CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Convex Optimization

Date: November 10th, 2025 *Author: Surbhi Goel*

The central goal of continuous optimization is to solve problems of the form:

$$\min_{x \in \mathbb{R}^n} f(x)$$

Critical points. From calculus, a necessary condition for x^* to be a local minimum is that the gradient vanishes:

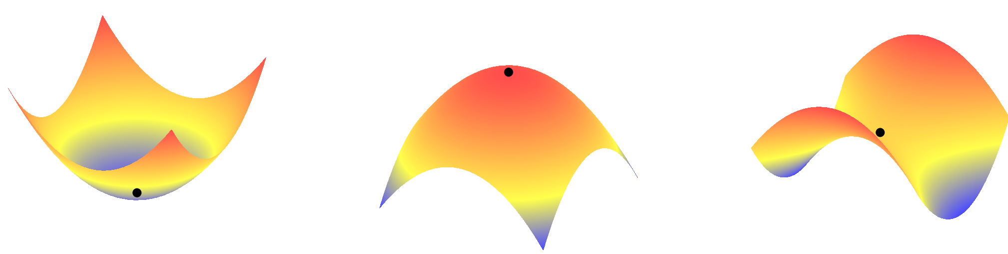
$$\nabla f(x^*) = 0$$

Such a point is called a *critical point*. However, this condition is not sufficient—a critical point could be a local minimum, a local maximum, or a saddle point.

The second derivative test. To classify a critical point x_0, we examine the Hessian $\nabla^2 f(x_0)$:

- If $\nabla^2 f(x_0) \succ 0$ (positive definite), x_0 is a *strict local minimum*.
- If $\nabla^2 f(x_0) \prec 0$ (negative definite), x_0 is a *strict local maximum*.
- If $\nabla^2 f(x_0)$ is *indefinite*, x_0 is a *saddle point*.

Note: The notation $\nabla^2 f(x_0) \succ 0$ means that the Hessian is positive definite, i.e., all its eigenvalues are positive. Similarly, $\nabla^2 f(x_0) \prec 0$ means that the Hessian is negative definite, i.e., all its eigenvalues are negative. The notation $\nabla^2 f(x_0)$ is indefinite if it is neither positive nor negative definite, i.e., it has both positive and negative eigenvalues.



Local Minimum (Curves up in all directions) | Local Maximum (Curves down in all directions) | Saddle Point (Up one way, down another)

Figure 1: Illustration of critical points on 2D surfaces.

The fundamental challenge. Even if we identify a local minimum, we have no guarantee it is the *global minimum*. A general function can have many local minima, and finding the best one is computationally intractable.

Is there a class of functions where local minima are automatically global?

The answer is yes: **convex functions**.

1 Convex Sets and Convex Functions

Definition 1 (Convex Set)**.** *A set $\mathcal{C} \subseteq \mathbb{R}^n$ is **convex** if for any two points $x, y \in \mathcal{C}$, the line segment connecting them is entirely contained in $\mathcal{C}$. Formally, for any $\alpha \in [0, 1]$,*

$$\alpha x + (1 - \alpha)y \in \mathcal{C}$$

Definition 2 (Convex Function)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ defined on a convex set $\mathcal{C}$ is **convex** if the line segment between any two points on its graph lies on or above the graph. Formally, for any $x, y \in \mathcal{C}$ and any $\alpha \in [0, 1]$,*

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

This is known as Jensen's inequality.

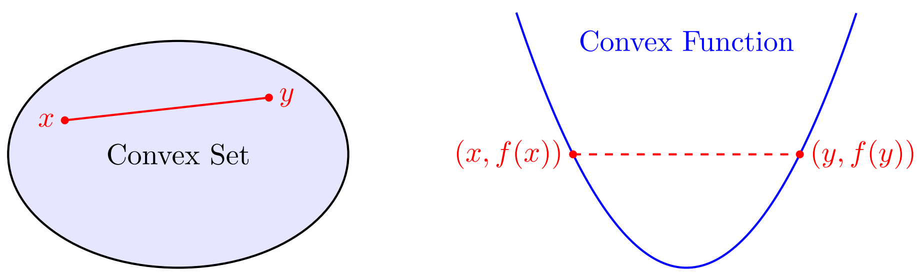


Figure 2: Illustration of a convex set (left) and a convex function (right).

2 Aside: Taylor's Theorem

Before we can characterize convex functions using derivatives, we need a more precise version of Taylor's theorem than what we saw in the previous lecture. The key difference is that we need an *exact* formula for the error, not just the asymptotic $o(\|h\|^2)$ notation. Taylor's theorem provides this exact formula for the remainder term.

Theorem 3 (Taylor's Theorem with Remainder)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. For any $x, y \in \mathbb{R}^n$, there exists a point z on the line segment between x and y such that:*

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(z)(y - x)$$

The final term, involving the Hessian evaluated at an intermediate point z, is the *exact* remainder. It is not an approximation.

Proof Sketch. The proof is a clever trick that reduces the multivariate problem to a single-variable one. We define a new function $g(t) = f(x + t(y - x))$ that represents the value of f along the straight line from x to y. Since $g(t)$ is a simple univariate function, we can apply the standard Taylor's theorem from single-variable calculus. By using the chain rule to relate the derivatives of g back to the gradient and Hessian of f, we recover the multivariate formula.

3 Equivalent Conditions for Convexity

We can now state equivalent conditions for convexity that are easier to check than the definition.

Definition 4 (First-Order Condition)**.** *A differentiable function f is convex if and only if it lies above its tangent hyperplane at any point:*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \textit{for all } x, y \in \mathcal{C}$$

Definition 5 (Second-Order Condition)**.** *A twice-differentiable function f is convex if and only if its Hessian is positive semidefinite everywhere:*

$$\nabla^2 f(x) \succeq 0 \quad \textit{for all } x \in \mathcal{C}$$

These three conditions—the definition, the first-order condition, and the second-order condition—are equivalent for differentiable functions. Here we prove a few of the implications.

Definition $\implies$ First-Order. Assume f is convex. By definition, for $\alpha \in [0, 1]$,

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

Rearranging this gives:

$$f(x + \alpha(y - x)) - f(x) \leq \alpha(f(y) - f(x))$$

Dividing by α (for $\alpha > 0$):

$$\frac{f(x + \alpha(y - x)) - f(x)}{\alpha} \leq f(y) - f(x)$$

Taking the limit as $\alpha \to 0^+$, the left side becomes the directional derivative of f at x in the direction $y - x$, which is $\nabla f(x)^\top (y - x)$. This yields:

$$\nabla f(x)^\top (y - x) \leq f(y) - f(x)$$

which is exactly the first-order condition. □

Second-Order $\implies$ First-Order. Using Taylor's Theorem with remainder:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(z)(y - x)$$

If $\nabla^2 f(x) \succeq 0$ for all x, then at the intermediate point z, the remainder term is non-negative:

$$\frac{1}{2}(y - x)^\top \nabla^2 f(z)(y - x) \geq 0$$

Therefore, $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$, which is the first-order condition. □

4 Properties of Convex Functions

Optimality Condition. Now we can state the optimality condition for convex functions.

Property 6 (Optimality Condition)**.** *For a convex function f, a point x^* is a global minimum if and only if $\nabla f(x^*) = 0$.*

Proof. If $\nabla f(x^*) = 0$, the first-order condition gives: $f(y) \geq f(x^*) + 0^\top (y - x^*) = f(x^*)$ for all y. Thus x^* is a global minimum. □

This is the main payoff: for convex functions, finding a critical point guarantees we have found a global minimum.

Operations that Preserve Convexity Convex functions are very broad class of functions that are very well-behaved. Following operations preserve convexity:

- **Non-negative weighted sum:** If $f_1, \ldots, f_k$ are convex and $w_1, \ldots, w_k \geq 0$, then $f(x) = \sum_{i=1}^k w_i f_i(x)$ is convex.
- **Pointwise maximum:** If $f_1, \ldots, f_k$ are convex, then $f(x) = \max(f_1(x), \ldots, f_k(x))$ is convex.
- **Composition with an affine function:** If f is convex, then $g(x) = f(Ax + b)$ is convex.

Example: Least Squares Loss The least squares loss is a convex function. It is defined as:

$$f(\theta) = \frac{1}{2}\|X\theta - y\|_2^2$$

The Hessian of the least squares loss is:

$$\nabla^2 f(\theta) = X^\top X$$

Since $X^\top X$ is always positive semidefinite, the least squares loss is always convex.

5 Strong Convexity

Some convex functions can be very flat, which can lead to slow convergence for optimization algorithms. *Strong convexity* is a stronger condition that rules this out by enforcing that the function has a certain minimum amount of curvature everywhere.

Definition 7 (Strong Convexity)**.** *A function f is μ-**strongly convex** for some $\mu > 0$ if its Hessian is uniformly positive definite:*

$$\nabla^2 f(x) \succeq \mu I \quad \textit{for all } x$$

This is equivalent to saying that the function $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

A strongly convex function is bounded below by a quadratic. Using Taylor's Theorem, this implies:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

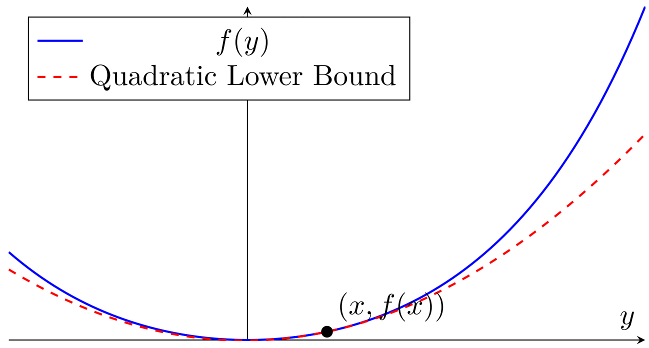


Figure 3: A μ-strongly convex function is lower-bounded by a quadratic function at every point.

Gradient descent on a strongly convex function converges faster than on a convex function.