Session 9: Convex functions

Optimization and Computational Linear Algebra for Data Science

Marylou Gabrié (based on material by Léo Miolane)

Contents

Optimization

In machine learning, we often have to minimize functions

$$f(\theta) = \mathrm{Loss}(\mathrm{data}, \mathrm{model}_\theta) \quad \text{with respect to} \quad \theta \in \mathbb{R}^n.$$

- For $n = 1, 2$, one could plot f to find the minimizer.
- This is intractable for larger dimension.

We will

- focus on convex cost functions f.
- study gradient descent algorithms to minimize f.

Convex vs non-convex

Convex	Non-convex

1. Functions of n variables

Functions of one variable

$$\begin{aligned} f : \quad \mathbb{R} &\to \mathbb{R} \\ x &\mapsto f(x) \end{aligned}$$

Functions of n variables

$$\begin{aligned} f: \quad \mathbb{R}^n &\to \mathbb{R} \\ x &\mapsto f(x) = f(x_1, \ldots, x_n) \end{aligned}$$

1.1.1 Derivative / Gradient

$$\begin{array}{rccc} f: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & f(x) \end{array}$$

Derivative at $x \in \mathbb{R}$:

$$f'(x) \in \mathbb{R}$$

$$\begin{array}{rccl} f: & \mathbb{R}^n & \to & \mathbb{R} \\ & x & \mapsto & f(x) = f(x_1, \ldots, x_n) \end{array}$$

Gradient at $x \in \mathbb{R}^n$:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^n$$

Gradient and contour lines

1.1.2 Hessian matrix

What is the equivalent of the second derivative for function of n variables ?

$$\begin{array}{rccl} f: & \mathbb{R}^n & \to & \mathbb{R} \\ & x & \mapsto & f(x) = f(x_1, \dots, x_n) \end{array}$$

Hessian at $x \in \mathbb{R}^n$:

$$H_f(x) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(x) & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_2}(x) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_n}(x) \\ \dfrac{\partial^2 f}{\partial x_2 \, \partial x_1}(x) & \dfrac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \, \partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \, \partial x_1}(x) & \dfrac{\partial^2 f}{\partial x_n \, \partial x_2}(x) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Example

Schwarz's Theorem

Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is «twice differentiable», then for all $x \in \mathbb{R}^n$ and all $i, j \in \{1, \ldots, n\}$ we have:

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(x) = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)(x).$$

1.2 Taylor's formulas

Functions of one variable, $\begin{array}{rccc} f: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & f(x) \end{array}$

Order 1 Taylor's formulas

Let $x \in \mathbb{R}^n$. Heuristically, for $h \in \mathbb{R}^n$ "small", we have

$$f(x+h) \;\simeq\; f(x) + \langle \nabla f(x), h \rangle.$$

Order 2 Taylor's formulas

Let $x \in \mathbb{R}^n$. Heuristically, for $h \in \mathbb{R}^n$ "small", we have

$$f(x+h) \;\simeq\; f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\mathsf{T} H_f(x) h.$$

2. Convexity

2.1 Convex sets

Definition

A set $S \subset \mathbb{R}^n$ is called a convex set if for all $x, y \in S$ and all $\alpha \in [0, 1]$,

$$\alpha x + (1 - \alpha)y \in S.$$

Properties/Exercises

1. Show that any subspace S of $\mathbb{R}^n$ is convex.
2. Let $\|\cdot\|$ be a (arbitrary) norm and $r \geq 0$. Show that the "ball" of radius r:
$$B(r) = \{x \in \mathbb{R}^n \mid \|x\| \leq r\}$$
is convex.

2.2 Convex / concave functions

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and all $\alpha \in [0, 1]$,

$$f(\alpha x + (1 - \alpha)y) \quad \leq \quad \alpha f(x) + (1 - \alpha)f(y). \tag{1}$$

2.2 Convex / concave functions

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and all $\alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \quad \leq \quad \alpha f(x) + (1-\alpha) f(y). \tag{1}$$

- We say that f is *strictly convex* is there is strict inequality in (1) whenever $x \neq y$ and $\alpha \in (0, 1)$.
- A function f is called concave if $-f$ is convex.

Properties/Exercises

1. Show that any linear map $f : \mathbb{R}^n \to \mathbb{R}$ is convex and concave.
2. Show that a norm $\|\cdot\|$ is convex.
3. Show that the sum of two convex functions is also a convex function.

3. Convex functions and derivatives

Convex functions vs their tangents

Proposition

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for all $x, y \in \mathbb{R}^n$

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

Proof. Exercise!

Minimizers of a convex function

Corollary

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function and $x \in \mathbb{R}^n$. Then

$$x \text{ is a minimizer of } f \iff \nabla f(x) = 0.$$

Hessian of convex function

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function. Then f is convex if and only if for all $x \in \mathbb{R}^n$, $H_f(x)$ is positive semi-definite.

Recall ways of proving positive definiteness.

Hessian of convex function

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function. Then f is convex if and only if for all $x \in \mathbb{R}^n$, $H_f(x)$ is positive semi-definite.

Remarks.

- Functions of 1 variables:
- Positive definite Hessian:
- Proof intuition:

4. Jensen’s inequality

Jensen's inequality

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then for all $x_1, \ldots, x_k \in \mathbb{R}^n$ and all $\alpha_1, \ldots, \alpha_k \geq 0$ such that $\sum_{i=1}^k \alpha_i = 1$ we have

$$f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leq \sum_{i=1}^{k} \alpha_i f(x_i).$$

More generally, if X is a random variable that takes value in $\mathbb{R}^n$ we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Example: entropy

- Consider a random variable X that takes values in $\{1, \cdots, k\}$.
- Denote by p_i, the probability $P(X = i)$ for $i \in \{1, \cdots, k\}$.
- The **entropy** of X is defined by

$$H(X) = \sum_{i=1}^{k} p_i \log\left(\frac{1}{p_i}\right)$$

.

- Observe that he logarithm is a concave function (exercise!).

Example: entropy

We just proved: $$0 \leq H(X) \leq \log(k)$$

.

- For the uniform distribution, $p_i = \frac{1}{k}$.

Questions?

Questions?