EPFL CS439: Optimization for Machine Learning

Note: I am not affiliated with EPFL. These notes are based on the online course material provided by Prof. Martin Jaggi.

Siddhartha Bhattacharya

January 16, 2025

Contents

0.1 High Level Overview

General unconstrained minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) \tag{1}$$

- $x \in \mathbb{R}^d$ refers to candidate solutions, variables, or parameters. $\mathbb{R}^d$ is the domain.

- $f : \mathbb{R}^d \to \mathbb{R}$ is the objective function.
- typical assumptions: f is continuous and differentiable.

0.1.1 How to Optimize

Two main steps:
Mathematical Modeling: Defining & modeling the optimization problem.
Computational Optimization: Running an (approximate) optimization algorithm.

1 Theory of Convex Optimization

1.1 Warm-up: Cauchy-Schwarz Inequality

Theorem 1.1 (Cauchy-Schwarz Inequality)**.** *Let $u, v \in \mathbb{R}^d$. Then*

$$\langle u, v\rangle^2 \leq \|u\|^2 \|v\|^2 \tag{2}$$

Equivalently,

$$|\langle u, v\rangle| \leq \|u\| \|v\| \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the standard Euclidean dot product.

For nonzero u, v, this is equivalent to

$$-1 \leq \frac{\langle u, v\rangle}{\|u\| \|v\|} \leq 1 \tag{4}$$

So, the angle α between u and v is given by $\cos(\alpha) = \frac{\langle u, v\rangle}{\|u\|\|v\|}$. Thus, equality holds in (2) if and only if u and v are scalar multiples of each other.

1.2 What is Convexity? Convex Sets and Functions

Definition 1.2. A set C is a **convex set** if the line segment between any two points of C lies entirely in C. Formally, for any $x, y \in C$ and $0 \leq \lambda \leq 1$, we have

$$\lambda x + (1 - \lambda) y \in C \tag{5}$$

Proposition 1.3. Intersections of convex sets are convex.

Remark 1.4. Unions of convex sets are not necessarily convex.

Proposition 1.5 (later: projections onto convex sets)**.**

$$P_C(x') = \arg\min_{y \in C} \|y - x'\|$$

Definition 1.6. A function $f : \mathbb{R}^d \to \mathbb{R}$ is a **convex function** if

1. $\mathbf{dom}(f)$ is a convex set
2. for all $x, y \in \mathrm{dom}(f)$ and $0 \leq \lambda \leq 1$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{6}$$

Geometrically: the line segment connecting $(x, f(x))$ and $(y, f(y))$ lies above the graph of f.

1.3 Proving Convexity

Convex optimization problems are of the form

$$\min f(x) \text{ s.t. } x \in C \tag{7}$$

where both f is a convex function and $C \subseteq \mathrm{dom}(f)$ is a convex set.

Crucial property of convex optimization problems: every local minimum is a global minimum.

1.3.1 Solving Convex Optimization – Provably

For convex optimization problems, all algorithms

- coordinate descent, gradient descent, SGD, projected and proximal gradient descent

do converge to the global optimum! (assuming f is differentiable)

Definition 1.7. A **graph** of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\{(x, f(x)) \in \mathbb{R}^{d+1} \mid x \in \mathrm{dom}(f)\}$$

Definition 1.8. The **epigraph** of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{epi}(f) = \{(x, \alpha) \in \mathbb{R}^{d+1} \mid x \in \mathrm{dom}(f), \alpha \geq f(x)\}$$

Visually, the epigraph is the set of points above the graph of f.

Proposition 1.9. A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if its epigraph is a convex set.

Proof. First, let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then, for any $x, y \in \text{dom}(f)$ and $0 \leq \lambda \leq 1$, we have

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

Thus, for any $(x, \alpha), (y, \beta) \in \text{epi}(f)$ and $0 \leq \lambda \leq 1$, we have

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

Therefore, $(\lambda x + (1-\lambda)y, \lambda\alpha + (1-\lambda)\beta) \in \text{epi}(f)$, so $\text{epi}(f)$ is convex.

For the converse, suppose that $\text{epi}(f)$ is convex. Let $x, y \in \text{dom}(f)$ and $0 \leq \lambda \leq 1$. We know that $(x, f(x)), (y, f(y)) \in \text{epi}(f)$ by definition of the epigraph (let $\alpha = f(x)$ and $\beta = f(y)$). By the convvexity of $\text{epi}(f)$, any convex combination of x, y is also in $\text{epi}(f)$.

So, the convex combination $(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda)f(y)) \in \text{epi}(f)$. Therefore, by the definition of $\text{epi}(f)$,

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

and thus, f is convex. $\square$

Example 1.10. Examples of convex functions:

- Linear functions: $f(x) = a^T x$
- Affine functions: $f(x) = a^T x + b$
- Exponential functions: $f(x) = e^{ax}$
- Norms: every norm on $\mathbb{R}^d$ is convex

Proof. Proof of convexity of norms:

By the triangle inequality, $\|x + y\| \leq \|x\| + \|y\|$ and homogeneity of a norm,

$\|\lambda x\| = |\lambda| \, \|x\|$, we have that for any $x, y \in \mathbb{R}^d$ and $0 \leq \lambda \leq 1$,

$$\begin{aligned} \|\lambda x + (1-\lambda)y\| &\leq \|\lambda x\| + \|(1-\lambda)y\| \\ &= |\lambda| \, \|x\| + |1-\lambda| \, \|y\| \\ &= \lambda \, \|x\| + (1-\lambda) \, \|y\| \end{aligned}$$

$\square$

Lemma 1.11 (Jensen's Inequality). *Let f be convex, $x_1, \ldots, x_m \in \mathbf{dom}(f)$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^m \lambda_i = 1$. Then*

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \leq \sum_{i=1}^m \lambda_i f(x_i) \tag{8}$$

Proof. First, since f is convex, we have that for any $x, y \in \mathbf{dom}(f)$ and $0 \leq \lambda \leq 1$,

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

We proceed by induction. For the 'base case', let $k = 2$ and $0 \leq \lambda_1, \lambda_2 \leq 1$ such that $\lambda_1 + \lambda_2 = 1$. Note that this implies $\lambda_2 = 1 - \lambda_1$.

$$\begin{aligned} f\left(\sum_{i=1}^2 \lambda_i x_i\right) =& f(\lambda_1 x_1 + \lambda_2 x_2) \\ =& f(\lambda_1 x_1 + (1-\lambda_1) x_2) \end{aligned}$$

Next, since f is convex,

$$f(\lambda_1 x_1 + (1-\lambda_1)x_2) \leq \lambda_1 f(x_1) + (1-\lambda_1) f(x_2)$$

Now, substituting $\lambda_2 = (1-\lambda_1)$, we get that

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2)$$

So, (8) holds for $k = 2$. Now, suppose for induction that (8) holds for an arbitrary $k \geq 2$. We show that it holds for $k+1$. Let $1 = \sum_{i=1}^{k+1} \lambda_i$. If we let $\beta = \sum_{i=1}^k \lambda_i$, then $\lambda_{k+1} = 1 - \beta$. Using this, we can rewrite the convex combination $\sum_{i=1}^{k+1} \lambda_i x_i$ as:

$$\begin{aligned}\sum_{i=1}^{k+1} \lambda_i x_i =& \beta \left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} x_i \right) + \lambda_{k+1} x_{k+1} \\ =& \beta \left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} x_i \right) + (1-\beta) x_{k+1}\end{aligned}$$

And since f is convex,

$$f\left(\beta \left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} x_i \right) + (1-\beta) x_{k+1} \right) \leq \beta f \left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} x_i \right) + (1-\beta) f(x_{k+1})$$

And by inductive hypothesis on the first k terms, we know that $f\left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} x_i\right) \leq \sum_{i=1}^{k} \frac{\lambda_i}{\beta} f(x_i)$, which is a valid convex combination, since we defined $\beta = \sum_{i=1}^{k} \lambda_i$, implying that $\sum_{i=1}^{k} \frac{\lambda_i}{\beta} = \frac{\sum_{i=1}^{k} \lambda_i}{\beta} = \frac{\beta}{\beta} = 1$.

Thus,

$$\beta f \left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} x_i \right) + (1-\beta) f(x_{k+1}) \leq \beta \left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} f(x_i) \right) + (1-\beta) f(x_{k+1})$$

And by the definition of the convex combination, $\beta\left(\sum_{i=1}^{k} \frac{\lambda_i}{\beta} f(x_i)\right) + (1-\beta) f(x_{k+1}) = \sum_{i=1}^{k+1} \lambda_i f(x_i)$, we are done. $\square$

Remark 1.12. For $m = 2$, Jensen's inequality reduces to the definition of convexity. Jensen's inequality is a general definition for convex combinations of any number of points in the domain.

Lemma 1.13. *Let f be convex and suppose that* $\mathbf{dom}(f)$ *is open. Then f is continuous.*

1.4 Characterizations of Convexity

Definition 1.14 (Differentiable Functions)**.** Graph of the affine function $f(x) + \nabla f(x)^T (y - x)$ is a tangent hyperplan to the graph of f at $(x, f(x))$.

Lemma 1.15 (First-order Characterization of Convexity)**.** *Suppose that* $\mathbf{dom}(f)$

is open and f is differentiable; in particular, the gradient (vector of partial derivatives)

$$\nabla f(x) := \left[\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_d}(x)\right]$$

exists at ever point $x \in \mathbf{dom}(f)$. Then f is convex if and only if $\mathbf{dom}(f)$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{9}$$

holds for all $x, y \in \mathbf{dom}(f)$.

Lemma 1.16 (Second-order Characterization of Convexity)**.** *Suppose that $\mathbf{dom}(f)$ is open and f is twice differentiable; In particular, the Hessian (matrix of second partial derivatives)*

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(x) \end{bmatrix}$$

exists at every point $x \in \mathbf{dom}(f)$ and is symmetric. Then f is convex if and only if for all $x \in \mathbf{dom}(f)$, we have

$$\nabla^2 f(x) \succeq 0 \;\; i.e. \nabla^2 f(x) is\ positive\ semidefinite \tag{10}$$

Recall that a matrix A is positive semidefinite if for all $z \in \mathbb{R}^d$, we have $z^T A z \geq 0$.

Connection to positive operators. Let us regard the hessian matrix A as the matrix representation of a linear operator $T \in \mathcal{L}(V)$, i.e. $A = M(T, B)$ for some basis B of $\mathbb{R}^d$. Then. f is convex if and only if T is a positive operator, i.e. $\langle Tx, x\rangle \geq 0$ for all $x \in \mathbb{R}^d$ and T is self adjoint (by definition of the Hessian). So, $\langle Tx, x\rangle = x^T A x$.

Example 1.17. Let $f(x_1, x_2) = x_1^2 + x_2^2$. Then, $\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. Which is positive semidefinite, so f is convex.

1.4.1 Operations that Preserve Convexity

Lemma 1.18. *Let $f_1, f_2, \ldots, f_m$ be convex functions and $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$. Then $f := \sum_{i=1}^m \lambda_i f_i$ is convex on $\mathbf{dom}(f) := \bigcap_{i=1}^m \mathbf{dom}(f_i)$.*

Lemma 1.19. *Let f be a convex function with $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ and let $g : \mathbb{R}^m \to \mathbb{R}^d$ be an affine function, meaning that $g(x) = Ax+b$ for some matrix $A \in \mathbb{R}^{d\times m}$ and $b \in \mathbb{R}^d$. Then the function $f \circ g$ (that maps $x \to f(Ax+b)$) is convex on $\mathbf{dom}(f \circ g) := \{x \in \mathbb{R}^m : g(x) \in \mathbf{dom}(f)\}$.*

1.5 Local Minima are Global Minima

Definition 1.20. A **local minimum** of $f : \mathbf{dom}(f) \to \mathbb{R}$ is a point x such that there exists $\epsilon > 0$ with

$$f(x) \leq f(x) \forall y \in \mathbf{dom}(f) \text{ satisfying } \|y - x\| \leq \epsilon \tag{11}$$

Lemma 1.21. *Let x^* be be a local minimum of a convex function $f : \mathbf{dom}(f) \to \mathbb{R}$. Then x^* is a global minimum, meaning that $f(x^*) \leq f(y) \forall y \in \mathbf{dom}(f)$.*

Proof. Let x^* be a local minimum to a convex function f. Then suppose for contradiction that there exists another $y \in \mathbf{dom}(f)$ such that $f(y) < f(x^*)$.

Then let $y' = \lambda y + (1-\lambda)x^*$ for some $0 < \lambda < 1$. Since $f(y) < f(x^*)$, it follos that $\lambda f(y) + (1-\lambda) f(x^*) < f(x^*)$, so $f(y') < f(x^*)$. Now we show that y' is within the ϵ-neighborhood of x^*. Recall that by the definition of a local minimum, there exists $\epsilon > 0$ such that $f(x^*) \leq f(x) \forall x \in \mathbf{dom}(f)$ satisfying $\|x - x^*\| \leq \epsilon$.

For any $\epsilon > 0$, we may choose a small enough λ such that $\|y' - x^*\| \leq \epsilon$. First, let us expand the norm of $y' - x^*$:

$$\begin{aligned} \|y' - x^*\| &= \|\lambda y + (1-\lambda)x^* - x^*\| \\ &= \|\lambda y - \lambda x^*\| \\ &= \lambda \|y - x^*\| \end{aligned}$$

Now, if we let $\lambda = \frac{\epsilon}{\|y - x^*\|} > 0$ (since $y \neq x^*$), then

$$\begin{aligned} \|y' - x^*\| &= \frac{\epsilon}{\|y - x^*\|} \|y - x^*\| \\ &= \epsilon \end{aligned}$$

Thus, $f(y') \leq f(x^*)$ and $\|y' - x^*\| \leq \epsilon$, which contradicts the assumption that x^* is a local minimum. $\square$

Lemma 1.22 (Critical Points are Global Minima)**.** *Suppose that f is convex and differentiable over an open domain* $\mathbf{dom}(f)$*. Let* $x \in \mathbf{dom}(f)$*. If* $\nabla f(x) = 0$ *(**critical point**), then x is the global minimum.*

Proof. Suppose that $\nabla f(x) = 0$. According to the lemme on the first-order characterization of convexity, we have that

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) = f(x)$$

for all $y \in \mathbf{dom}(f)$. Thus, x is a global minimum. □

1.6 Strict Convexity

Definition 1.23. A function $f : \mathbf{dom}(f) \to \mathbb{R}$ is **strictly convex** if for all $x \neq y \in \mathbf{dom}(f)$ and all $\lambda \in (0, 1)$, we have

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \tag{12}$$

This differs from the definition of convexity in that the inequality is strict.

Lemma 1.24. *If f is strictly convex, then f has at most one global minimum.*

1.7 Constrained Minimization

Definition 1.25. Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and let $X \subseteq \mathbf{dom}(f)$ denote the constraint (or feasible) set. be a convex set. A point $x \in X$ is a **minimizer** of f **over** X if

$$f(x) \leq f(y) \forall y \in X$$

Lemma 1.26. *Suppose that* $f : \mathbf{dom}(f) \to \mathbb{R}$ *is convex and differentiable over an open domain* $\mathbf{dom}(f) \subseteq \mathbb{R}^d$*, and let* $X \subseteq \mathbf{dom}(f)$ *be a convex set. A point* $x^* \in X$ *is a **minimizer of*** $\mathbf{f}$ ***over*** $\mathbf{X}$ *if and only if*

$$\nabla f(x^*)^T (x - x^*) \geq 0 \;\; \forall x \in X \tag{13}$$