Optimization for Machine Learning
CS-439

Lecture 1: Introduction & Convexity

Martin Jaggi

EPFL – `github.com/epfml/OptML_course`

February 22, 2018

Outline

- Convexity, Gradient Methods, Constrained Optimization, Proximal algorithms, Subgradient Methods, **Stochastic Gradient Descent**, Coordinate Descent, Frank-Wolfe, Accelerated Methods, Primal-Dual context and certificates, Lagrange and Fenchel Duality, Second-Order Methods including Quasi-Newton, Derivative-Free Optimization.
- Advanced Contents:
  - Parallel and Distributed Optimization Algorithms, Synchronous and Asynchronous Communication.
  - Computational and Statistical Trade-Offs (Time vs Data vs Accuracy). Variance Reduced Methods, and Lower Bounds.
  - Non-Convex Optimization: Convergence to Critical Points, Saddle-Point methods, Alternating minimization for matrix and tensor factorizations

Course Organization

- Lectures
- Exercises
- Mini-Project

Grading: Written final exam, closed book

See details on course webpage on github

Optimization

- General optimization problem (**unconstrained minimization**)

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{with} \quad & \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

  - candidate solutions, variables, parameters $\mathbf{x} \in \mathbb{R}^d$
  - objective function $f : \mathbb{R}^d \to \mathbb{R}$
  - typically: technical assumption: f is continuous and differentiable

Why? And How?

Optimization is everywhere

machine learning, big data, statistics, data analysis of all kinds, finance, logistics, planning, control theory, mathematics, search engines, simulations, and many other applications ...

- **Mathematical Modeling**:
  - *defining & modeling the optimization problem*
- **Computational Optimization**:
  - *running an (appropriate) optimization algorithm*

Optimization for Machine Learning

- **Mathematical Modeling**:
  - defining & and measuring the machine learning model
- **Computational Optimization**:
  - learning the model parameters

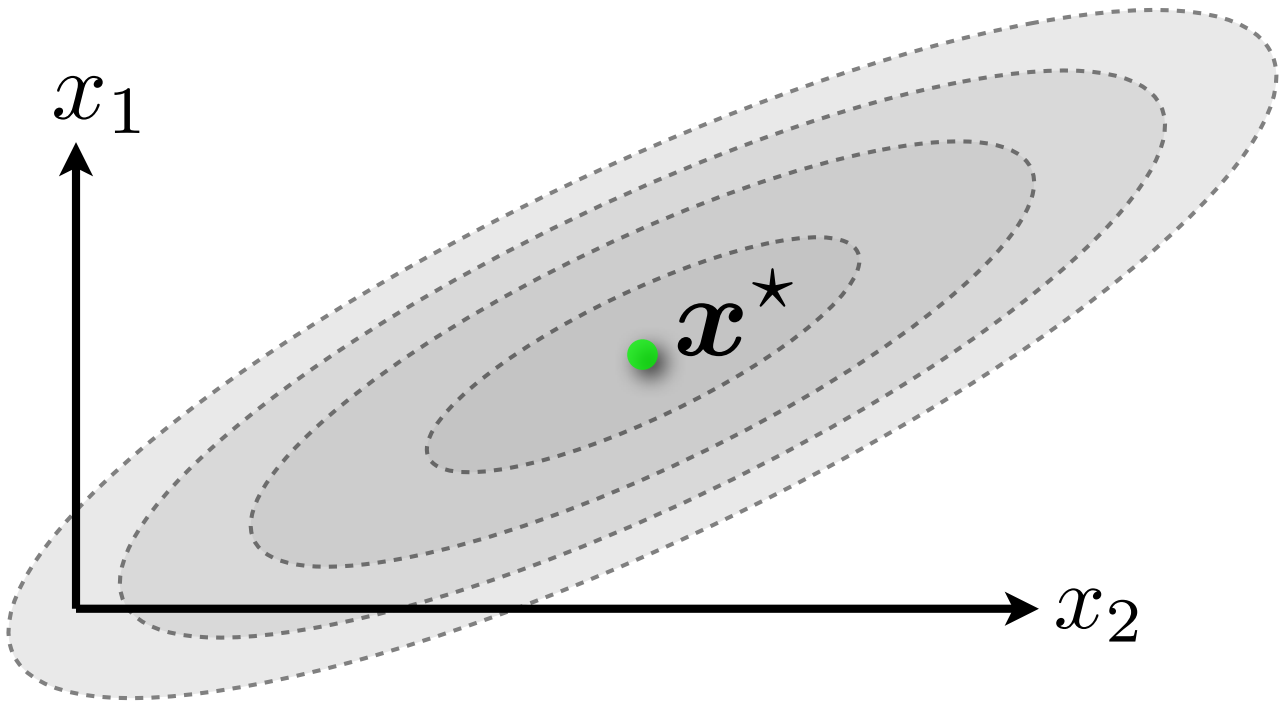
But what about deep learning?
Convex theory does not apply, so is useless?

Optimization Algorithms

- Optimization at large scale: **simplicity** rules!
- Main approaches:
  - **Gradient Descent**
  - **Stochastic Gradient Descent** (SGD)
  - **Coordinate Descent**
- History:
  - 1847: Cauchy proposes gradient descent
  - 1950s: Linear Programs, soon followed by non-linear, SGD
  - 1980s: General optimization, convergence theory
  - 2005-today: Large scale optimization, convergence of SGD

Example: Coordinate Descent

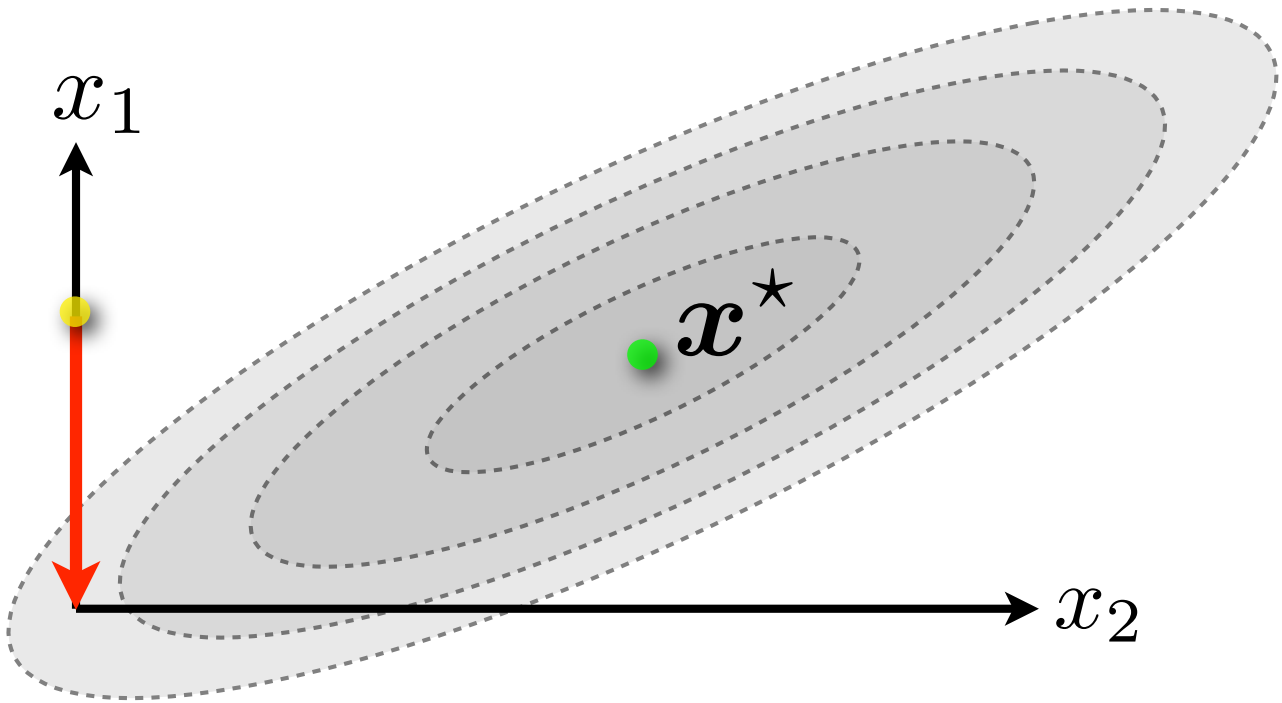
Goal: Find $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing $f(\mathbf{x})$. (Example: $d = 2$)



Idea: Update one coordinate at a time, while keeping others fixed.

Example: Coordinate Descent

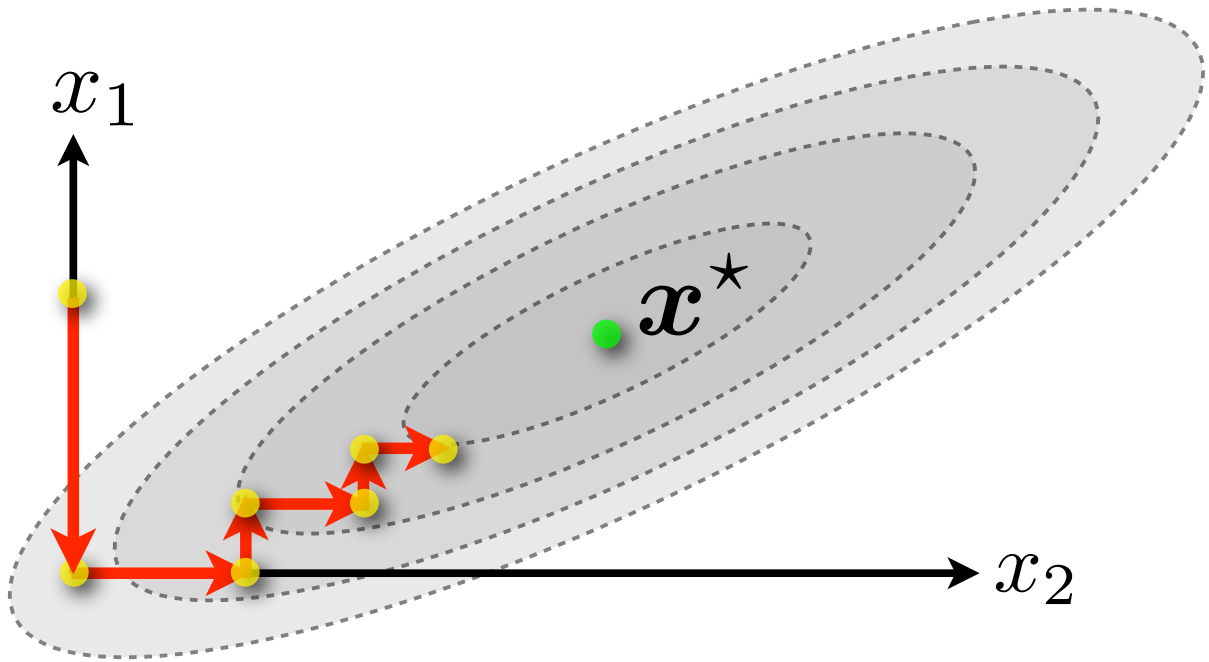
Goal: Find $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing $f(\mathbf{x})$.



Idea: Update one coordinate at a time, while keeping others fixed.

Example: Coordinate Descent

Goal: Find $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing $f(\mathbf{x})$.



Idea: Update one coordinate at a time, while keeping others fixed.

Chapter 1

Theory of Convex Functions

Convex Sets

A set C is **convex** if the line segment between any two points of C lies in C, i.e., if for any $\mathbf{x}, \mathbf{y} \in C$ and any λ with $0 \leq \lambda \leq 1$, we have

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C.$$

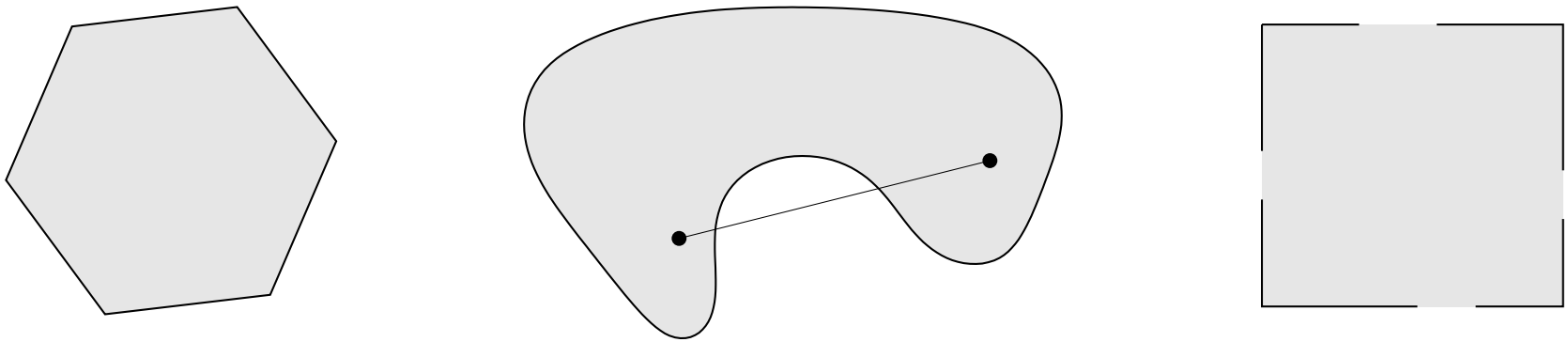


*Figure 2.2 from S. Boyd, L. Vandenberghe

Left Convex.

Middle Not convex, since line segment not in set.

Right Not convex, since some, but not all boundary points are contained in the set.

Properties of Convex Sets

- Intersections of convex sets are convex

  Observation 1.2. Let $C_i, i \in I$ be convex sets, where I is a (possibly infinite) index set. Then $C = \bigcap_{i \in I} C_i$ is a convex set.

- (later) Projections onto convex sets are *unique*, and *often* efficient to compute

$$P_C(\mathbf{x}') := \operatorname{argmin}_{\mathbf{y} \in C} \|\mathbf{y} - \mathbf{x}'\|$$

Convex Functions

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if (i) $\mathbf{dom}(f)$ is a convex set and (ii) for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, and λ with $0 \leq \lambda \leq 1$, we have

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}).$$

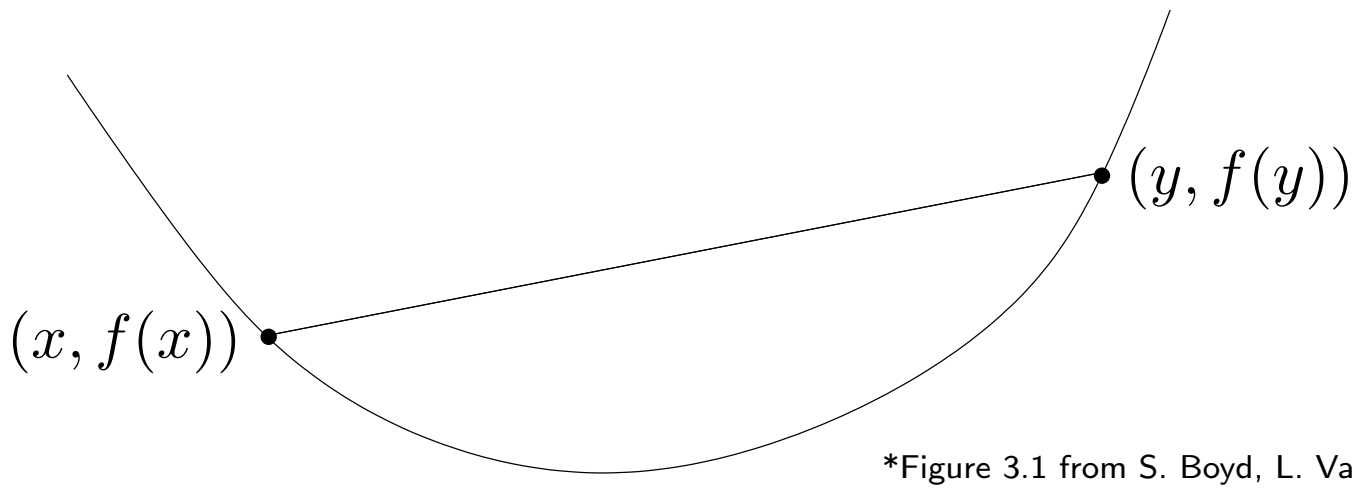


*Figure 3.1 from S. Boyd, L. Vandenberghe

Geometrically: The line segment between $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$ lies above the graph of f.

Motivation: Convex Optimization

Convex Optimization Problems are of the form

$$\min \; f(\mathbf{x}) \qquad \text{s.t.} \qquad \mathbf{x} \in C$$

where both

- f is a convex function
- C is a convex set (note: $\mathbb{R}^d$ is convex)

Properties of Convex Optimization Problems

- Every local minimum is a **global minimum**, see next...

Motivation: Solving Convex Optimization - Provably

For convex optimization problems, all algorithms

- Coordinate Descent
- Gradient Descent
- Stochastic Gradient Descent
- Projected [Stoch.] Gradient Descent

do **converge** to the global optimum! (assuming f differentiable)

Example Theorem: For convex problems, the **convergence rate** of "most" of the above algorithms is proportional to $\frac{1}{t}$, i.e.

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{c}{t}$$

(where $\mathbf{x}^\star$ is some optimal solution to the problem.)

Motivation: Convergence Theory

f	Algorithm	Rate	# Iter	Cost/iter
non-smooth	center of gravity	$\exp\left(-\frac{t}{n}\right)$	$n \log\left(\frac{1}{\varepsilon}\right)$	1 ∇, 1 n-dim $\int$
non-smooth	ellipsoid method	$\frac{R}{r}\exp\left(-\frac{t}{n^2}\right)$	$n^2 \log\left(\frac{R}{r\varepsilon}\right)$	1 ∇, mat-vec $\times$
non-smooth	Vaidya	$\frac{Rn}{r}\exp\left(-\frac{t}{n}\right)$	$n \log\left(\frac{Rn}{r\varepsilon}\right)$	1 ∇, mat-mat $\times$
quadratic	CG	exact $\exp\left(-\frac{t}{\kappa}\right)$	$\kappa \log\left(\frac{1}{\varepsilon}\right)$	1 ∇
non-smooth, Lipschitz	PGD	$RL/\sqrt{t}$	R^2L^2/ε^2	1 ∇, 1 proj.
smooth	PGD	$\beta R^2/t$	$\beta R^2/\varepsilon$	1 ∇, 1 proj.
smooth	AGD	$\beta R^2/t^2$	$R\sqrt{\beta/\varepsilon}$	1 ∇
smooth (any norm)	FW	$\beta R^2/t$	$\beta R^2/\varepsilon$	1 ∇, 1 LP
strong. conv., Lipschitz	PGD	$L^2/(\alpha t)$	$L^2/(\alpha\varepsilon)$	1 ∇ , 1 proj.
strong. conv., smooth	PGD	$R^2\exp\left(-\frac{t}{\kappa}\right)$	$\kappa \log\left(\frac{R^2}{\varepsilon}\right)$	1 ∇ , 1 proj.
strong. conv., smooth	AGD	$R^2\exp\left(-\frac{t}{\sqrt{\kappa}}\right)$	$\sqrt{\kappa}\log\left(\frac{R^2}{\varepsilon}\right)$	1 ∇
$f+g$, f smooth, g simple	FISTA	$\beta R^2/t^2$	$R\sqrt{\beta/\varepsilon}$	1 ∇ of f Prox of g
$\max_{y\in\mathcal{Y}} \varphi(x,y)$, φ smooth	SP-MP	$\beta R^2/t$	$\beta R^2/\varepsilon$	MD on $\mathcal{X}$ MD on $\mathcal{Y}$
linear, $\mathcal{X}$ with F ν-self-conc.	IPM	$\nu\exp\left(-\frac{t}{\sqrt{\nu}}\right)$	$\sqrt{\nu}\log\left(\frac{\nu}{\varepsilon}\right)$	Newton step on F
non-smooth	SGD	$BL/\sqrt{t}$	B^2L^2/ε^2	1 stoch. ∇, 1 proj.
non-smooth, strong. conv.	SGD	$B^2/(\alpha t)$	$B^2/(\alpha\varepsilon)$	1 stoch. ∇, 1 proj.
$f=\frac{1}{m}\sum f_i$ f_i smooth strong. conv.	SVRG	–	$(m+\kappa)\log\left(\frac{1}{\varepsilon}\right)$	1 stoch. ∇

(Bubeck [Bub15])

Convex Functions & Sets

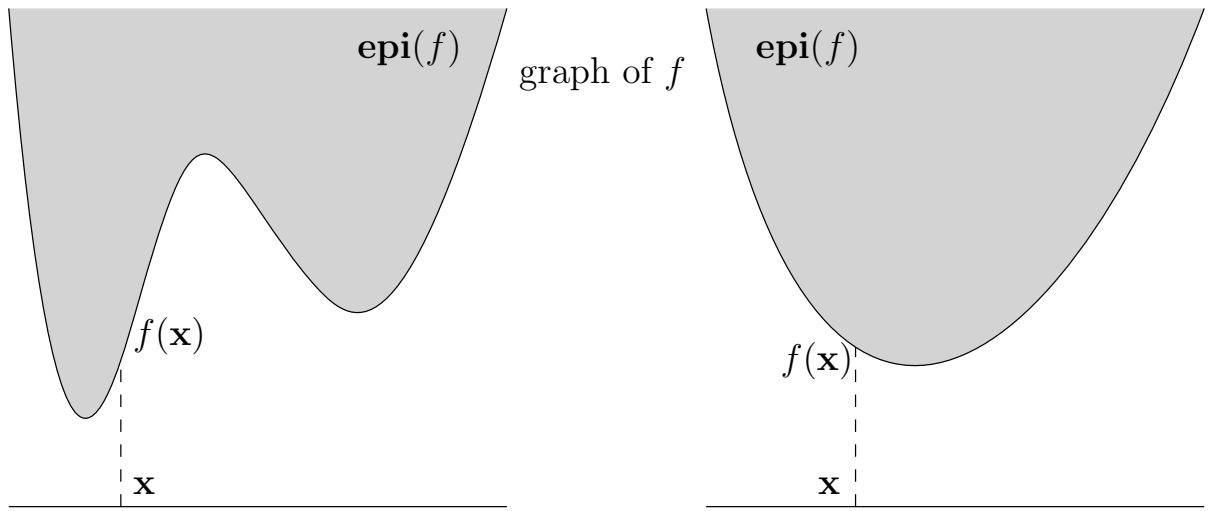
Epigraph: The *graph* of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\{(\mathbf{x}, f(\mathbf{x})) \,|\, \mathbf{x} \in \mathbf{dom}(f)\},$$

The **epigraph** of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathbf{epi}(f) := \{(\mathbf{x}, \alpha) \in \mathbb{R}^{d+1} \,|\, \mathbf{x} \in \mathbf{dom}(f), \alpha \geq f(\mathbf{x})\},$$

Observation 1.4. A function is convex *iff* its epigraph is a convex set.



Convex Functions & Sets

Proof: recall $\mathbf{epi}(f) := \{(\mathbf{x}, \alpha) \in \mathbb{R}^{d+1} \,|\, \mathbf{x} \in \mathbf{dom}(f), \alpha \geq f(\mathbf{x})\}$

Convex Functions

Examples of convex functions

- Linear functions: $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$
- Affine functions: $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$
- Exponential: $f(\mathbf{x}) = e^{\alpha \mathbf{x}}$
- Norms. Every norm on $\mathbb{R}^d$ is convex.

Convexity of a norm $f(\mathbf{x})$

By the triangle inequality $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ and homogeneity of a norm $f(a\mathbf{x}) = |a| f(\mathbf{x})$, a scalar:

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq f(\lambda \mathbf{x}) + f((1-\lambda)\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

We used the triangle inequality for the inequality and homogeneity for the equality.

Jensen's inequality

Lemma (Jensen's inequality)

Let f be convex, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbf{dom}(f)$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^d \lambda_i = 1$. Then

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

For $m = 2$, this is convexity. The proof of the general case is **Exercise 1**.

First-order characterization of convexity

Lemma ([BV04, 3.1.3])

Suppose that $\mathbf{dom}(f)$ *is open and that* f *is differentiable; in particular, the* **gradient** *(vector of partial derivatives)*

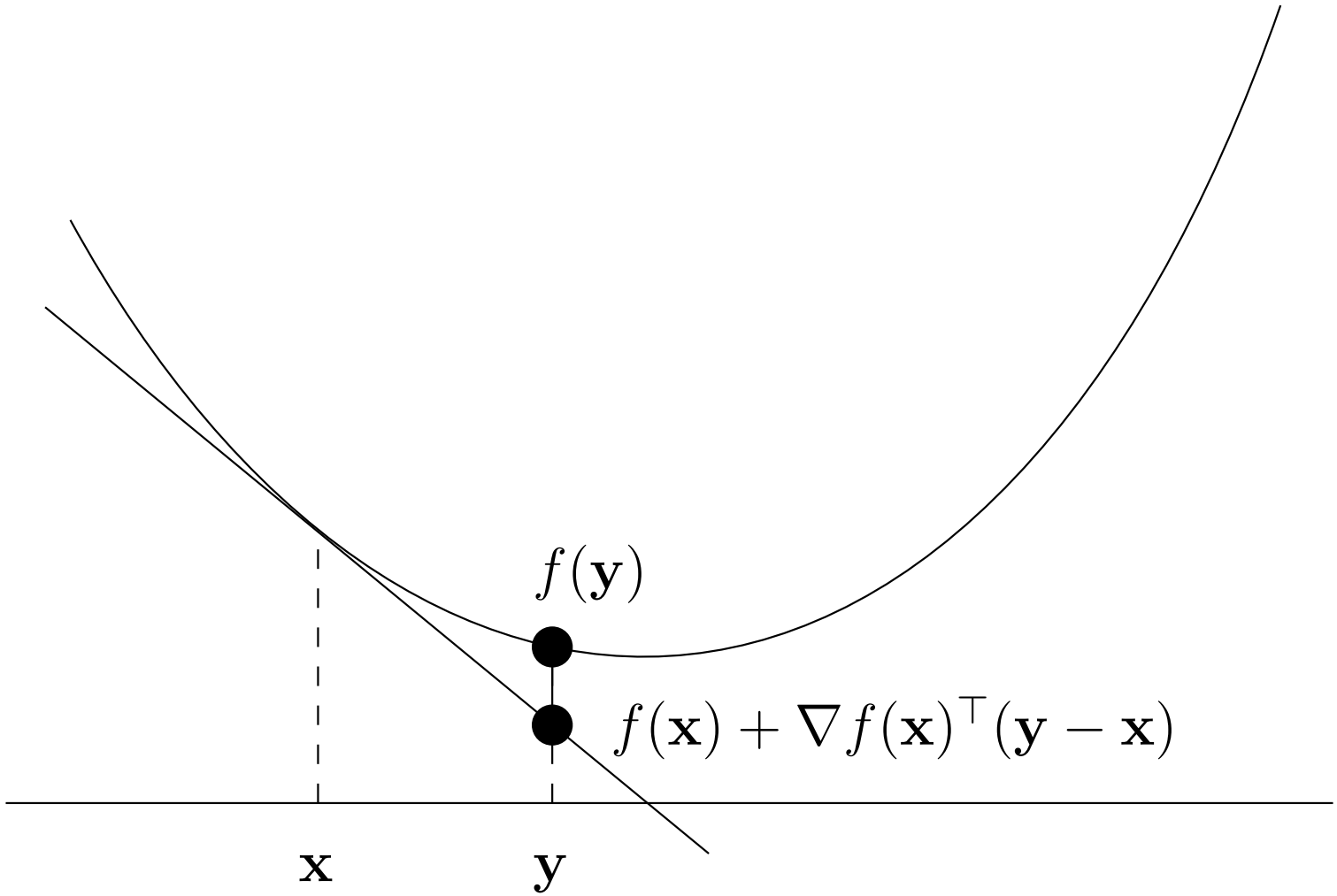
$$\nabla f(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d} \right)$$

exists at every point $\mathbf{x} \in \mathbf{dom}(f)$*. Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex and*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \tag{1}$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$*.*

First-order characterization of convexity



Second-order characterization of convexity

Lemma ([BV04, 3.1.4])

Suppose that $\mathbf{dom}(f)$ *is open and that* f *is twice differentiable; in particular, the* **Hessian** *(matrix of second partial derivatives)*

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_d} \end{pmatrix}$$

exists at every point $\mathbf{x} \in \mathbf{dom}(f)$ *and is symmetric. Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex, and for all* $\mathbf{x} \in \mathbf{dom}(f)$*, we have*

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{(i.e. } \nabla^2 f(\mathbf{x}) \textit{ is positive semidefinite).}$$

(A symmetric matrix M *is positive semidefinite if* $\mathbf{x}^\top M \mathbf{x} \geq 0$ *for all* $\mathbf{x}$*, and positive definite if* $\mathbf{x}^\top M \mathbf{x} > 0$ *for all* $\mathbf{x}$*.)*

Operations that preserve convexity

Lemma (Exercise 4)

(i) *Let* $f_1, f_2, \ldots, f_m$ *be convex functions,* $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. *Then* $f := \sum_{i=1}^m \lambda_i f_i$ *is convex on* $\mathbf{dom}(f) := \bigcap_{i=1}^m \mathbf{dom}(f_i)$.

(ii) *Let* f *be a convex function with* $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, $g : \mathbb{R}^m \to \mathbb{R}^d$ *an affine function, meaning that* $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, *for some matrix* $A \in \mathbb{R}^{d \times m}$ *and some vector* $\mathbf{b} \in \mathbb{R}^d$. *Then the function* $f \circ g$ *(that maps* $\mathbf{x}$ *to* $f(A\mathbf{x} + \mathbf{b})$*) is convex on* $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$.

Solving Convex Optimization Problems - Provably

Definition

A **local minimum** of $f : \mathbf{dom}(f) \to \mathbb{R}$ is a point $\mathbf{x}$ such that there exists $\varepsilon > 0$ with

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f) \text{ satisfying } \|\mathbf{y} - \mathbf{x}\| < \varepsilon.$$

Lemma

Let $\mathbf{x}^\star$ be a **local minimum** *of a convex function $f : \mathbf{dom}(f) \to \mathbb{R}$. Then $\mathbf{x}^\star$* **is a global minimum**, *meaning that*

$$f(\mathbf{x}^\star) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f).$$

Proof.

Solving Convex Optimization Problems - Provably

Lemma

Suppose that f is convex and differentiable over an open domain $\mathbf{dom}(f)$. Let $\mathbf{x} \in \mathbf{dom}(f)$. If $\nabla f(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is a **global minimum**.

Proof.

Suppose that $\nabla f(\mathbf{x}) = \mathbf{0}$. According to our Lemma on the first-order characterization of convexity, we have

Strictly convex functions

Definition ([BV04, 3.1.1])

A function $f : \mathbf{dom}(f) \to \mathbb{R}$ is **strictly convex** if (i) $\mathbf{dom}(f)$ is convex and (ii) for all $\mathbf{x} \neq \mathbf{y} \in \mathbf{dom}(f)$ and all $\lambda \in (0, 1)$, we have

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}). \tag{2}$$

Lemma

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be strictly convex. Then f has at most one global minimum.

Bibliography


Sébastien Bubeck.
Convex Optimization: Algorithms and Complexity.
Foundations and Trends in Machine Learning, 8(3-4):231–357, 2015.


Stephen Boyd and Lieven Vandenberghe.
Convex Optimization.
Cambridge University Press, New York, NY, USA, 2004.
`https://web.stanford.edu/~boyd/cvxbook/`.