Probabilistic Generative Models — Convex analysis

Caio Corro

These notes contain the basic knowledge about convex analysis required for the Probabilistic Generative Models course. Some notions are not properly defined and can be ignored for the scope of this course (*e.g.* the relative interior of a set). The objective is to stay on simple notions to understand the methods studied in the course, without getting burdened with more complex notions that do not improve the clarity of the presentation. More details on convex analysis and optimization can be found in [Boyd et al., 2004], [Beck, 2017] and [Borwein and Lewis, 2006].

1 Convex sets and functions

Convex functions are interesting to study for (at least) two reasons:

1. any local minimizer of a convex function is also a global minimizer (if it exists!), which allows to derive "simple" optimality conditions that are both necessary and sufficient (Section 2);
2. we can derive parameterized bounds on convex functions that will be the basis of variational methods for inference.

Convex functions are defined as follows.

Definition 1 (Convex set)**.** Let $U \subseteq \mathbb{R}^k$ be a set. The set U is convex if and only if:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in U, \epsilon \in [0,1]: \quad \epsilon \boldsymbol{u} + (1-\epsilon)\boldsymbol{v} \in U.$$

Definition 2 (Convex function)**.** Let $f : U \to \mathbb{R}$, $U \subseteq \mathbb{R}^k$, be a function. The function f is convex if and only if:

1. U is a convex set;
2. the function satsifies the following inequality:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in U, \epsilon \in [0,1]: \quad f(\epsilon \boldsymbol{u} + (1-\epsilon)\boldsymbol{v}) \leq \epsilon f(\boldsymbol{u}) + (1-\epsilon) f(\boldsymbol{v})$$

A function f is concave if and only if $-f$ is convex.

If the domain of a function is an open set and it is (twice) differentiable everywhere, convexity can also be characterized in terms of its gradient (or Hessian).

Proposition 1 (Convexity: first-order condition)**.** Let $U \subseteq \mathbb{R}^k$ be a convex open set and $f : U \to \mathbb{R}$ be a differentiable function. Then, f is convex if and only if:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in U: \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \nabla f(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle.$$

Proposition 2 (Convexity: second-order condition)**.** Let $U \subseteq \mathbb{R}^k$ be a convex open set and $f : U \to \mathbb{R}$ be a twice differentiable function. Then, f is convex if and only if the Hessian $\nabla^2 f(\boldsymbol{u})$ is a positive semidefinite matrix, for all $\boldsymbol{u} \in U$.

Proof of Propositions 1 and 2 can be found in [Boyd et al., 2004, Section 3.1.3 and Exercise 3.8]. Note that the second order condition simplifies to non-negative second-order derivatives in the case of $U \subseteq \mathbb{R}$.

Two other important characteristics of functions are properness and closedness. A function $f : U \to [-\infty, +\infty]$ is a proper function if and only if there is no $\boldsymbol{u} \in U$ s.t. $f(\boldsymbol{u}) = -\infty$ and there exists at least one $\boldsymbol{u}' \in U$ s.t. $f(\boldsymbol{u}') \neq +\infty$. The closedness property is equivalent to lower semicontinuity. We won't study this property in this course, more information can be found in [Beck, 2017, Setion 2.1].

Note that a convex function may not be differentiable everywhere. This the case of the absolute value function $f(u) = |u|$, which is not differentiable at 0. In the case of convex functions, the concept of subgradient is a very useful generalization of the gradient,[1] that will be used to define optimality conditions, among others.

Definition 3 (Subgradient)**.** Let $U \in \mathbb{R}^k$ be a set and $f : U \to \mathbb{R}$ a convex function. Then, a vector $\boldsymbol{g} \in \mathbb{R}^k$ is a subgradient of f at $\boldsymbol{u} \in U$ if and only if:

$$\forall \boldsymbol{v} \in U : \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{v} - \boldsymbol{u} \rangle.$$

The set of subgradient at a given point is called the subdifferential and is denoted as $\partial f(\boldsymbol{u})$.

Proposition 3 (Subgradient properties)**.** Let U be a convex set and $f : U \to \mathbb{R}$ be a convex function. The subgradients of f satisfy the following properties:

1. $\partial f(\boldsymbol{u}), \boldsymbol{u} \in U$, is a convex set;
2. if f is a proper function and $\boldsymbol{u} \in \operatorname{int}(\operatorname{dom} f)$, then $\partial f(\boldsymbol{u}) \neq \emptyset$, *i.e.* there exists at least one subgradient at $\boldsymbol{u}$ [Beck, 2017, Theorem 3.14];
3. if f is a proper function, $\boldsymbol{u} \in \operatorname{int}(\operatorname{dom} f)$ and f differentiable at $\boldsymbol{u}$, then $\partial f(\boldsymbol{u}) = \{\nabla f(\boldsymbol{u})\}$.

It practice, it is often useful to work with extended-value extensions of a function. Let $U \subseteq \mathbb{R}^k$ be a set and $f : U \to \mathbb{R}$ be a function. The extended-value extensions of f, denoted $\widetilde{f} : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$, is defined as follows:

$$\widetilde{f}(\boldsymbol{u}) = \begin{cases} f(\boldsymbol{u}) & \text{if } \boldsymbol{u} \in \operatorname{dom} f, \\ \infty & \text{otherwise.} \end{cases}$$

The domain of a function is then defined as the set of input whose images are not equal to infinity (but potentially equal to minus infinity), *i.e.* $\operatorname{dom} \widetilde{f} = \{\boldsymbol{u} \in \mathbb{R}^k | \widetilde{f}(\boldsymbol{u}) \neq \infty\}$. To simplify notation, it is usual to simply refer to the extended-value extension as f.

Definition 4 (Indicator function)**.** Let $S \subseteq \mathbb{R}^k$ be a set. The indicator function $\delta_S : \mathbb{R}^k \to \infty$ of

[1] Note that for a nonconvex function, there may exists points where the function is differentiable but where no subgradient exists. Hence, the term generalization is an exaggeration in general.

the set S is defined as follows:

$$\delta_S(\boldsymbol{u}) = \begin{cases} 0 & \text{if } \boldsymbol{u} \in S, \\ \infty & \text{otherwise.} \end{cases}$$

Indicator functions can be used to introduce constraints in an optimization problem as a term in the objective. For example, the two following problems are equivalent:

$$\begin{aligned} \min_{\boldsymbol{u}} \ & f(\boldsymbol{u}) \\ \text{s.t. } & \boldsymbol{u} \in S. \end{aligned} \qquad = \qquad \min_{\boldsymbol{u}} \ f(\boldsymbol{u}) + \delta_S(\boldsymbol{u})$$

2 Optimality conditions

In this section, we present optimality conditions for optimization (un)constrained optimization problems. They will be useful to compute closed-form expression of minimizers of these optimization problems.

Proposition 4 (Fermat's optimality conditions)**.** Let $f : U \to \mathbb{R} \cup \{\infty\}$ be a proper convex function. Then, $\hat{\boldsymbol{u}} \in \arg\min_{\boldsymbol{u} \in U} f(\boldsymbol{u})$ is a minimizer of f is and only if $\mathbf{0} \in \partial f(\hat{\boldsymbol{u}})$.

Proof. For any $\boldsymbol{g} \in \partial f(\hat{\boldsymbol{u}})$, the subgradient inequality holds:

$$\forall \boldsymbol{v} \in U : \quad f(\boldsymbol{v}) \geq f(\hat{\boldsymbol{u}}) + \langle \boldsymbol{g}, \boldsymbol{v} - \boldsymbol{u} \rangle$$

In particular, we have $\mathbf{0} \in \partial f(\hat{\boldsymbol{u}})$ by assumption, therefore setting $\boldsymbol{g} = \mathbf{0}$ we obtain:

$$f(\boldsymbol{v}) \geq f(\hat{\boldsymbol{u}}),$$

hence $\hat{\boldsymbol{u}}$ is a minimizer of f. □

In the constrained case, we restrict ourselves to problems of the following form:

$$\begin{aligned} \min_{\boldsymbol{u} \in \mathbb{R}^k} \quad & f(\boldsymbol{u}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b} \\ & \boldsymbol{C}\boldsymbol{u} \leq \boldsymbol{d} \end{aligned}$$

where f is a convex function and $\boldsymbol{A} \in \mathbb{R}^{m \times k}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{C} \in \mathbb{R}^{n \times k}$ and $\boldsymbol{d} \in \mathbb{R}^m$ defines a set of m equality and n inequality constraints. We assume there exists at least one feasible solution to the problem. The Lagrangian $L : \mathbb{R}^k \times \mathbb{R}^m \times \mathbb{R}^n_{++} \to \mathbb{R}$ associated with this problem is defined as:

$$L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{u}) + \langle \lambda, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{u} - \boldsymbol{d} \rangle$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \in \mathbb{R}^n_{++}$ are dual variables associated with inequalities and equalities, respectively. Importantly, $\boldsymbol{\mu}$ must contains only non-negative values. By analogy, the vector $\boldsymbol{u}$ is called the primal variable. Although outside the scope of this course, for any couple of dual variables, the minimizing the Lagrangian over the primal variable defines an lower bound on the original problem:

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}^n : \quad \min_{\boldsymbol{u} \in \mathbb{R}^k} \quad L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \leq \quad \begin{aligned} \min_{\boldsymbol{u} \in \mathbb{R}^k} \quad & f(\boldsymbol{u}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b} \\ & \boldsymbol{C}\boldsymbol{u} \leq \boldsymbol{d} \end{aligned}$$

Proposition 5 (Karush–Kuhn–Tucker optimality conditions)**.** Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function and assume the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{u} \in \mathbb{R}^k} \quad & f(\boldsymbol{u}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b} \\ & \boldsymbol{C}\boldsymbol{u} \leq \boldsymbol{d} \end{aligned}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times k}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{C} \in \mathbb{R}^{n \times k}$ and $\boldsymbol{d} \in \mathbb{R}^m$. Then, optimal primal and dual variable $\hat{\boldsymbol{u}} \in \mathbb{R}^k$, $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^m$ and $\hat{\boldsymbol{\mu}} \in \mathbb{R}^k_{++}$ satisfy the following constraints:

$$\begin{aligned} &\text{(stationarity)} && \partial_{\boldsymbol{u}} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \ni \boldsymbol{0} \quad \Longleftrightarrow \quad \partial f(\boldsymbol{u}) + \boldsymbol{A}^\top \boldsymbol{\lambda} + \boldsymbol{C}^\top \boldsymbol{\mu} \ni \boldsymbol{0} \\ &\text{(primal feasibility)} && \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b} \\ &&& \boldsymbol{C}\boldsymbol{u} \leq \boldsymbol{d} \\ &\text{(dual feasibility)} && \boldsymbol{\mu} \geq \boldsymbol{0} \\ &\text{(complementary slackness)} && \langle \boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{u} - \boldsymbol{d} \rangle = \boldsymbol{0} \end{aligned}$$

Note that the KKT conditions are in general only necessary conditions, but they become sufficiant conditions in the case of convex problems as presented in Proposition 5

3 Fenchel conjugates

Fenchel conjugates are a very useful tools that appears in many area of machine learning (duality for convex learning problems like SVMs, loss functions, variational approximations, ...).

Definition 5 (Fenchel conjugate)**.** Let $f : U \to \mathbb{R} \cup \{\infty\}$ be a function. The Fenchel conjugate of f, denoted f^*, is defined as follows:

$$f^*(\boldsymbol{t}) = \sup_{\boldsymbol{u} \in \operatorname{dom} f} \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f(\boldsymbol{u}).$$

Proposition 6 (Fenchel-Young inequality)**.** Let $f : U \to \mathbb{R} \cup \{\infty\}$ be a function. Then, the following inequality holds:

$$\forall \boldsymbol{u} \in U, \boldsymbol{t} \in \operatorname{dom} f^* : \quad f^*(\boldsymbol{t}) \geq \langle \boldsymbol{u}, \boldsymbol{t} \rangle - f(\boldsymbol{u}).$$

Proof. The inequality directly follows from the definition of the Fenchel conjugate. □

Proposition 7 (Fenchel biconjugation)**.** Let $f : U \to \mathbb{R} \cup \{\infty\}$ be a closed convex function. Then:

$$f = f^{**},$$

that is conjugate of the conjugate of f is f.

4 Jensen's inequality

Proposition 8 (Jensen's inequality (simplified))**.** Let $f : U \to \mathbb{R} \cup \{\infty\}$ be a convex function, $\boldsymbol{u}^{(1)}...\boldsymbol{u}^{(n)}$ be n datapoints and $\boldsymbol{\mu} \in \triangle(n)$. Then:

$$\sum_{i=1}^{n} \mu_i f\left(\boldsymbol{u}^{(i)}\right) \geq f\left(\sum_{i=1}^{n} \mu_i \boldsymbol{u}^{(i)}\right).$$

Proof. Let $\boldsymbol{u} = \sum_{i=1}^{n} \mu_i \boldsymbol{u}^{(i)}$ and $\boldsymbol{g} \in \partial f(\boldsymbol{u})$. By convexity of f, we have:

$$\forall i \in \{1...n\} : \quad f(\boldsymbol{u}^{(i)}) \geq f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}^{(i)} - \boldsymbol{u} \rangle.$$

As $\boldsymbol{\mu}$ contains only positive values, by multipling and then summing all inequalities we obtain:

$$\sum_i \mu_i f\left(\boldsymbol{u}^{(i)}\right) \geq \sum_i \mu_i \left(f(\boldsymbol{u}) + \langle \boldsymbol{g}, \boldsymbol{u}^{(i)} - \boldsymbol{u} \rangle\right),$$

$$\sum_i \mu_i f\left(\boldsymbol{u}^{(i)}\right) \geq \sum_i \mu_i f(\boldsymbol{u}) + \left\langle \boldsymbol{g}, \sum_i \mu_i \left(\boldsymbol{u}^{(i)} - \boldsymbol{u}\right) \right\rangle,$$

Note that as $\boldsymbol{\mu} \in \triangle(n)$, we have $\sum_i \mu_i f(\boldsymbol{u}) = f(\boldsymbol{u}) \sum_i \mu_i = f(\boldsymbol{u})$. A similar argument cancels the second term of the right-hand side of the inequality. Replacing $\boldsymbol{\mu}$ by its definition, we obtain:

$$\sum_{i=1}^{n} \mu_i f\left(\boldsymbol{u}^{(i)}\right) \geq f\left(\sum_{i=1}^{n} \mu_i \boldsymbol{u}^{(i)}\right).$$

□

From the proposition above, we can see that it really looks like an expectation. This is indeed the case: the Jensen's inequality holds can be defined over expectation, including continuous random variables and transformed random variables. However, we don't prove this in these notes.

Proposition 9 (Jensen's inequality)**.** Let $f : X \to \mathbb{R} \cup \{\infty\}$ be a convex function, $\mathcal{X}$ a random variables taking values in X and p a probability distribution over $\mathcal{X}$. Then:

$$\mathbb{E}_{p(\mathcal{X})}\left[f(\mathcal{X})\right] \geq f\left(\mathbb{E}_{p(\mathcal{X})}\left[\mathcal{X}\right]\right).$$

Under mild conditionsm, the inequality also holds if $\mathcal{X}$ is transformed via a (possibly nonconvex) function g:

$$\mathbb{E}_{p(\mathcal{X})}\left[f(g(\mathcal{X}))\right] \geq f\left(\mathbb{E}_{p(\mathcal{X})}\left[g(\mathcal{X})\right]\right).$$

5 Exercises

In the following, σ refers to the sigmoid function and $\triangle(n)$ to the simplex of dimension n.

1. Let U be a convex set and n a strictly positive integer. Prove that:

$$\forall \boldsymbol{u}^{(1)}...\boldsymbol{u}^{(n)}, \boldsymbol{\mu} \in \triangle(n) : \quad \sum_{i=1}^{n} \mu_i \boldsymbol{u}^{(i)} \in U.$$

2. Prove that the following functions are convex:
   (a) $f : \mathbb{R}^k \to \mathbb{R}$ defined as $f(\boldsymbol{u}) = \langle \boldsymbol{a}, \boldsymbol{u} \rangle + b$ where $\boldsymbol{a} \in \mathbb{R}^k$ and $b \in \mathbb{R}$
   (b) $f : \mathbb{R} \to \mathbb{R}$ defined as $f(u) = \exp(u)$
   (c) $f : \mathbb{R}_{++} \to \mathbb{R}$ defined as $f(u) = -\log(u)$
   (d) $f : \mathbb{R}^k \to \mathbb{R}$ defined as $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{u}\|_2^2$
   (e) $f : \mathbb{R}^k \to \mathbb{R}$ defined as $f(\boldsymbol{u}) = \log \sum_i \exp(u_i)$
3. Compute (explicitly!) the conjugates and the biconjugates of the above functions.
4. Prove that the function $-\log \sigma(u)$ is convex.
5. Compute the Fenchel conjugate of function $-\log \sigma(u)$.
6. Use results of two previous questions to derive a parameterized upper bound to the sigmoid function.

References

[Beck, 2017] Beck, A. (2017). *First-order methods in optimization*. SIAM.

[Borwein and Lewis, 2006] Borwein, J. and Lewis, A. (2006). *Convex Analysis*. Springer.

[Boyd et al., 2004] Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.