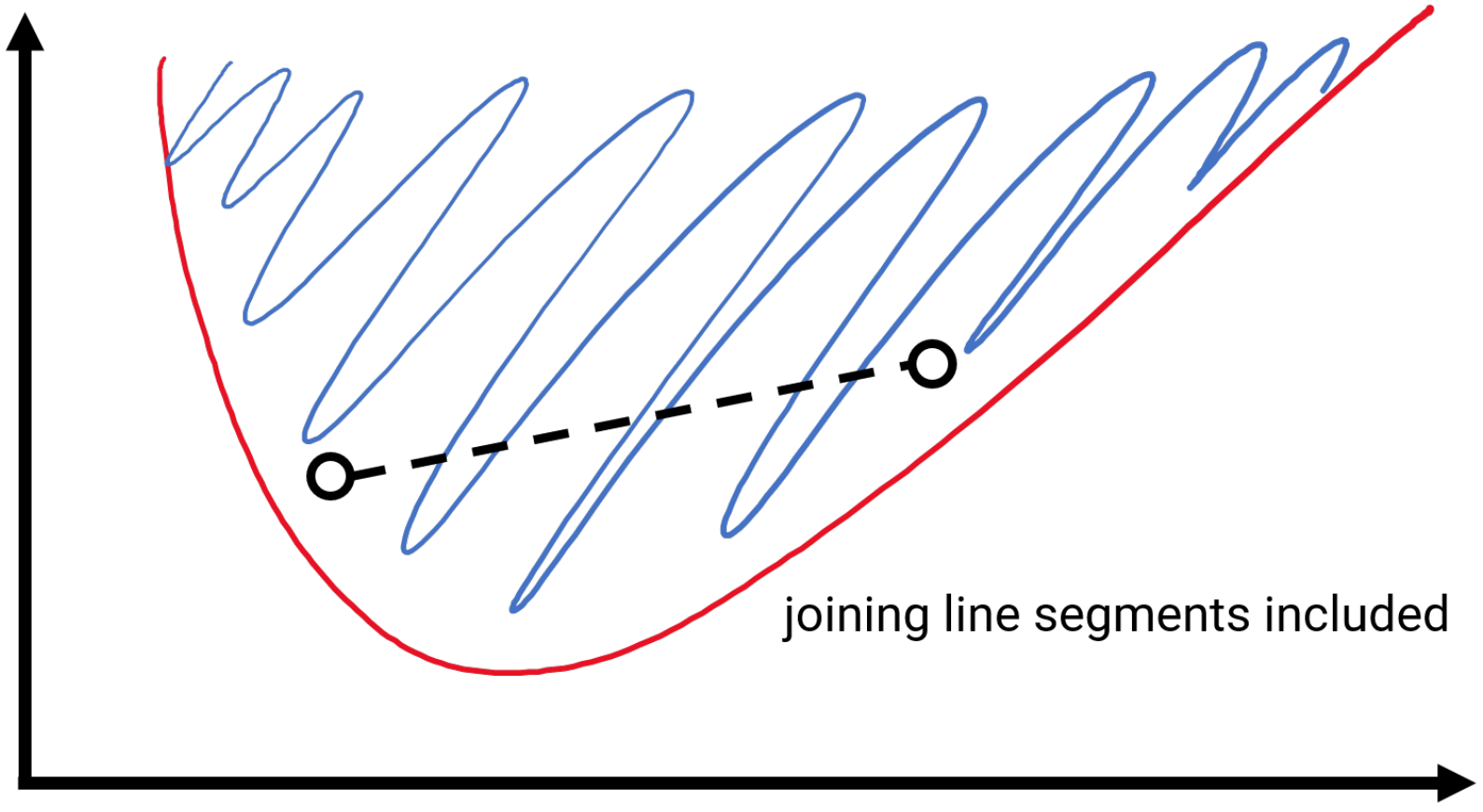
Illustrations of convexity (1/4)



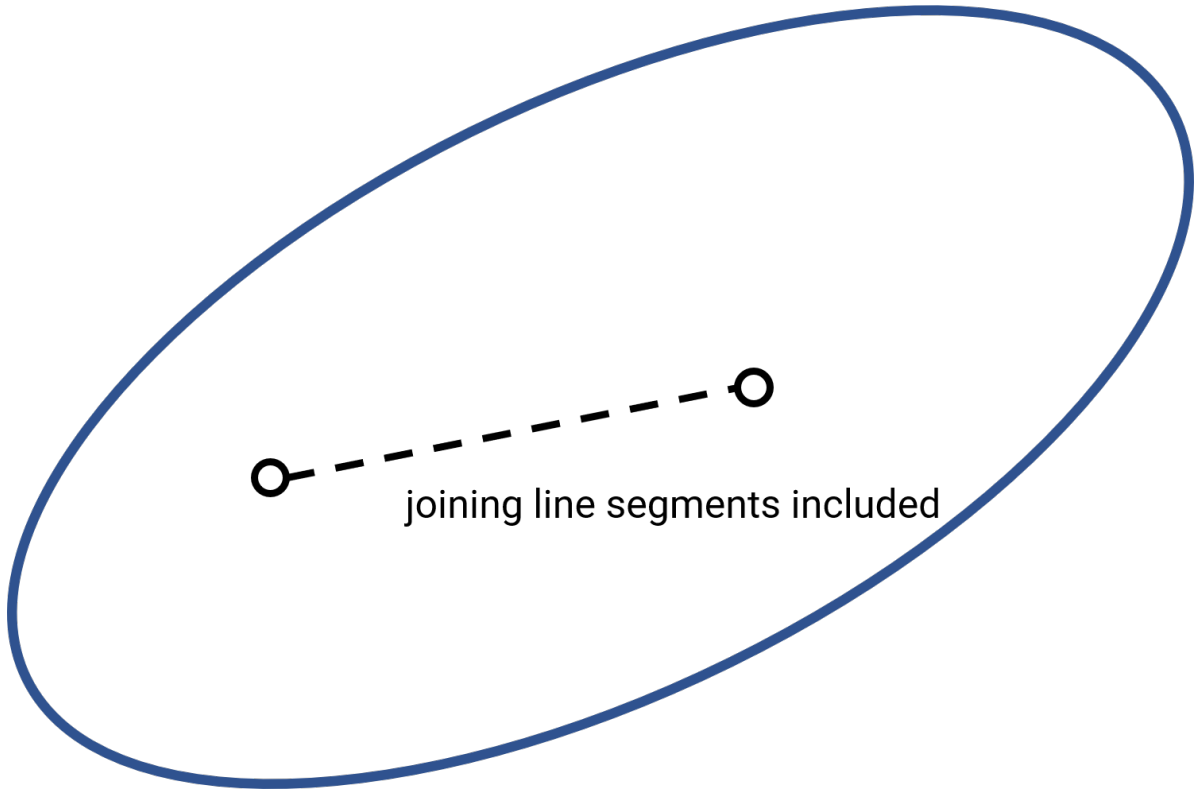
Illustrations of convexity (2/4)



Illustrations of convexity (3/4)



Illustrations of convexity (4/4)



Convex vs. non-convex optimization problems

$$f^\star = \min_x \{\, f(x) \mid x \in \mathcal{X} \,\}.$$

Definition. If both f and $\mathcal{X}$ are convex, we say that this is a *convex* problem; otherwise, we say that this is a *non-convex* problem.

Bad news for non-convexity (1/2)

Theorem. Computing the global minimum of a *non-convex* optimization problem is NP-hard.

Proof. Consider the subset-sum problem, which is NP-complete: Let $a \in \mathbb{N}^p$ and $b \in \mathbb{N}$; is there any $x \in \{0, 1\}^p$, such that $\langle a, x \rangle = b$? Now consider the *non-convex* optimization problem:

$$f^\star = \min \left\{ \left(\langle a, x \rangle - b \right)^2 + \sum_{i=1}^{p} x_i (1 - x_i) \;\middle|\; x \in [0, 1]^p \right\}.$$

The answer to the subset-sum problem is yes, if and only if $f^\star = 0$.

K. G. Murty and S. N. Kabadi. 1987. Some NP-complete problems in quadratic and nonlinear programming.

Bad news for non-convexity (2/2)

Theorem. Checking whether a point achieves a local minimum of a *non-convex* optimization problem is NP-hard.

Proof. The 3-satisfiability (3-SAT) problem reduces to one such problem.

Theorem. If the *non-convex* objective function is not differentiable, then even to find a descent direction is NP-hard.

Proof. The knapsack problem reduces to one such problem.

P. M. Pardalos and G. Schnitger. 1988. Checking local optimality in constrained quadratic programming is NP-hard.
Yu. Nesterov. 2013. Gradient methods for minimizing composite functions.

Addressing non-convexity (1/4)

Theorem. If the objective function has bounded derivatives up to the k-th order for some $k \in \mathbb{N}$, and some technical conditions hold, then a *non-convex* optimization problem can be solved in polynomial time given a noisy zeroth-order oracle.

Theorem. Let $y_1, \ldots, y_n \in \mathbb{R}^p$. The principal component analysis (PCA) problem

$$x^\star \in \arg\min_x \left\{ -\left\langle x, \left(\sum_{i=1}^n y_i y_i^{\mathrm{T}} \right) x \right\rangle \,\middle|\, x \in \mathbb{R}^p, \|x\|_2 = 1 \right\},$$

which is *non-convex*, can be solved in polynomial time.

J. Dippon. 2003. Accelerated randomized stochastic optimization.

O. Shamir. 2016. Convergence of stochastic gradient descent for PCA.

Addressing non-convexity (2/4)

Example. Consider the *phase retrieval* problem, which asks one to recover $x^\natural \in \mathbb{R}^p$, given $a_1, \ldots, a_n \in \mathbb{R}^p$, and

$$y_i := |\langle a_i, x^\natural \rangle|^2, \quad i = 1, \ldots, n.$$

A natural approach is to consider the *non-convex* optimization problem:

$$\hat{x} \in \arg\min_x \left\{ 0 \;\middle|\; x \in \mathbb{R}^p, y_i = |\langle a_i, x \rangle|^2 \; \forall i \right\}.$$

Y. Chen and E. J. Candès. 2016. Solving random quadratic systems of equations is nearly as easy as solving linear systems.

Addressing non-convexity (3/4)

Proposition. Solving the non-convex optimization problem is NP-hard.

Proof. Consider the NP-complete *stone problem*: Let $w \in \mathbb{N}^p$. Check if there exists some $x := (x_1, \ldots, x_p) \in \{-1, +1\}^p$, such that $\langle w, x \rangle = 0$.

The answer is yes, if and only if the optimization problem

$$\hat{x} \in \arg\min_x \left\{ 0 \;\middle|\; x \in \mathbb{R}^p, \langle w, x \rangle^2 = 0, x_i^2 = 1 \; \forall i = 1, \ldots, p \right\}$$

has a solution.

Y. Chen and E. J. Candès. 2016. Solving random quadratic systems of equations is nearly as easy as solving linear systems.

Addressing non-convexity (4/4)

Theorem. Suppose that A_i's satisfy *some technical conditions*. Consider the *convex* optimization problem:

$$\hat{X} \in \arg\min_{X} \left\{ \|X\|_{S^1} \;\middle|\; X \in \mathbb{R}^{p\times p}, X \geq 0, y_i = \mathrm{Tr}(A_i X)\ \forall i \right\}.$$

Then it holds that $\hat{X} = x^\natural (x^\natural)^{\mathrm{T}}$.

Remark. Recall the low-rank matrix recovery problem in Lecture 1.

E. J. Candès *et al.* 2012. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming.

Good news for convexity

Theorem. A convex optimization problem can be solved in polynomial time, if any of the following holds.

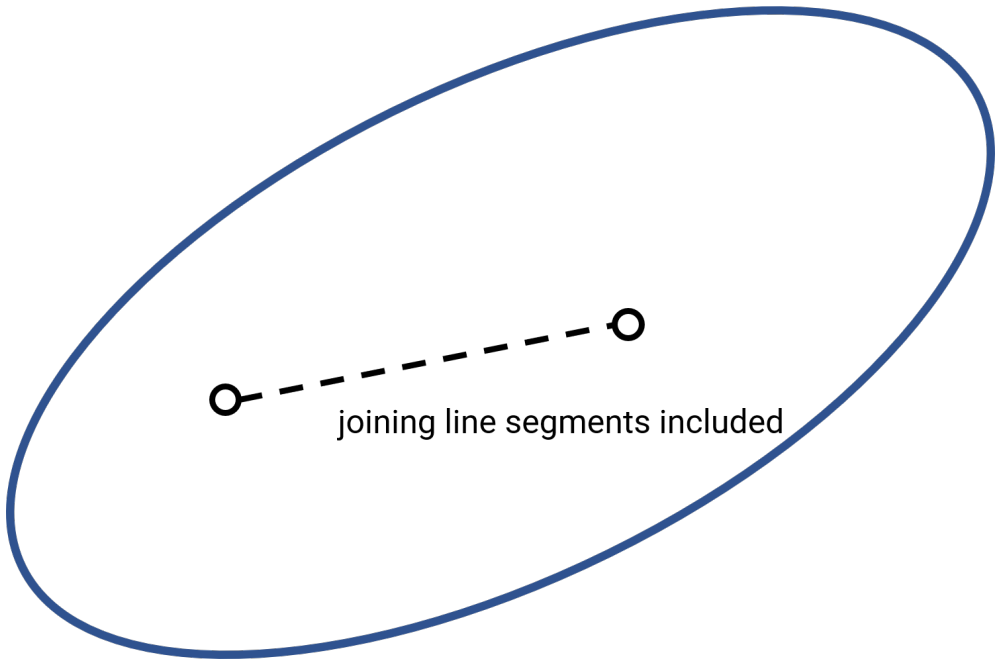
- A membership oracle is available.
- A zeroth-order oracle and a membership oracle are available.
- A separation oracle is available.

Definition. For any point x, a separation oracle for a closed convex set $\mathcal{X}$ returns 'yes' if $x \in \mathcal{X}$, and a hyperplane that separates x and $\mathcal{X}$ otherwise.

A. T. Kalai and S. Vempala. 2006. Simulated annealing for convex optimization.
Y. T. Lee *et al.* 2018. Efficient convex optimization with membership oracles.
D. Bertsimas and S. Vempala. 2004. Solving convex programs by random walks.

Basic concepts in convex analysis

Convex set



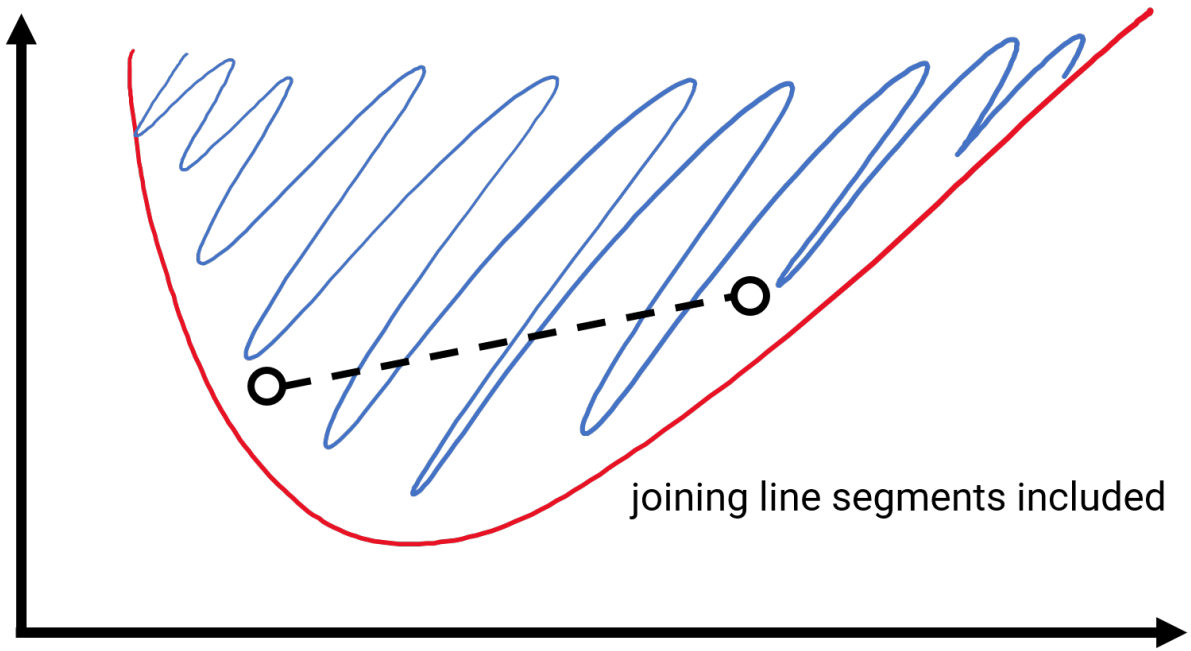
Definition. We say that a set $\mathcal{X}$ is convex, if and only if for any $x, y \in \mathcal{X}$,

$$\alpha x + (1 - \alpha) y \in \mathcal{X}, \quad \forall \alpha \in]0, 1[.$$

Exercise

1. $\mathcal{X} := [0, 1]$.
2. $\mathcal{X} :=]0, 1[$.
3. $\mathcal{X} := \{ x \mid x \in \mathbb{R}^p, \|x\|_2 \leq 1 \}$.
4. $\mathcal{X} := \{ x \mid x \in \mathbb{R}^p, \|x\|_2 = 1 \}$.
5. $\mathcal{X} := \{ x \mid x \in \mathbb{R}^p, Ax = b \}$, for some $A \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^n$.
6. $\mathcal{X} := \left\{ (x, y) \,\middle|\, x \in [0, 1], y \in [0, 1], x^2 + y^2 \geq 1 \right\}$.
7. $\mathcal{X} := \left\{ (x, y) \,\middle|\, x \in [0, 1], y \in \mathbb{R}, y \geq x^2 \right\}$.
8. $\mathcal{X} := \{ X \mid X \in \mathbb{R}^{p \times p}, X \geq 0, \mathrm{Tr}(X) = 1 \}$.

Convex function



Definition. The *epigraph* of a function $f : \mathcal{X} \to \mathbb{R} \cup \{ +\infty \}$ is defined as $\operatorname{epi} f := \{ (x, t) \mid x \in \mathcal{X}, t \in \mathbb{R}, t \geq f(x) \}$.

Definition. We say that the function f is convex, if and only if $\operatorname{epi} f$ is convex.

Equivalent definition (1/2)

Proposition. A function $f : \mathcal{X} \to \mathbb{R} \cup \{ +\infty \}$ is convex, if and only if $\mathcal{X}$ is convex, and for any $x, y \in \mathcal{X}$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall \alpha \in]0,1[.$$

Proof. ($\Rightarrow$) By convexity of f, we write

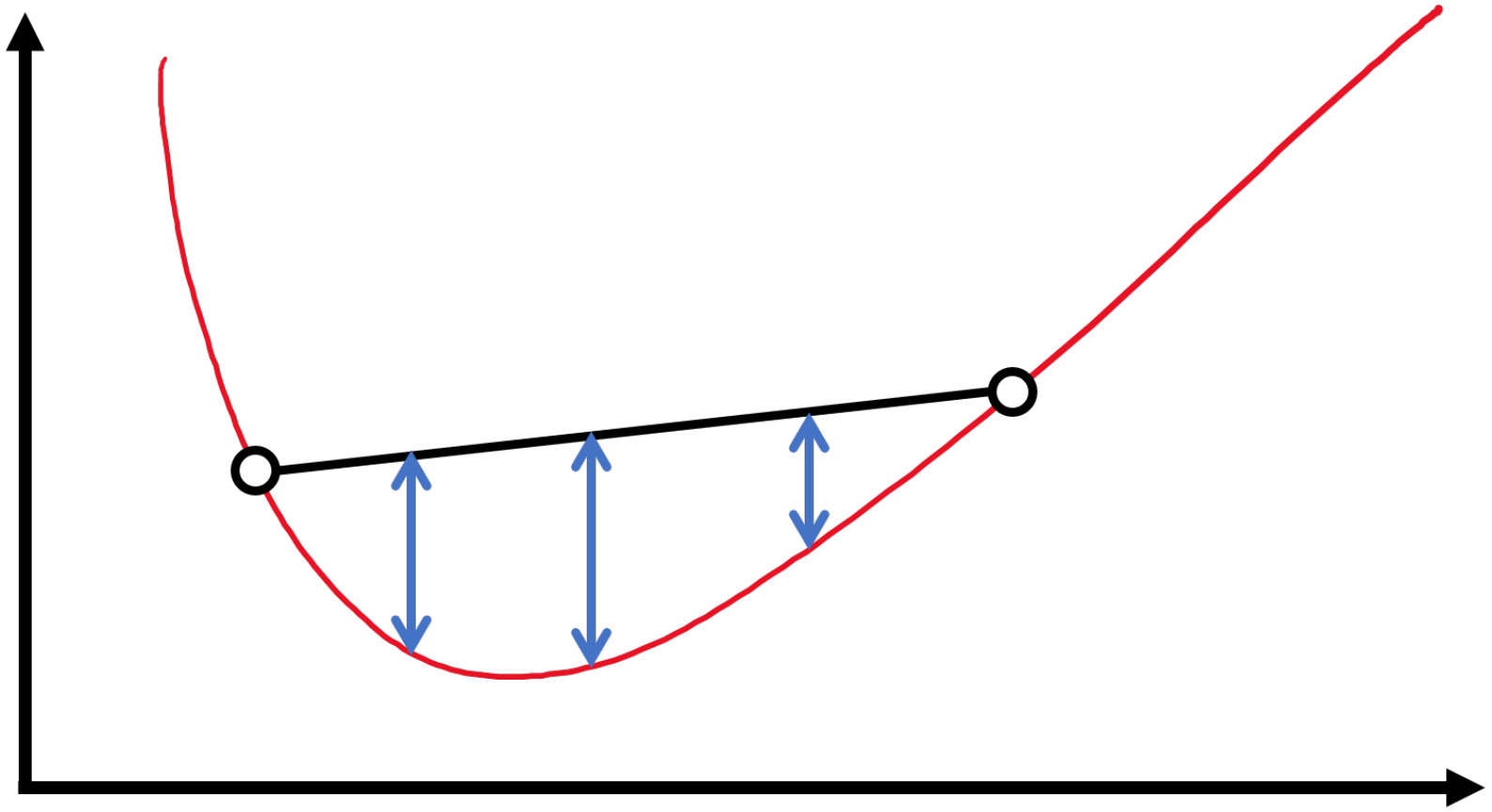
$$\alpha(x, f(x)) + (1-\alpha)(y, f(y)) \in \operatorname{epi} f.$$

($\Leftarrow$) For any (x, t_x), $(y, t_y) \in \operatorname{epi} f$, we have

$$\alpha x + (1-\alpha)y \in \mathcal{X},$$
$$\alpha t_x + (1-\alpha) t_y \geq \alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha)y).$$

Therefore, $\alpha(x, t_x) + (1-\alpha)(y, t_y) \in \operatorname{epi} f$.

Equivalent definition (2/2)

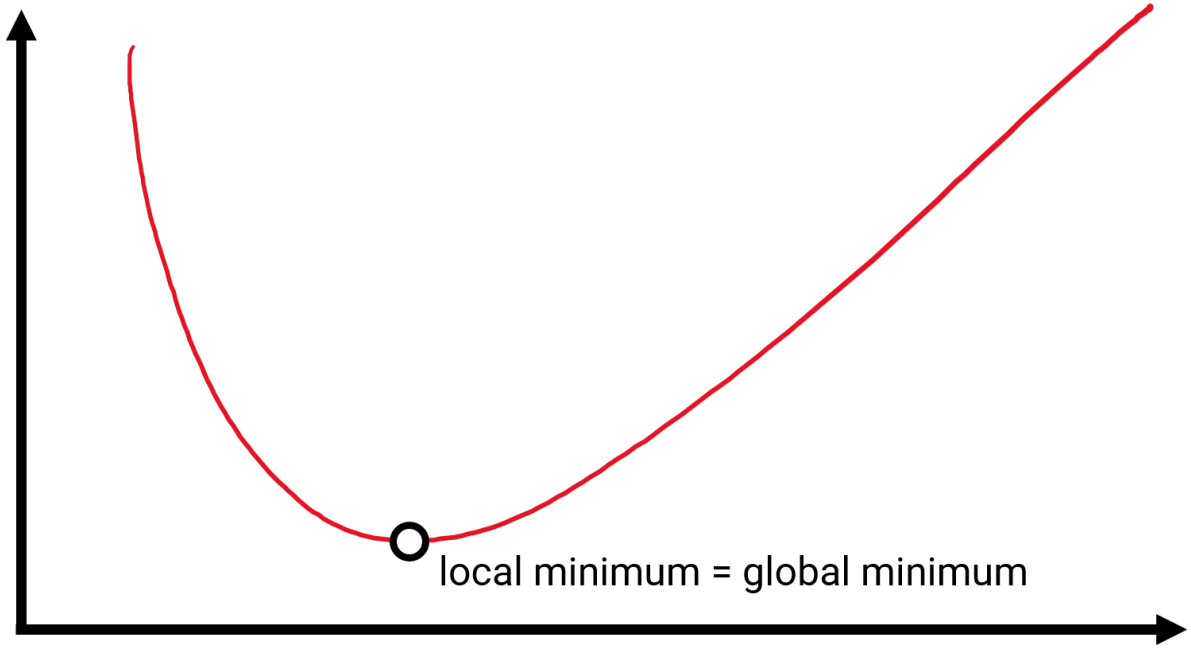


Exercise

Let $\mathcal{X} \subseteq \mathbb{R}^p$ be a convex set, $\mathcal{Y} \subset \mathbb{R}^p$ be a non-convex set, $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, and $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$.

1. $f(x) := 1$.
2. $f(x) := x^2$.
3. $f(x) := x^3$.
4. $g(x) := \|x\|_2^2$.
5. $g(x) := \|x\|_1$.
6. $g(x) := \|Ax - b\|_2^2 + \|x\|_1$, for some $A \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$.
7. $g(x) := \chi_{\mathcal{X}}(x)$ (*characteristic function*).
8. $g(x) := \chi_{\mathcal{Y}}(x)$.

Implications (1/2)



Proposition. Let $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then every local minimum of f is also a global minimum.

Proof

Proof. Let x be a local minimizer of f on $\mathcal{X}$. Suppose that there exists some global minimizer $x^\star \neq x \in \mathcal{X}$ such that $f(x^\star) < f(x)$. Then we have

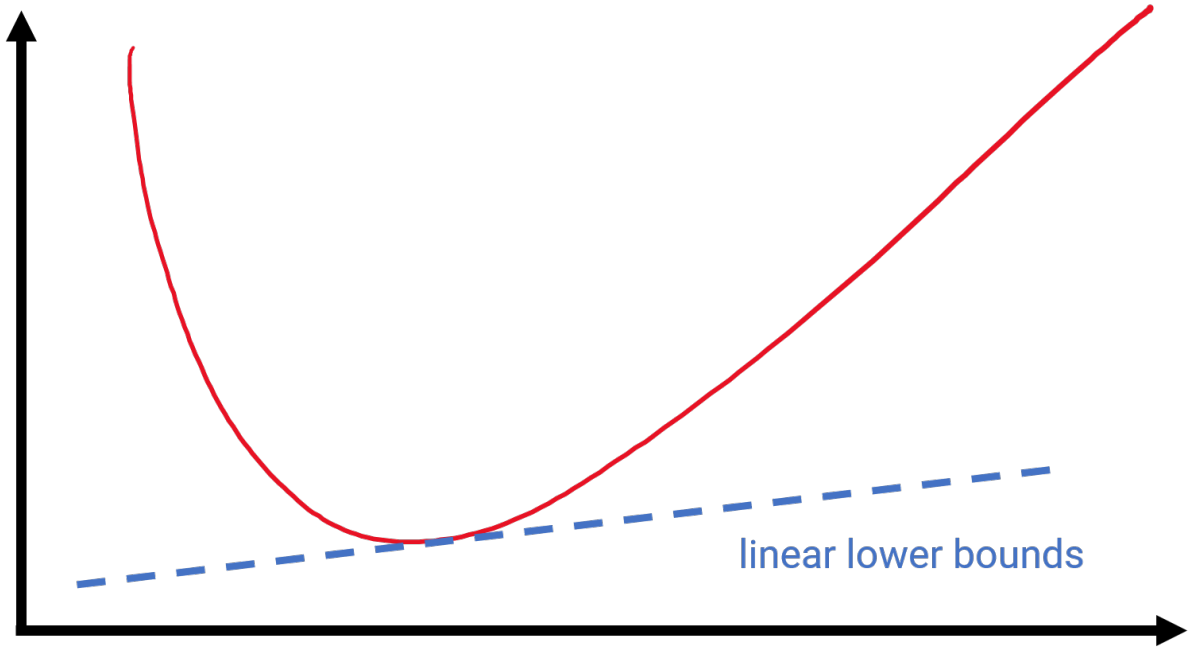
$$f(\alpha x + (1-\alpha)x^\star) \leq \alpha f(x) + (1-\alpha) f(x^\star) < f(x), \quad \forall \alpha \in]0,1[.$$

However, since x is a local minimizer, we have

$$f(\alpha x + (1-\alpha)x^\star) \geq f(x), \quad \text{for } \alpha \in]0,1[\text{ close enough to } 1,$$

a contradiction.

Implications (2/2)



Proposition. Let $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be *differentiable*. For any $x \in \mathcal{X}$, there exists some affine function f_x, such that $f_x(x) = f(x)$, and

$$f(y) \geq f_x(y), \quad \forall y \in \mathcal{X}.$$

Detour: Differentiability

Definition. We say that a function $f : \mathcal{X} \subset \mathbb{R}^p \to [-\infty, +\infty]$ is differentiable at $x \in \mathcal{X}$ where $f(x)$ is finite, if and only if there exists some $g_x \in \mathbb{R}^p$, such that

$$f(y) = f(x) + \langle g_x, y - x \rangle + o(\|y - x\|_2).$$

That is,

$$\lim_{y \to x} \frac{f(y) - f(x) - \langle g_x, y - x \rangle}{\|y - x\|_2} = 0.$$

Definition. If such a vector g_x exists, we call it the *gradient* of f at x, and write it as $\nabla f(x)$.

Detour: Computation of a gradient

Proposition. If f is differentiable at x, we have

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_p}(x) \right).$$

Proof. Let $\{ e_1, \ldots, e_p \}$ be the canonical basis of $\mathbb{R}^p$. Then we have

$$\lim_{\lambda \to 0} \frac{f(x + \lambda e_i) - f(x) - \lambda \langle \nabla f(x), e_i \rangle}{\lambda} = 0, \quad \forall i = 1, \ldots, p.$$

That is,

$$\frac{\partial f}{\partial x_i}(x) = \langle \nabla f(x), e_i \rangle, \quad \forall i = 1, \ldots, p.$$

Existence of linear lower bounds: Proof

Theorem. Let $f : \mathcal{X} \to [-\infty, \infty]$ be a convex function differentiable at x. Then we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall y \in \mathcal{X}.$$

Proof. For any $x, y \in \mathcal{X}$, by convexity, we have

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y), \quad \forall \alpha \in]0, 1[.$$

That is,

$$f(x + \alpha(y - x)) - f(x) \leq \alpha \left(f(y) - f(x)\right), \quad \forall \alpha \in]0, 1[.$$

Dividing both sides by α, and letting $\alpha \downarrow 0$, the theorem follows.

Optimality condition

Definition. We say that a set $\mathcal{X}$ is closed, if every converging sequence in $\mathcal{X}$ converges to a point in $\mathcal{X}$.

Theorem. Let f be convex and differentiable on $\mathbb{R}^p$. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be closed and convex. Then we have

$$x^\star \in \arg\min_x \{\, f(x) \mid x \in \mathcal{X} \,\},$$

if and only if

$$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Proof of the optimality condition

($\Leftarrow$) We have

$$f(y) \geq f(x^\star) + \langle \nabla f(x^\star), y - x^\star \rangle \geq f(x^\star), \quad \forall y \in \mathcal{X}.$$

($\Rightarrow$) Suppose there exists some $z \in \mathcal{X}$, such that

$$\langle \nabla f(x^\star), z - x^\star \rangle < 0.$$

Consider the function $\varphi(\alpha) := f(x^\star + \alpha(z - x^\star))$, $\alpha \in [0, 1]$. Then $\varphi(0) = f(x^\star)$, and

$$\varphi'(0) = \langle \nabla f(x^\star), z - x^\star \rangle < 0.$$

Therefore, we have $f(x^\star + \alpha(z - x^\star)) < f(x^\star)$ with small enough $\alpha > 0$, a contradiction.

Other characterizations of convexity

Characterization by linear lower bounds (1/2)

Theorem. A function f differentiable on $\mathbb{R}^p$ is convex, if and only if

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

Remark. Therefore, one may consider the inequality above as the definition of convexity (with differentiability). Nesterov adopted an interesting axiomatic approach to *derive* this alternative definition.

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

Characterization by linear lower bounds (2/2)

Proof. We have proved the "only if" direction.

For any $x, y \in \mathbb{R}^p$, we define $x_\alpha := \alpha x + (1-\alpha)y$ for every $\alpha \in [0,1]$. Then we write

$$f(y) \geq f(x_\alpha) + \langle \nabla f(x_\alpha), y - x_\alpha \rangle = f(x_\alpha) + \alpha \langle \nabla f(x_\alpha), y - x \rangle$$
$$f(x) \geq f(x_\alpha) + \langle \nabla f(x_\alpha), x - x_\alpha \rangle = f(x_\alpha) + (1-\alpha) \langle \nabla f(x_\alpha), x - y \rangle .$$

Multiply the first inequality by $(1-\alpha)$, and the second inequality by α, and sum up the two inequalities. Then the theorem follows.

Monotonicity of the gradient (1/2)

Theorem. A function f continuously differentiable on $\mathbb{R}^p$ is convex, if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^p.$$

Remark. Therefore, we say that the gradient is a *monotone operator*.

Proof. ($\Rightarrow$) By convexity, we write

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

Summing up the two inequalities, we get the desired result.

Monotonicity of the gradient (2/2)

Proof continued. ($\Leftarrow$) Define $\varphi(t) := f(x + t(y - x))$ for all $t \in [0, 1]$. Then we have $\varphi'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$. Recall the Taylor formula with the integral remainder

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

We write

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

Positive semidefiniteness of the Hessian (1/2)

Definition. Let f be a function twice continuously differentiable on $\mathbb{R}^p$. Its *Hessian matrix* at $x \in \mathbb{R}^p$ is given by $\nabla^2 f(x) := (H_{i,j}) \in \mathbb{R}^{p \times p}$, where

$$H_{i,j} := \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \quad \forall i, j.$$

Theorem. A function f twice continuously differentiable on $\mathbb{R}^p$ is convex, if and only if

$$\nabla^2 f(x) \geq 0, \quad \forall x \in \mathbb{R}^p.$$

Positive semidefiniteness of the Hessian (2/2)

Proof. For the one-dimensional case ($p = 1$), we have $f'' > 0$ if and only if f' is monotonically increasing. The theorem follows.

In general, for any $x, y \in \mathbb{R}^p$, define

$$\varphi(t; x, y) := f(x + t(y - x)), \quad \forall t \in \mathbb{R}.$$

Then f is convex, if and only if φ is convex for any x, y (Why?). By the discussion for the one-dimensional case above, f is convex if and only if

$$\varphi''(t; x, y) = \langle y - x, \left[\nabla^2 f(x + t(y - x))\right](y - x)\rangle \geq 0, \quad \forall x, y, t.$$

The theorem follows.

Example: Logistic regression

Let $(x_i, y_i) \in \mathbb{R}^p \times \{ -1, +1 \}$, $i = 1, \ldots, n$, be image-label pairs. Logistic regression corresponds to solving the optimization problem

$$w^\star \in \arg\min_w \{ f(w) \mid w \in \mathbb{R}^p \},$$

where

$$f(w) := \frac{1}{n} \sum_{i=1}^{n} \log \left[1 + \exp \left(-y_i \langle x_i, w \rangle \right) \right].$$

Question. Is f a convex function?

Subgradient & subdifferential

What if the function is not differentiable?

Idea. Extend the linear-lower-bounds characterization.

Definition. Let $f : \mathbb{R}^p \to]-\infty, +\infty]$ be convex. We say that a vector $g_x \in \mathbb{R}^p$ is a *subgradient* of f at a point $x \in \mathcal{X}$, if and only if

$$f(y) \geq f(x) + \langle g_x, y - x \rangle, \quad \forall y \in \mathbb{R}^p.$$

The set of all such g_x's is called the *subdifferential* of f at x, and is written as $\partial f(x)$.

Remark. We will study the notions in a few weeks.

Conclusions

Summary

Several characterizations of convexity:

- Convexity of a set: $\alpha x + (1-\alpha)y \in \mathcal{X}$.
- Convexity of a function: $\operatorname{epi} f$ being convex.
- Equivalent definition:
  $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$.
- Monotonicity of the gradient: $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$.
- Positive semi-definiteness of the Hessian: $\nabla^2 f(x) \geq 0$.

Optimality condition: $\langle \nabla f(x^\star), x - x^\star \rangle \geq 0$.

Next lecture

- Smoothness & strong convexity
- Gradient descent