Session 9: Convex functions

Optimization and Computational Linear Algebra for Data Science

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

Gradient/Hessian

For $f : \mathbb{R}^n \to \mathbb{R}$:

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

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

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

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

Taylor's formulas

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

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

Taylor's formulas

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

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

Convex sets

Convex set

Definition

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

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

Exercise

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.

Convex functions

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 $\alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \quad \leq \quad \alpha f(x) + (1-\alpha) f(y). \qquad (1)$$

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.

Exercise

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.

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

Proof

Proof

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.

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.

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

Example: entropy

Questions?

Questions?