MATH3027: Optimization (UK 22/23)

Week 9: Convex Functions and Convex Optimization

Prof. Richard Wilkinson
School of Mathematical Sciences
University of Nottingham, United Kingdom
Please send any comments or mistakes to
r.d.wilkinson@nottingham.ac.uk

This week we study convex functions and their characterization. Having studied convex sets and functions, we will be in a position to study convex optimization problems, that is, the optimization of convex functions constrained to convex sets. Convex problems are far much more general than linear programmes (LPs), but like LPs, they can be solved quickly and reliably for very large problems. Instead, the difficulty is determining whether or not a problem is convex.

We introduce the concept of stationarity as a characterization of optimality in convex problems, and finally, we will study projection operators and propose an algorithm for convex optimization, known as the projected gradient method.

Convex Functions

We begin by giving a definition of a convex function.

Definition (Convex Function). *A function $f : C \to \mathbb{R}$ defined*[1] *on a convex set $C \subseteq \mathbb{R}^n$ is* ***convex*** *if*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \text{ for any } \mathbf{x}, \mathbf{y} \in C, \lambda \in [0,1].$$

[1] In cases where no domain is specified, we assume that f is defined over the entire space $\mathbb{R}^n$.

*f is **strictly convex** if*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \text{ for any } \mathbf{x} \neq \mathbf{y} \in C, \lambda \in (0,1).$$

*We say f is **concave** if $-f$ is convex*[2].

Examples of Convex Functions

- Affine Functions. $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$, where $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Then

$$\begin{aligned} f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) &= \mathbf{a}^\top (\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) + b \\ &= \lambda \left(\mathbf{a}^\top \mathbf{x}\right) + (1-\lambda) \left(\mathbf{a}^\top \mathbf{y}\right) + \lambda b + (1-\lambda) b \\ &= \lambda \left(\mathbf{a}^\top \mathbf{x} + b\right) + (1-\lambda) \left(\mathbf{a}^\top \mathbf{y} + b\right) \\ &= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) . \end{aligned}$$

  Note that affine functions are also concave.

- Norms. $g(\mathbf{x}) = \|\mathbf{x}\|$, take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Then

$$\begin{aligned} g(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) &= \|\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\| \\ &\leq \|\lambda \mathbf{x}\| + \|(1-\lambda)\mathbf{y}\| \text{ by the triangle inequality} \\ &= \lambda \|\mathbf{x}\| + (1-\lambda)\|\mathbf{y}\| \\ &= \lambda g(\mathbf{x}) + (1-\lambda) g(\mathbf{y}) . \end{aligned}$$

We now state[3] a fundamental result for convex functions, that generalizes the definition of convexity from two to many points.

Theorem (Jensen's Inequality). *Let $f : C \longrightarrow \mathbb{R}$ be a convex function where $C \subseteq \mathbb{R}^n$ is a convex set. Then, for any $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in C$ and $\boldsymbol{\lambda} \in \Delta_k$, the following inequality holds:*

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

First Order Characterization of Convex Functions

Convex functions are not necessarily differentiable,but when they are, we can characterize them as functions for which the tangent hyperplane[4] underestimates the function. See Figure 1.

[2] Swap the direction of the inequalities.

[3] You may have seen this result stated in probability courses as

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$$

where X is a random variable.

[4] The tangent hyperplane of f at x is, by the linear approximation theorem,

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Theorem (The Gradient Inequality)**.** *Let $f : C \longrightarrow \mathbb{R}$ be a continuously differentiable function defined on a convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if*

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) \textit{ for any } \mathbf{x}, \mathbf{y} \in C\,. \tag{1}$$

An analogous result holds for strictly convex functions (with a strict inequality).

Proof. Suppose first that f is convex. Let $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in (0,1]$. If $\mathbf{x} = \mathbf{y}$, then the inequality trivially holds. Assume that $\mathbf{x} \neq \mathbf{y}$. Then by the definition of convexity for f

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

Taking $\lambda \longrightarrow 0^+$, we obtain

$$f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x})$$

where $f'(\mathbf{x}; \mathbf{d})$ is the directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{d}$. Since f is continuously differentiable, $f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) = \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$, and the inequality follows.

To prove the converse, assume that that the gradient inequality holds. Let $\mathbf{z}, \mathbf{w} \in C$, and let $\lambda \in (0,1)$. We will show that $f(\lambda\mathbf{z} + (1-\lambda)\mathbf{w}) \leq \lambda f(\mathbf{z}) + (1-\lambda)f(\mathbf{w})$, Let $\mathbf{u} = \lambda\mathbf{z} + (1-\lambda)\mathbf{w} \in C$. Then

$$\mathbf{z} - \mathbf{u} = \frac{\mathbf{u} - (1-\lambda)\mathbf{w}}{\lambda} - \mathbf{u} = -\frac{1-\lambda}{\lambda}(\mathbf{w} - \mathbf{u})\,.$$

By invoking the gradient inequality on the pairs $\mathbf{u}, \mathbf{z}$ and $\mathbf{u}, \mathbf{w}$, and substituting for $\mathbf{w} - \mathbf{u}$ using the above, we have

$$\begin{aligned} f(\mathbf{u}) + \nabla f(\mathbf{u})^\top(\mathbf{z} - \mathbf{u}) &\leq f(\mathbf{z})\,, \\ f(\mathbf{u}) - \frac{\lambda}{1-\lambda}\nabla f(\mathbf{u})^\top(\mathbf{z} - \mathbf{u}) &\leq f(\mathbf{w})\,. \end{aligned}$$

Thus,

$$f(\mathbf{u}) \leq \lambda f(\mathbf{z}) + (1-\lambda)f(\mathbf{w})$$

□

Convexity + Stationarity ⇒ Global Optimality!

A direct result of the gradient inequality is that the first order optimality condition $\nabla f(\mathbf{x}^*) = 0$ is sufficient for global optimality.

Theorem (Stationarity Implies Global Optimality)**.** *Let f be a continuously differentiable function which is convex over a convex set $C \subseteq \mathbb{R}^n$. Suppose that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ for some $\mathbf{x}^* \in C$. Then $\mathbf{x}^*$ is the global minimizer of f over C.*

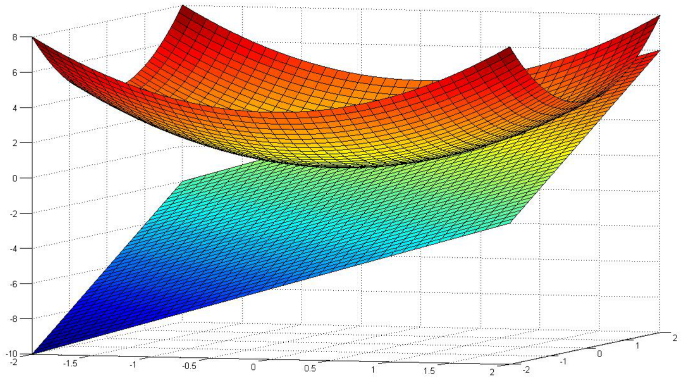

Figure 1: For a convex function f, the tangent plane at every point is always below f.

Proof. This is a direct consequence of the gradient inequality. □

For 1d functions, convex functions are characterized by having a non-decreasing gradient. The analogue of this for functions of multiple variables is the following characterization:

Theorem (Monotonicity of the Gradient)**.** *Suppose that f is a continuously differentiable function over a convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if*

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq 0 \textit{ for any } \mathbf{x}, \mathbf{y} \in C\,.$$

We can now extend our link between convexity and optimality conditions to second-order characterizations.

Theorem (Second-Order Characterization of Convexity)**.** *Let f be a twice continuously differentiable function over an open convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ for any $\mathbf{x} \in C$.*

In addition, if $\nabla^2 f(\mathbf{x}) \succ 0 \; \forall \; \mathbf{x}$ then f is strictly convex. The converse is false[5].

We can now revisit optimality conditions for quadratic functions. Let $f : \mathbb{R}^n \to \mathbb{R}$ be the quadratic function given by $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$ where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then f is convex if and only if $\mathbf{A} \succeq \mathbf{0}$ (and strictly convex iff $\mathbf{A} \succ \mathbf{0}$).

Example. Convexity of the log-sum-exp function:

$$f(\mathbf{x}) = \log\left(e^{x_1} + e^{x_2} + \ldots + e^{x_n}\right), \quad \mathbf{x} \in \mathbb{R}^n\,.$$

The gradient is given by:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, \quad i = 1, 2, \ldots, n\,.$$

[5] Consider $f(x) = x^4$. This is a strictly convex function, but $f''(0) = 0$.

Therefore, the Hessian is

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \begin{cases} -\frac{e^{x_i}e^{x_j}}{\left(\sum_{j=1}^n e^{x_j}\right)^2}, & i \neq j \\ -\frac{e^{x_i}e^{x_j}}{\left(\sum_{j=1}^n e^{x_j}\right)^2} + \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, & i = j \end{cases} .$$

We can thus write the Hessian matrix as

$$\nabla^2 f(\mathbf{x}) = \operatorname{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^\top, \qquad \text{with} \qquad \mathbf{w} = \left(\frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}\right)_{i=1}^n \in \Delta_n .$$

For any $\mathbf{v} \in \mathbb{R}^n$:

$$\mathbf{v}^\top \nabla^2 f(\mathbf{x}) \mathbf{v} = \sum_{i=1}^n w_i v_i^2 - \left(\mathbf{v}^\top \mathbf{w}\right)^2 \geq 0 ,$$

since defining $s_i = \sqrt{w_i} v_i$, $t_i = \sqrt{w_i}$, we have

$$\left(\mathbf{v}^\top \mathbf{w}\right)^2 = \left(\mathbf{s}^\top \mathbf{t}\right)^2 \leq \|\mathbf{s}\|^2 \|\mathbf{t}\|^2 = \left(\sum_{i=1}^n w_i v_i^2\right)\left(\sum_{i=1}^n w_i\right) = \sum_{i=1}^n w_i v_i^2 .$$

Thus, $\nabla^2 f(\mathbf{x}) \geq \mathbf{0}$ and hence f is convex over $\mathbb{R}^n$.

Example E.1 Show the convexity of the quad-over-lin function

$$f(x_1, x_2) = \frac{x_1^2}{x_2}$$

defined over $\mathbb{R} \times \mathbb{R}_{++} = \{(x_1, x_2) : x_2 > 0\}$.

Operations Preserving Convexity

- Let f be a convex function defined over a convex set $C \subseteq \mathbb{R}^n$ and let $\alpha \geq 0$. Then αf is a convex function over C.
- Let $f_1, f_2, \ldots, f_p$ be convex functions over a convex set $C \subseteq \mathbb{R}^n$. Then the sum function $f_1 + f_2 + \ldots + f_p$ is convex over C.
- Let f be a convex function defined on a convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{b} \in \mathbb{R}^n$. Then the function g defined by

  $$g(\mathbf{y}) = f(\mathbf{A}\mathbf{y} + \mathbf{b})$$

  is convex over the convex set $D = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{A}\mathbf{y} + \mathbf{b} \in C\}$.

- Let $f : C \longrightarrow \mathbb{R}$ be a convex function defined over the convex set $C \subseteq \mathbb{R}^n$. Let $g : I \longrightarrow \mathbb{R}$ be a one-dimensional nondecreasing convex function over the interval $I \subseteq \mathbb{R}$. Assume that the image of C under f is contained in $I : f(C) \subseteq I$. Then the composition of g with f defined by

$$h(\mathbf{x}) \equiv g(f(\mathbf{x}))$$

is convex over C. E.2 Find a counter example when g is not nondecreasing?

E.3 Prove these statements.

Examples E.4 Using the properties above, prove the following functions are convex.

- The generalized quad-over-lin function

$$g(\mathbf{x}) = \frac{\|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2}{\mathbf{c}^\top \mathbf{x} + d} \qquad \left(\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{c} \in \mathbb{R}^n, d \in \mathbb{R}\right)$$

is convex over $D = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{c}^\top \mathbf{x} + d > 0\}$.

- $f(x_1, x_2) = -\log(x_1 x_2)$, over $\mathbb{R}^2_{++}$.
- $f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2 + 2x_1 - 3x_2 + e^{x_1}$.
- $h(\mathbf{x}) = e^{\|\mathbf{x}\|^2}$.

In the computer lab, we will look at software for solving convex optimization problems. To do this, we have to write functions in such a way that the software is able to validate that the function is indeed convex.

Further Results for Convex Functions

Theorem (Point-Wise Maximum of Convex Functions)**.** *Let $f_1, f_2, \dots, f_p : C \rightarrow \mathbb{R}$ be p convex functions over the convex set $C \subseteq \mathbb{R}^n$. Then the maximum function*

$$f(\mathbf{x}) \equiv \max_{i=1,2,\dots,p} \{f_i(\mathbf{x})\}$$

is convex over C.

Examples E.5 Proving the following functions are all convex

- $f(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\}$.
- For a given vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$, let $x_{[i]}$ denote the i -th largest value in $\mathbf{x}$. For any $k \in \{1, 2, \ldots, n\}$ let

$$h_k(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[k]}.$$

Theorem (Preservation of Convexity Under Partial Minimization). *Let $f : C \times D \to \mathbb{R}$ be a convex function defined over the set $C \times D$ where $C \subseteq \mathbb{R}^m$ and $D \subseteq \mathbb{R}^n$ are convex sets. Let*

$$g(\mathbf{x}) = \min_{\mathbf{y} \in D} f(\mathbf{x}, \mathbf{y}), \quad \mathbf{x} \in C$$

where we assume that the minimum is finite. Then g is convex over C.

Example. The distance function from a convex set $d_C(\mathbf{x}) \equiv \inf_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|$ is convex.

Four Important Theorems for Convex Functions

We now state four important results for convex functions. The first relates to the continuity of convex functions. In particular, convex functions are continuous at interior points of their domain:

Theorem (Continuity of Convex Functions). *Let $f : C \longrightarrow \mathbb{R}$ be a convex function defined over a convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x}_0 \in \operatorname{int}(C)$. Then there exist $\varepsilon > 0$ and $L > 0$ such that $B[\mathbf{x}_0, \varepsilon] \subseteq C$ and*

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| \leq L\|\mathbf{x} - \mathbf{x}_0\| \text{ for any } \mathbf{x} \in B[\mathbf{x}_0, \varepsilon]$$

In addition, all the directional derivatives of a convex function exist at the interior of the domain.

Theorem (Existence of Directional Derivatives of Convex Functions). *Let $f : C \to \mathbb{R}$ be a convex function over the convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x} \in \operatorname{int}(C)$. Then for any $\mathbf{d} \neq 0$, the directional derivative $f'(\mathbf{x}; \mathbf{d})$ exists.*

The last two theorems relate to the problem of **maximizing** a non-constant convex function over a convex set. As we'll soon see, *minimizing* a convex function over a convex set is in some sense relatively easy, whereas in contrast, maximizing a convex function is hard as the stationarity condition does not hold.

Theorem (No Maximum Inside the Convex Set). *Let $f : C \longrightarrow \mathbb{R}$ be convex and nonconstant over the nonempty convex set $C \subseteq \mathbb{R}^n$. Then f does not attain a maximum at a point in* $\operatorname{int}(C)$.

Finally, we state that the maximum of convex function over compact convex sets can be found at the extreme points of the set.

Theorem (Maximum of a Convex Function Over a Compact Convex Set). *Let $f : C \longrightarrow \mathbb{R}$ be convex over the nonempty convex and compact set $C \subseteq \mathbb{R}^n$. Then there exists at least one maximizer of f over C that is an extreme point of C.*

Proof. Let $\mathbf{x}^*$ be a maximizer of f over C. If $\mathbf{x}^*$ is an extreme point of C, then the result is established. Otherwise, by Krein-Milman, $C = \operatorname{conv}(\operatorname{ext}(C))$ implies the existence of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in \operatorname{ext}(C)$ such that

$$\mathbf{x}^* = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i$$

for $\boldsymbol{\lambda} \in \Delta_k$. By the convexity of f,

$$f(\mathbf{x}^*) \leq \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i)$$

or equivalently

$$\sum_{i=1}^{k} \lambda_i \left(f(\mathbf{x}_i) - f(\mathbf{x}^*)\right) \geq 0$$

as $\lambda_i \geq 0$. Since $\mathbf{x}^*$ is a maximizer of f over C, we have $f(\mathbf{x}_i) \leq f(\mathbf{x}^*)$ for all $i = 1, \ldots, k$. This implies that $f(\mathbf{x}_i) = f(\mathbf{x}^*)$. Consequently, the extreme points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are all maximizers of f over C. □

E.6

Find the maximum of the following convex maximization problems:

1. $$\max\{\mathbf{x}^\top \mathbf{Q}\mathbf{x} : ||\mathbf{x}||_\infty \leq 1\}$$

   where $\mathbf{Q} \succeq 0$.

2. $$\max\{\ ||\mathbf{A}\mathbf{x}||_1 : ||\mathbf{x}||_1 \leq 1\}$$

   where $\mathbf{A} \in \mathbb{R}^{m \times n}$.

Convex Optimization Problems

A **convex optimization** problem consists of minimizing a convex function $f(\mathbf{x})$ over a convex set C:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in C \end{aligned} \tag{CVX}$$

A more explicit way of writing this is in a functional form:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0, \quad && i = 1, 2, \dots, m \\ & h_j(\mathbf{x}) = 0, \quad && j = 1, 2, \dots, p\,, \end{aligned}$$

where $f, g_1, \dots, g_m : \mathbb{R}^n \longrightarrow \mathbb{R}$ are convex functions, and $h_1, h_2, \dots, h_p : \mathbb{R}^m \longrightarrow \mathbb{R}$ are affine functions.

E.7

Show that this functional form of the problem does fit into the general formulation (CVX), i.e., show that $C = \{\mathbf{x} : g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0\}$ is a convex set.

A very important feature of convex optimization problems is that local minima are global minima!

Theorem (Local minima are global minima in convex problems.). *Let $f : C \longrightarrow \mathbb{R}$ be a convex function defined on the convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x}^* \in C$ be a local minimum of f over C. Then x^* is a global minimum of f over C.*

Proof. Assume $\mathbf{x}^*$ is a local minimum of f over C. This implies that there exists $r > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for any $\mathbf{x} \in C \cap B[\mathbf{x}^*, r]$. Let $\mathbf{x}^* \neq \mathbf{y} \in C$. We will show that $f(\mathbf{y}) \geq f(\mathbf{x}^*)$. Let $\lambda \in (0,1)$ be such that $\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*) \in B[\mathbf{x}^*, r]$. Since $\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*) \in B[\mathbf{x}^*, r]$, it follows that $f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*))$ and hence

$$\begin{aligned} f(\mathbf{x}^*) &\leq f(\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*)) \\ &= f((1-\lambda)\mathbf{x}^* + \lambda\mathbf{y}) \\ &\leq (1-\lambda) f(\mathbf{x}^*) + \lambda f(\mathbf{y})\,. \end{aligned}$$

by Jensen's inequality. The desired inequality $f(\mathbf{x}^*) \leq f(\mathbf{y})$ follows. □

A small variation of the proof of the last theorem yields the following.

Theorem. *Let $f : C \longrightarrow \mathbb{R}$ be a strictly convex function defined on the convex set C. Let $x^* \in C$ be a local minimum of f over C. Then x^* is a strict global minimum of f over C.*

Another important and easily deduced property of convex problems is that the set of optimal solutions is also convex.

Theorem. *Let $f : C \longrightarrow \mathbb{R}$ be a convex function defined over the convex set $C \subseteq \mathbb{R}^n$. Then the set of optimal solutions of the problem*

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}$$

is convex. If, in addition, f is strictly convex over C, then there exists at most one optimal solution of the problem.

E.8 Prove this result using Jensen's inequality.

Finally, note that maximizing a concave function over a convex set is also a convex optimization problem.

Examples:

- A Convex Problem: E.9

$$\begin{array}{ll} \min & -2x_1 + x_2 \\ \text{s.t.} & x_1^2 + x_2^2 \leq 3 \end{array}$$

- A Nonconvex Problem: E.10

$$\begin{array}{ll} \min & x_1^2 - x_2 \\ \text{s.t.} & x_1^2 + x_2^2 = 3 \end{array}$$

- Linear Programming

$$\begin{array}{lll} & \min & \mathbf{c}^\top \mathbf{x} \\ (\mathbf{LP}) : & \text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & & \mathbf{B}\mathbf{x} = \mathbf{g} \end{array}$$

  (LP) is a convex optimization problem (constraints and objective function are linear / affine and hence convex). It is also equivalent to a problem of maximizing a convex (linear) function subject to a convex constraints set. Hence, if the feasible set C is compact and nonempty, then by the theorems from last week, **there exists at least one optimal solution which is an extreme point, or equivalently, a basic feasible solution**.

- Convex Quadratic Problems consist of minimizing a convex quadratic function subject to affine constraints. The general form is

$$\begin{array}{ll} \min & \mathbf{x}^\top \mathbf{Q}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{c} \end{array}$$

  $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is positive semidefinite, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{c} \in \mathbb{R}^m$. Convex QP problems frequently occur in statistics and machine learning.

Optimization over a Convex Set and Stationarity

We will consider the constrained optimization problem given by

$$\min_{\mathbf{x}} \quad \{f(\mathbf{x}) : \mathbf{x} \in C\}\,, \tag{P}$$

where C is a closed convex subset of $\mathbb{R}^n$, and f is continuously differentiable over C, not necessarily convex.

When we looked at unconstrained optimization, we showed that a necessary condition for $\mathbf{x}^*$ to be a local optima is that $\nabla f(\mathbf{x}^*) = 0$. We called points for which the gradient was zero *stationary points.* For constrained optimizations problems such as (P), we need an alternative definition of what it means to be stationary. Instead of defining stationarity solely in terms of the function f, we have to consider stationary points of the problem (P). Once we've defined this new concept of stationarity for convex problems, we will then prove similar results as in the unconstrained case, namely that stationarity is a necessary condition for optimality, and that under certain conditions (convexity of f) it is also a sufficient condition for optimality.

Definition (Stationarity)**.** *Let f be a continuously differentiable function over a closed and convex set C. Then $\mathbf{x}^*$ is called a stationary point of* (P) *if*

$$\nabla f\left(\mathbf{x}^*\right)^\top \left(\mathbf{x} - \mathbf{x}^*\right) \geq 0\,, \qquad \textit{for any } \mathbf{x} \in C\,.$$

Intuition: this condition says that if $\mathbf{x}^*$ is stationary, then there are no feasible descent directions of f at $\mathbf{x}^*$, i.e., there is no direction $\mathbf{d}$ such that the directional derivative $f'(\mathbf{x}^*, \mathbf{d}) = \nabla f(\mathbf{x}^*)^\top \mathbf{d} < 0$ and $\mathbf{x}^* + \mathbf{d} \in C$. This suggests that stationarity is a necessary condition for $\mathbf{x}^*$ to be a local minima of (P).

Theorem (Stationarity as a Necessary Optimality Condition)**.** *Let f be a continuously differentiable function over a nonempty closed convex set C, and let $\mathbf{x}^*$ be a local minimum of* (P) *Then $\mathbf{x}^*$ is a stationary point of* (P)*.*

Proof. Let $\mathbf{x}^*$ be a local minimum of (P), and assume in contradiction that $\mathbf{x}^*$ is not a stationary point of (P). This implies that there exists $\mathbf{x} \in C$ such that

$$\nabla f\left(\mathbf{x}^*\right)^\top \left(\mathbf{x} - \mathbf{x}^*\right) < 0\,.$$

Thus, $f'\left(\mathbf{x}^*; \mathbf{d}\right) < 0$, where $\mathbf{d} = \mathbf{x} - \mathbf{x}^*$. Therefore, there exists $\varepsilon \in (0, 1)$ such that $f\left(\mathbf{x}^* + t\mathbf{d}\right) < f\left(\mathbf{x}^*\right)$, $\forall t \in (0, \varepsilon)$. Finally, since $\mathbf{x}^* + t\mathbf{d} = (1-t)\mathbf{x}^* + t\mathbf{x} \in C$, $\forall t \in (0, \varepsilon)$, we conclude that $\mathbf{x}^*$ is not a local optimum point of (**P**). This contradicts out initial assumption. □

Examples of Stationarity Conditions

- For unconstrained problems, i.e., where $C = \mathbb{R}^n$, we can show that the new concept of stationarity is equivalent to the stationarity condition we studied previously for unconstrained problems (i.e. that $\nabla f(\mathbf{x}^*) = 0$).

  Firstly, assume that $\mathbf{x}^*$ is a stationary point of (P), i.e., that

  $$\nabla f\left(\mathbf{x}^*\right)^\top \left(\mathbf{x} - \mathbf{x}^*\right) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n .$$

  Plugging $\mathbf{x} = \mathbf{x}^* - \nabla f\left(\mathbf{x}^*\right)$ in the above implies

  $$-\left\|\nabla f\left(\mathbf{x}^*\right)\right\|^2 \geq 0$$

  and thus $\nabla f\left(\mathbf{x}^*\right) = 0$.

  Conversely, if $\nabla f\left(\mathbf{x}^*\right) = 0$, then obviously the inequality in the new definition of stationarity is statisified. We've thus proved that when $C = \mathbb{R}^n$

  $$\nabla f\left(\mathbf{x}^*\right) = 0 \text{ if and only if } \nabla f\left(\mathbf{x}^*\right)^\top \left(\mathbf{x} - \mathbf{x}^*\right) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n .$$

- For $C = \mathbb{R}^n_+$, $\mathbf{x}^*$ is a stationary point iff

  $$\nabla f\left(\mathbf{x}^*\right)^\top \left(\mathbf{x} - \mathbf{x}^*\right) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n_+ .$$

  This is equivalent to

  $$\nabla f\left(\mathbf{x}^*\right)^\top \mathbf{x} - \nabla f\left(\mathbf{x}^*\right)^\top \mathbf{x}^* \geq 0 \text{ for all } \mathbf{x} \geq 0.$$

  Noting that

  $$\mathbf{a}^\top \mathbf{x} + b \geq 0 \text{ for all } \mathbf{x} \geq 0 \text{ if and only if } \mathbf{a} \geq 0 \text{ and } b > 0$$

  we can see that this is equivalent to $\nabla f\left(\mathbf{x}^*\right) \geq 0$ and $\nabla f\left(\mathbf{x}^*\right)^\top \mathbf{x}^* \leq 0$. This is then equivalent to

  $$\nabla f\left(\mathbf{x}^*\right) \geq 0 \qquad \text{and} \qquad x_i^* \frac{\partial f}{\partial x_i}\left(\mathbf{x}^*\right) = 0, \quad i = 1, 2, \ldots, n,$$

  which can be summarized as

  $$\frac{\partial f}{\partial x_i}\left(\mathbf{x}^*\right) \begin{cases} = 0 & x_i^* > 0 \\ \geq 0 & x_i^* = 0. \end{cases}$$

  If x_i^* is in the interior of C then $\mathbf{x}^*$ is only stationary if the partial derivative $\frac{\partial f}{\partial x_i}\left(\mathbf{x}^*\right)$ is zero, but if x_i^* is on the boundary of C (i.e. $x_i^* = 0$) then we only require the partial derivative to be non-negative[6].

[6] If it was negative, it would suggest that moving away from $x_i = 0$ to $x_i > 0$ would yield a smaller value of f.

The following table summarizes some important stationarity conditions:

Feasible Set	Explicit Stationarity Condition
$\mathbb{R}^n$	$\nabla f(\mathbf{x}^*) = \mathbf{0}$
$\mathbb{R}^n_+$	$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) \begin{cases} = 0 & x_i^* > 0 \\ \geq 0 & x_i^* = 0 \end{cases}$
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{e}^\top \mathbf{x} = 1\}$	$\frac{\partial f}{\partial x_1}(\mathbf{x}^*) = \ldots = \frac{\partial f}{\partial x_n}(\mathbf{x}^*)$
$B[\mathbf{0}, 1]$	$\nabla f(\mathbf{x}^*) = \mathbf{0}$ or $\lVert\mathbf{x}^*\rVert = 1$ and $\exists \lambda \leq 0 : \nabla f(\mathbf{x}^*) = \lambda \mathbf{x}^*$

where $\mathbf{e} = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}^\top$.

E.11

Prove the third result above.

We've shown for constrained optimization problems where the feasible set C is convex, that stationarity is a necessary condition for a point to be a local minima. For convex problems, i.e., when the objective f is also a convex function, stationarity is also a sufficient condition.

Theorem (Stationarity in Convex Optimization)**.** *Let f be a continuously differentiable convex function over a nonempty closed and convex set $C \subseteq \mathbb{R}^n$. Then $\mathbf{x}^*$ is a stationary point of* (P) *iff $\mathbf{x}^*$ is an optimal solution of* (P).

Proof. If $\mathbf{x}^*$ is an optimal solution of (P), then we already showed that it is a stationary point of (P). Assume that $\mathbf{x}^*$ is a stationary point of (P). Let $\mathbf{x} \in C$. Then

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq f(\mathbf{x}^*)$$

establishing the optimality of $\mathbf{x}^*$. The first inequality above follows from the gradient inequality for convex functions (Equation 1), and the second from the definition of stationarity. □

The Orthogonal Projection Operator

We now introduce the orthogonal projection operator $P_C(\mathbf{x})$, which returns the point in C closest to $\mathbf{x}$.

Definition (Orthogonal Projection)**.** *Given a nonempty closed convex set C, the orthogonal projection operator $P_C : \mathbb{R}^n \to C$ is defined by*

$$P_C(\mathbf{x}) = \underset{\mathbf{y} \in C}{\operatorname{argmin}} \left\{ \lVert \mathbf{y} - \mathbf{x} \rVert^2 \right\}.$$

Note that as $f(\mathbf{y}) = ||\mathbf{y} - \mathbf{x}||^2$ is a strictly convex function of $\mathbf{y}$, finding the orthogonal projection operator requires us to solve a convex optimization problem. Because the feasible set C is convex, we can prove that this problem has a unique optimal solution.

Theorem (The First Projection Theorem). *Let $C \subseteq \mathbb{R}^n$ be a nonempty closed and convex set. Then for any $\mathbf{x} \in \mathbb{R}^n$, the orthogonal projection $P_C(\mathbf{x})$ exists and is unique.*

Proof. $f(\mathbf{y}) = \|\mathbf{y} - \mathbf{x}\|^2$ is a coercive function, hence the minimum is attained in C, and moreover, $f(\mathbf{y})$ is strictly convex[7] hence there is only one optimal solution. □

Examples of Orthogonal Projections: In general, computing the orthogonal projection operator $P_C(\mathbf{x})$ can be difficult, but in some cases it can be computed explicitly.

- For $C = \mathbb{R}^n$ we have $P_C(\mathbf{x}) = \mathbf{x}$.
- For $C = \mathbb{R}^n_+$, we need to solve

  $$\begin{aligned} \min_{\mathbf{y}} ||\mathbf{y} - \mathbf{x}|| &= \sum (y_i - x_i)^2 \\ \text{s.t. } y_i &\geq 0 \text{ for all } i \end{aligned}$$

  The objective is a sum, and the constraints are separable, and so we need to just solve the problems

  $$\begin{aligned} \min_{y_i} &\quad (y_i - x_i)^2 \\ \text{s.t. } y_i &\quad \geq 0. \end{aligned}$$

  These are solved at

  $$y_i^* = \max(x_i, 0).$$

  Thus we have

  $$P_{\mathbb{R}^n_+}(\mathbf{x}) = [\mathbf{x}]_+$$

  where $[\mathbf{x}]_+ = (\max\{x_1, 0\}, \max\{x_2, 0\}, \ldots, \max\{x_n, 0\})^\top$.
- A box is a subset of $\mathbb{R}^n$ of the form

  $$B = [\ell_1, u_1] \times [\ell_2, u_2] \times \cdots \times [\ell_n, u_n] = \{\mathbf{x} \in \mathbb{R}^n : \ell_i \leq x_i \leq u_i\},$$

  where $\ell_i \leq u_i$ for all $i = 1, 2, \ldots, n$.

  E.12

  For this set show that

  $$[P_B(\mathbf{x})]_i = \begin{cases} u_i & x_i \geq u_i \\ x_i & \ell_i < x_i < u_i \\ \ell_i & x_i \leq \ell_i \end{cases}$$

[7] It is a quadratic function with positive definite matrix

$$f(\mathbf{y}) = (\mathbf{y} - \mathbf{x})^\top \mathbf{I}(\mathbf{y} - \mathbf{x})$$

.

- For the closed ball in $\mathbb{R}^n$, $C = B[0, r]$, we can show that

$$P_{B[0,r]} = \begin{cases} \mathbf{x} & \|\mathbf{x}\| \leq r \\ r\frac{\mathbf{x}}{\|\mathbf{x}\|} & \|\mathbf{x}\| > r \end{cases}$$

There is an important geometric characterization of the projection operator. It says that for $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in C$, the angle between $\mathbf{x} - P_C(\mathbf{x})$ and $\mathbf{y} - P_C(\mathbf{x})$ is greater than or equal to 90°.

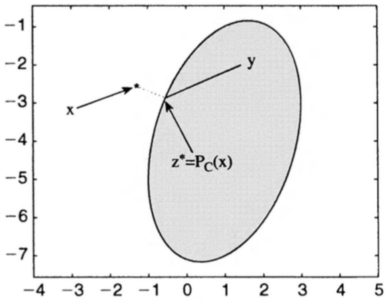

Figure 2: The orthogonal projection operator onto a convex set C. If $\mathbf{z}^* = P_C(\mathbf{x})$, then the angle between $\mathbf{x} - \mathbf{z}^*$ and $\mathbf{y} - \mathbf{z}^*$ for $\mathbf{y} \in C$ is always greater than 90°.

Theorem (The Second Projection Theorem). *Let C be a nonempty closed convex set and let $\mathbf{x} \in \mathbb{R}^n$. Then $\mathbf{z} = P_C(\mathbf{x})$ if and only if*

$$(\mathbf{x} - \mathbf{z})^\top(\mathbf{y} - \mathbf{z}) \leq 0\,, \quad \textit{for any } \mathbf{y} \in C$$

Proof. $P_C(\mathbf{x})$ is the solution of the convex optimization problem

$$\begin{array}{ll} \min_{\mathbf{y}} & f(\mathbf{y}) = ||\mathbf{y} - \mathbf{x}||^2 \\ \text{s.t.} & \mathbf{y} \in C. \end{array}$$

Thus $\mathbf{z} = P_C(\mathbf{x})$ if and only if it is a stationary point, i.e.,

$$\nabla f(\mathbf{z})^\top(\mathbf{y} - \mathbf{z}) \geq 0 \text{ for all } \mathbf{y} \in C.$$

But $\nabla f(\mathbf{z}) = 2(\mathbf{z} - \mathbf{x})$ giving the result as desired. □

Finally, an important result in convex optimization is the representation of stationarity using the orthogonal projection operator.

Theorem (Representation of Stationarity via the Orthogonal Projection Operator). *Let f be a continuously differentiable function over the nonempty closed convex set C, and let $s > 0$. Then x^* is a stationary point of* (P) *if and only if*

$$\mathbf{x}^* = P_C\left(\mathbf{x}^* - s\nabla f\left(\mathbf{x}^*\right)\right)\,.$$

Proof. By the second projection theorem, $\mathbf{x}^* = P_C\left(\mathbf{x}^* - s\nabla f\left(\mathbf{x}^*\right)\right)$ iff

$$\left(\mathbf{x}^* - s\nabla f\left(\mathbf{x}^*\right) - \mathbf{x}^*\right)^\top \left(\mathbf{y} - \mathbf{x}^*\right) \leq 0 \text{ for any } \mathbf{y} \in C,$$

which is equivalent to

$$\nabla f\left(\mathbf{x}^*\right)^\top \left(\mathbf{y} - \mathbf{x}^*\right) \geq 0 \text{ for any } \mathbf{y} \in C,$$

namely, the definition of stationarity. □

The Gradient Projection Method

The result that x^* is a stationary point of (P) if and only if

$$\mathbf{x}^* = P_C\left(\mathbf{x}^* - s\nabla f\left(\mathbf{x}^*\right)\right) .$$

suggests we can find stationary points by solving this fixed-point problem. The gradient projection method is a gradient descent type algorithm that uses the orthogonal projection operator to solve this problem:

Algorithm 1: The Gradient Projection Method

```
Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in C$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:
1 Pick a stepsize $t^k$ by a line search procedure.
2 Set $\mathbf{x}^{k+1} = P_C\left(\mathbf{x}^k - t^k\nabla f\left(\mathbf{x}^k\right)\right)$ .
3 If $\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.
```

Another way of writing this is to split step 2 into two steps:

2a $\mathbf{y}^{k+1} = \mathbf{x}^k - t^k\nabla f\left(\mathbf{x}^k\right)$

2b $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in C} ||\mathbf{x} - \mathbf{y}^{k+1}||$

Note that in the unconstrained case $P_C(\mathbf{x}) = \mathbf{x}$ and so the algorithm simplifies to

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k\nabla f\left(\mathbf{x}^k\right),$$

i.e., the gradient descent method we studied at the start of the module. In the constrained case, the algorithm essentially does the usual gradient descent step $\mathbf{x}^k - t^k\nabla f\left(\mathbf{x}^k\right)$ and then projects this point back into C using the orthogonal projection P_C if it lies outside of C.

As before, there are several strategies for choosing the stepsizes t^k. When $f \in C_L^{1,1}$, we can choose t^k to be constant and equal to $\frac{1}{L}$. An alternative is to use backtracking.

Before describing backtracking, lets first define the gradient mapping to be

$$G_M(\mathbf{x}) = M\left[\mathbf{x} - P_C\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right)\right],$$

where $M > 0$ is a positive constant. In the unconstrained case $G_M(\mathbf{x}) = \nabla f(\mathbf{x})$, so the gradient mapping is an extension of the usual gradient operation. By the previous theorem, $G_M(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x}$ is a stationary point of (P). This means that we can consider $\|G_M(\mathbf{x})\|^2$ to be an optimality measure, i.e., for the stationary points $\|G_M(\mathbf{x})\|^2 = 0$. Otherwise $\|G_M(\mathbf{x})\|^2 > 0$.

The constrained version of the backtracking rule we studied in the unconstrained case, gives the following algorithm:

```
Algorithm 2: The Gradient Projection Method with Backtracking
Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in C$. Parameters $s > 0$, $\alpha \in (0,1)$,
                and $\beta \in (0,1)$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:
Pick $t^k = s$.
While $f(\mathbf{x}^k) - f(P_C(\mathbf{x}^k - t^k \nabla f(\mathbf{x}^k))) < \alpha t^k \left\| G_{\frac{1}{t^k}}(\mathbf{x}^k) \right\|^2$, set $t^k := \beta t^k$.
Set $\mathbf{x}^{k+1} = P_C\left(\mathbf{x}^k - t^k \nabla f\left(\mathbf{x}^k\right)\right)$.
If $\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.
```

Finally, if we make some assumptions about f then it is possible to prove (but beyond the scope of the module) that the gradient projection method will converge.

Theorem (Convergence of the Gradient Projection Method). *Let $\left\{\mathbf{x}^k\right\}$ be the sequence generated by the gradient projection method for solving problem* (P) *with either a constant stepsize $\bar{t} \in \left(0, \frac{2}{L}\right)$, where L is a Lipschitz constant of ∇f (i.e. $f \in C_L^{1,1}(C)$) or a backtracking stepsize strategy. Assume that f is bounded below. Then:*

1. *The sequence $\left\{f\left(\mathbf{x}^k\right)\right\}$ is nonincreasing.*
2. *$G_d\left(\mathbf{x}^k\right) \longrightarrow 0$ as $k \longrightarrow \infty$, where*

$$d = \begin{cases} 1/\bar{t} & \textit{constant stepsize} \\ 1/s & \textit{backtracking.} \end{cases}$$

Checklist

The idea of this checklist is to help you to self-evaluate your progress and understanding of the subject, and to give you some guidance on where to focus. If you can tick all the boxes it means you’re doing alright, otherwise you need to study a bit more, grab a book, watch the videos, or seek help from classmates, the lecturers, or the demonstrators. Try to fill as many gaps as quickly as possible.

And remember to do the E.13 ’s!

Learning Outcome	Check
I can state the definition of convex function and its different characterizations.	
I can show whether a function is convex or not.	
I understand the links between convexity, minimizers, and maximizers.	
I can identify a convex optimization problem.	
I understand stationarity in convex constrained optimization as a generalization of the stationarity concept studied in Week 3.	
I can compute stationarity conditions for different constraints.	
I understand the link between stationarity, orthogonal projection, and projected gradient descent.	