

Deep learning and continuous optimization

Spring semester 2025/26

Kristóf Bérczi

Eötvös Loránd University
Institute of Mathematics
Department of Operations Research

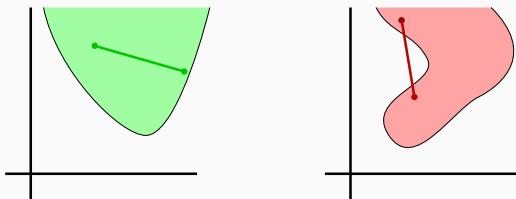


Lecture 2: Convexity

Convex sets

A set $K \subseteq \mathbb{R}^n$ is **convex** if for all $x, y \in K$ and $\theta \in [0, 1]$, we have

$$\theta x + (1 - \theta)y \in K.$$



Examples:

- **Polytopes:** $K = \{x \in \mathbb{R}^n : \langle a_i, x \rangle \leq b_i \text{ for } i = 1, \dots, m\}$, where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for $i = 1, \dots, m$.
- **Ellipsoids:** $K = \{x \in \mathbb{R}^n : x^T A x \leq 1\}$ where $A \in \mathbb{R}^{n \times n}$ is a positive definite matrix.
- **Balls (in ℓ_p norms for $p \geq 1$):** $K = \{x \in \mathbb{R}^n : \sqrt[p]{\sum_{i=1}^n |x_i - a_i|^p} \leq 1\}$, where $a \in \mathbb{R}^n$ is a vector.

Convex functions

A function $f : K \rightarrow \mathbb{R}$ is **convex** if its domain is a convex set and for all $x, y \in K$ and $\theta \in [0, 1]$, we have

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y).$$

If the inequality always holds as strict inequality, the function is **strictly convex**.

The function f is **concave** or **strictly concave** if $-f$ is convex or strictly convex, respectively.

Remark: If $f : K \rightarrow \mathbb{R}^n$ is a convex function, then setting $f(x) = +\infty$ for $x \notin K$ results in a convex function when the arithmetic operations on $\mathbb{R} \cup \{+\infty\}$ are interpreted in the reasonable way.

Semidefinite matrices

A matrix $M \in \mathbb{R}^{n \times n}$ is **symmetric** if $M^T = M$.

The **identity matrix** of size $n \times n$ is denoted by I_n .

A symmetric matrix M is **positive semidefinite (PSD)** if $x^T M x \geq 0$ holds for all $x \in \mathbb{R}^n$, and this is denoted by $M \succeq 0$.

M is **positive definite (PD)** if $x^T M x > 0$ holds for all non-zero $x \in \mathbb{R}^n$, and this is denoted by $M \succ 0$.

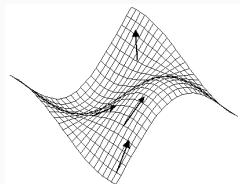
We define $M \succeq N \Leftrightarrow M - N \succeq 0$ and $M \succ N \Leftrightarrow M - N \succ 0$.

Calculus I

We are working with 'sufficiently smooth' functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

The derivative of $f(x_1, \dots, x_n)$ is called the **gradient**, and is defined as

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right]$$



The **directional derivative** of f in the direction d is $\langle \nabla f(x), d \rangle$.

The second derivatives of f can be summarized in the **Hessian** matrix

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Remark: The Hessian is symmetric if f is sufficiently differentiable.

Taylor expansion

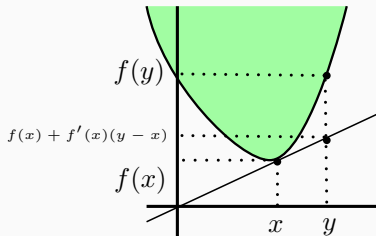
The Taylor series expansion of f around $x = a$ is

$$f(x) = \underbrace{f(a) + \langle \nabla f(a), x - a \rangle}_{\text{first order approximation}} + \underbrace{\frac{1}{2}(x - a)^T \nabla^2 f(a)(x - a) + \dots}_{\text{second order approximation}}$$

Consider a function in one dimension, i.e. $f : \mathbb{R} \rightarrow \mathbb{R}$.

When f is convex, the tangent is 'below' the graph, i.e.

$$f(y) \geq f(x) + f'(x)(y - x).$$



First order condition

First order condition

Let f be a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ over a convex set K . Then f is convex if and only if for all $x, y \in K$

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

Proof.

⇒ For any $\theta \in [0, 1]$, we have

$$(1 - \theta)f(x) + \theta f(y) \geq f(\theta y + (1 - \theta)x) = f(x + \theta(y - x)).$$

Subtracting $(1 - \theta)f(x)$ and dividing by θ yields

$$f(y) \geq f(x) + \frac{f(x + \theta(y - x)) - f(x)}{\theta}.$$

Taking limit $\theta \rightarrow 0$ gives $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$. □

First order condition

First order condition

Let f be a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ over a convex set K . Then f is convex if and only if for all $x, y \in K$

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

Proof.

⇐ Let $z = \theta x + (1 - \theta)y$. The first order approximation underestimates both $f(x)$ and $f(y)$, hence

$$f(x) \geq f(z) + \nabla f(z)^T(x - z),$$

$$f(y) \geq f(z) + \nabla f(z)^T(y - z).$$

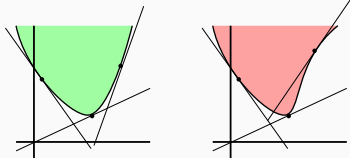
Therefore

$$\theta f(x) + (1 - \theta)f(y) \geq f(z) + \nabla f(z)^T(\theta x + (1 - \theta)y - z) = f(x + (1 - \theta)y).$$

□

Second order condition

In the one-dimensional case, $f''(x) \geq 0$ when f is convex, that is, the slope of the tangent is non-decreasing, as otherwise when the slope decreases the function becomes non-convex.



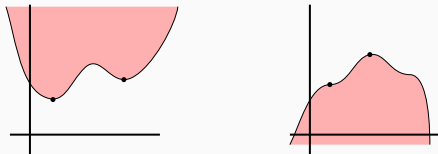
Second order condition

Let f be twice differentiable such that $\text{dom } f$ is open. Then f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \text{dom } f$.

Local vs. global optimum I

Convex optimization problem:

$$\inf_{x \in K} f(x) \xrightarrow{\text{usually}} \min_{x \in K} f(x)$$



Intuition: $\nabla f(x) = \underline{0}$ when x is optimal.

Problem: $\nabla f(x) = \underline{0}$ may correspond to a local optimum/maximum.

Global optimum

If the domain of a convex differentiable function f is \mathbb{R}^n , then x^* is an optimal solution to $\inf_{x \in \mathbb{R}^n} f(x)$ if and only if $\nabla f(x^*) = \underline{0}$.

Local vs. global optimum II

Proof of the 'if' direction.

Assume that $\nabla f(x_0) = \underline{0}$. Since f is convex, we know that for all $y \in \mathbb{R}^n$ we have

$$\begin{aligned} f(y) &\geq f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle \\ &= f(x_0) + \langle \underline{0}, y - x_0 \rangle \\ &= f(x_0). \end{aligned}$$

□

Remark: In the constrained setting, i.e. when $K \neq \mathbb{R}^n$, the following holds.

Global optimum

If f is a convex differentiable function, then x^* is an optimal solution to $\inf_{x \in K} f(x)$ if and only if $\langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in \mathbb{R}^n$.

Convex programs

A **convex program** can be written as follows.

Convex program

$$\inf f_0(x)$$

$$\text{s.t. } f_i(x) \leq 0 \text{ for } 1 \leq i \leq m$$

$$h_j(x) = 0 \text{ for } 1 \leq j \leq p$$

- f_i is convex for $i = 0, \dots, m$
- h_j is affine for $j = 1, \dots, p$

Remark: The domain of the problem is $D := \left(\bigcap_{i=0}^m \text{dom } f_i\right) \cap \left(\bigcap_{j=1}^p \text{dom } h_j\right)$, which is a convex set \Rightarrow Roughly speaking, this makes the problem tractable.

Question: Can we define a dual program? How to give a lower bound?

Dual programs I

Idea: “move the constraints into the objective function”

The **Lagrangian** associated with the problem is

$$L(x, \lambda, \mu) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \mu_j h_j(x),$$

where the λ_i s and μ_j s are called **Lagrangian multipliers**, and $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^p$ are called the **dual variables**.

The **Lagrangian dual function** is the min value of the Lagrangian over x ,

$$g(\lambda, \mu) := \inf_x L(x, \lambda, \mu).$$

Dual programs II

Let OPT_P denote the optimum value of the primal problem, and let \hat{x} be an arbitrary feasible solution. Furthermore, assume that $\lambda \geq 0$. Then

$$\begin{aligned}g(\lambda, \mu) &\leq f_0(\hat{x}) + \sum_{i=1}^m \lambda_i f_i(\hat{x}) + \sum_{j=1}^p \mu_j h_j(\hat{x}) \\ &\leq f_0(\hat{x}),\end{aligned}$$

hence $g(\lambda, \mu) \leq \inf_{x \text{ feasible}} f_0(x) = OPT_P$.

Conclusion:

- This gives a lower bound when $\lambda \geq 0$ and $g(\lambda, \mu) > -\infty \Rightarrow$ Such a pair λ, μ is called **dual feasible**.

Weak duality

The goal is to get the best lower bound on OPT_P using the Lagrangian dual.

The **dual program** is thus defined as

Dual program

$$\begin{aligned} \max \quad & g(\lambda, \mu) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

Let OPT_D denote the optimal value of the dual. Then **weak duality** holds by construction, that is, $OPT_D \leq OPT_P$.

Remarks:

- The dual program is always convex, regardless of the primal.
- That is, for any primal program (even though non-convex), we can always write a convex program that gives a lower bound on the primal objective value.

Strong duality

Question: Does $OPT_D = OPT_P$ always hold?

Answer: Unfortunately **NOT**. But!

Slater's condition requires that there is $x \in \text{relint}(D)$ such that $f_i(x) < 0$ if f_i is non-affine for $1 \leq i \leq m$, and $h_j(x) = 0$ for $1 \leq j \leq p$.

(That is, there exists an **interior** point in the domain, which is a feasible solution, and satisfies the non-affine inequality constraints **strictly**.)

Strong duality

If Slater's condition holds, then $OPT_D = OPT_P$.

Complementary slackness

Assume that $OPT_D = OPT_P$. Let x^* be a primal, λ^*, μ^* be dual optimal solutions. Then

$$\begin{aligned}f_0(x^*) &= g(\lambda^*, \mu^*) \\&= \inf_x L(x, \lambda^*, \mu^*) \\&\leq L(x^*, \lambda^*, \mu^*) \\&= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{j=1}^p \mu_j^* h_j(x^*) \\&\leq f_0(x^*),\end{aligned}$$

as $\lambda \geq 0$ (dual feasible) and $f_i(x^*) \leq 0$, $h_j(x^*) = 0$ (primal feasible).

Therefore

- x^* is a minimizer of $L(x, \lambda^*, \mu^*)$, and
- $\lambda_i^* f_i(x^*) = 0$ for $1 \leq i \leq m$, called the **complementary slackness condition**, meaning that the non-zero pattern of λ_i^* and $f_i(x^*)$ must be complementary.

Karush-Kuhn-Tucker (KKT) conditions I

Assume that $f_0, f_1, \dots, f_m, h_1, \dots, h_p$ are all differentiable.

Since x^* minimizes $L(x, \lambda^*, \mu^*)$ by the above, the gradient of L at x^* must be zero, that is,

$$\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(x^*) = \underline{0}.$$

To sum up, the following are some necessary conditions for any pair of primal and dual optimal solutions.

Primal feasibility: $f_i(x^*) \leq 0$ for $1 \leq i \leq m$, $h_j(x^*) = 0$ for $1 \leq j \leq p$.

Dual feasibility: $\lambda_i^* \geq 0$ for $1 \leq i \leq m$.

Complementary slackness: $\lambda_i^* f_i(x^*) = 0$ for $1 \leq i \leq m$.

Lagrangian optimality: $\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(x^*) = \underline{0}$.

This set of conditions is called the **KKT conditions**.

Karush-Kuhn-Tucker (KKT) conditions II

When the primal problem is convex, the KKT conditions are also **sufficient**!

⇒ Any x^*, λ^*, μ^* satisfying KKT must be primal and dual optimal solutions.

Reason: If the primal is convex, then $L(x, \lambda, \mu)$ is convex in x when λ, μ are fixed. Hence a local optimal solution is also a global optimal solution. More precisely:

$$\begin{aligned}g(\lambda^*, \mu^*) &= \inf_x L(x, \lambda^*, \mu^*) \\&= L(x^*, \lambda^*, \mu^*) \\&= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{j=1}^p \mu_j^* h_j(x^*) \\&= f_0(x^*).\end{aligned}$$

Summary: For a convex problem with differentiable functions, if Slater's condition is satisfied, then the KKT conditions are **necessary** and **sufficient** for optimality.