Session 9: Convex functions

Optimization and Computational Linear Algebra for Data Science

Marylou Gabrié (based on material by Léo Miolane)

Contents

- 1. Functions of multiple variables
 - 1.1 Gradients and Hessian
 - 1.2 Taylor's formula
- 2. Convexity definitions
 - 2.1 Convex sets
 - 1.2 Convex functions
- 3. Convexity and derivatives
- 4. Jensen's Inequality

Optimization

In machine learning, we often have to minimize functions

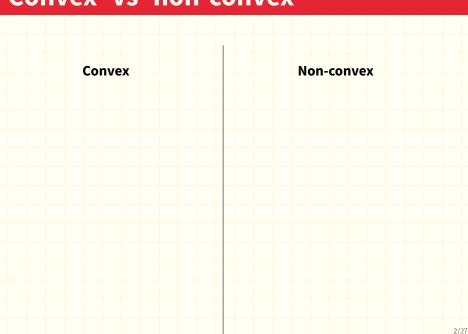
$$f(\theta) = \operatorname{Loss}(\operatorname{data}, \operatorname{model}_{\theta})$$
 with respect to $\theta \in \mathbb{R}^n$.

- 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.
- ightharpoonup study gradient descent algorithms to minimize f.

Convex vs non-convex



1. Functions of *n* variables

1. Functions of n variables 3/27

Functions of one variable

1. Functions of n variables

	£		TIDD	,		TDD										
	J	•	$\mathbb{I}\mathbb{Z}$	$\overset{\rightarrow}{\mapsto}$		$\mathbb{I}\mathbb{Z}$										
			œ		4	· (m	١									
			\boldsymbol{x}	\neg	J	(u))									

Functions of n variables

		TTD m			TTD.												
	f	\mathbb{R}^n															
		x	-	\rightarrow	f(:	x) =	= f	(x)	$_{1},\cdot$, :	(x_n)						

1. Functions of n variables

1.1.1 Derivative / Gradient

$$f: \mathbb{R} \to \mathbb{R}$$

$$x \mapsto f(x)$$

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

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

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

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

İ													

1. Functions of n variables 1.1 Gradient and Hessian 7/27

1.1.2 Hessian matrix

What is the equivalent of the second derivative for function of nvariables?

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

$$x \mapsto f(x) = f(x_1, \dots, x_n)$$

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

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

8/27

Example

9/27

1. Functions of n variables 1.1 Gradient and Hessian

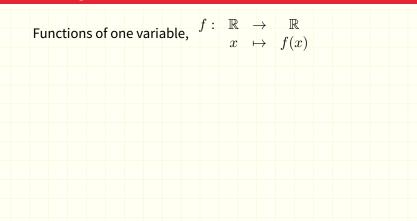
Schwarz's Theorem

Theorem

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



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. Convexity 14/27

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:

$$B(r) = \{ x \in \mathbb{R}^n \mid ||x|| \le r \}$$

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

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

2.2 Convex / concave functions

Definition

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

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

- 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.
- 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 $x,y\in\mathbb{R}^n$

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

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

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

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

4. Jensen's inequality 24/27

Jensen's inequality

Theorem

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

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

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

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

4. Jensen's inequality 25/27

Example: entropy

- Consider a random variable X that takes values in $\{1, \cdots, k\}$.
- Denote by p_i , the probability P(X=i) for $i \in \{1, \dots, k\}$.
- 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!).

4. Jensen's inequality 26/27

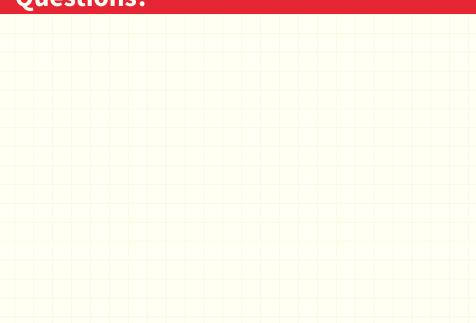
Example: entropy

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

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

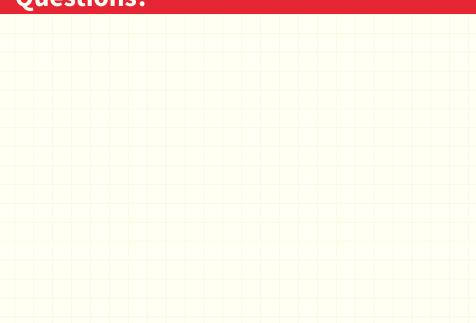
4. Jensen's inequality

Questions?



28/27

Questions?



28/27