Labs
Optimization for Machine Learning
Spring 2025

EPFL
School of Computer and Communication Sciences
Nicolas Flammarion
github.com/epfml/OptML_course

Problem Set 10 — Solutions (Lower Bounds & Convex conjugate)

1 Lower Bounds for a Non-smooth Function

1. The $\max$ function is convex, while $\|.\|^2$ is strongly convex, hence the function f is strongly convex. The function is not smooth, since the $\max$ function is not differentiable. The sub-differential of $\max$ at $\mathbf{x}$ is given by $\operatorname{conv}\left\{e_i \,:\, i \in \operatorname{argmax}_{1\le j\le t} \mathbf{x}_j\right\}$. This can be verified as follows:

$$\max_{1\le i\le t}(\mathbf{x}_i + \mathbf{h}_i) \ge \max_{1\le i\le t} \mathbf{x}_i + g^\top \mathbf{h},$$

   for any $g \in \partial f(\mathbf{x})$, $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$.

2. To show that

$$\operatorname{Span}\{\partial f(\mathbf{x}_0), \partial f(\mathbf{x}_1), \dots, \partial f(\mathbf{x}_{s-1})\} = \operatorname{Span}\{e_1, e_2, \dots, e_s\},$$

   we can use induction. For $s = 1$, we have $\mathbf{x}_0 = 0$ and $\partial f(\mathbf{x}_0) = \gamma e_1$ and the statement holds. Assume that the statement holds for $s = k$, i.e.,

$$\mathbf{x}_k \in \operatorname{Span}\{\partial f(\mathbf{x}_0), \partial f(\mathbf{x}_1), \dots, \partial f(\mathbf{x}_{k-1})\} = \operatorname{Span}\{e_1, e_2, \dots, e_k\}.$$

   For $s = k+1$, the sub-gradient at $f(\mathbf{x}_k)$ is given by

   - if $\operatorname{argmax}_{1\le i\le t}(\mathbf{x}_k)_i \in \{1, \dots, k\}$, then $\partial f(\mathbf{x}_k) = \alpha \mathbf{x}_k + \gamma e_j$ for some $j \in \{1, \dots, k\}$ and we have that $\partial f(\mathbf{x}_k) \in \operatorname{Span}\{e_1, e_2, \dots, e_k\}$ as $\mathbf{x}_k \in \operatorname{Span}\{e_1, e_2, \dots, e_k\}$ from the induction hypothesis.
   - If $\operatorname{argmax}_{1\le i\le t}(\mathbf{x}_k)_i \notin \{1, \dots, k\}$, then $\partial \max$ is e_{k+1} (as $(\mathbf{x}_k)_i = 0$ for $i \ge k+1$ and we choose the smallest index). Hence $\partial f(\mathbf{x}_k) = \alpha \mathbf{x}_k + \gamma e_{k+1}$. We have that $\partial f(\mathbf{x}_k) \in \operatorname{Span}\{e_1, e_2, \dots, e_{k+1}\}$ as $\mathbf{x}_k \in \operatorname{Span}\{e_1, e_2, \dots, e_k\}$ from the induction hypothesis.

3. Note that $\|\partial f(\mathbf{x})\| \le \alpha\|\mathbf{x}\| + \gamma$, hence we have that for any $\mathbf{x} \in B(R)$, f is Lipschitz with $L = \alpha R + \gamma$.

4. Note that the sub-differential of f at y is given by

$$\partial f(y) = \alpha y + \gamma \operatorname{conv}\left\{e_i \,:\, i \in \operatorname*{argmax}_{1\le j\le t} y_j\right\}. \qquad = \alpha y + \gamma \operatorname{conv}\{e_i \,:\, i = 1, \dots, t\}.$$

   Now, that $(\frac{1}{t}, \dots, \frac{1}{t}, 0, \dots, 0) \in \operatorname{conv}\{e_i \,:\, i = 1, \dots, t\}$, we have that $0 = \alpha y_i + \gamma e_i$ for $i = 1, \dots, t$. Hence $0 \in \partial f(y)$.

5. As $0 \in \partial f(y)$, so $\min f(x) = f(y) = -\frac{\gamma}{\alpha t} + \frac{\alpha}{2}\|y\|^2 = -\frac{\gamma^2}{2\alpha t}$ as $\|y\|^2 = \sum_{i=1}^t \frac{\gamma^2}{\alpha^2 t^2} = \frac{\gamma^2}{\alpha^2 t}$. Coming to the algorithm, for $s < t$, we have that

$$\mathbf{x}_s \in \operatorname{Span}\{e_1, \dots e_s\}.$$

   Hence $f(\mathbf{x}_s) = \gamma \max_{1\le j\le t}(\mathbf{x}_s)_i + \frac{\alpha}{2}\|\mathbf{x}_s\|^2$. First $\|\mathbf{x}_s\|^2 > 0$ and $\max_{1\le j\le t}(\mathbf{x}_s)_i \ge (\mathbf{x}_s)_t = 0$ since $\mathbf{x}_s \in \operatorname{Span}\{e_1, \dots e_s\}$ and $s < t$. Hence $f(\mathbf{x}_s) \ge 0$

$$f(\mathbf{x}_s) - \min_{\|x\|\le R} f(x) \ge 0 - f(y) = \frac{\gamma^2}{2\alpha t}.$$

Choose

$$\alpha = \frac{L}{R}\frac{1}{\sqrt{t}+1}, \quad \gamma = \frac{L\sqrt{t}}{\sqrt{t}+1},$$

then f is Lipschitz with L as $\alpha R + \gamma = L$. Hence, we have,

$$f(\mathbf{x}_s) - \min_{\|x\|\le R} f(x) \ge 0 - f(y) = \frac{\gamma^2}{2\alpha t} = \frac{RL}{2(\sqrt{t}+1)}.$$

2 Convex conjugate

For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ (which is not necessarily convex !), we consider its **convex conjugate** which for $y \in \mathbb{R}^d$ is defined as

$$f^*(y) = \sup_{x\in\mathbb{R}^d} (\langle x, y\rangle - f(x)) \in \mathbb{R} \cup \{+\infty\}$$

Prove the following properties.

1. Show that f^* is convex.

   Proof: Note that f^* is the pointwise supremum of **affine functions** $y \mapsto \langle x, y\rangle - f(x)$. As seen in the first class, the pointwise supremum of convex functions is convex. Therefore f^* is convex.

2. Show that for $x, y \in \mathbb{R}^d$, $f(x) + f^*(y) \ge \langle x, y\rangle$. This is known as the Fenchel inequality.

   Proof: For $y \in \mathbb{R}^d$, $f^*(y) = \sup_{x\in\mathbb{R}^d}(\langle x, y\rangle - f(x)) \ge \langle x, y\rangle - f(x)$ for all $x \in \mathbb{R}^d$.

3. Show that the biconjugate f^{**} (the conjugate of the conjugate) is such that $f^{**} \le f$.

   Proof: From the previous inequality we have that for all $x, y \in \mathbb{R}^d$, $f(x) \ge \langle x, y\rangle - f^*(y)$, we can therefore take the supremum over y of the left hand side: $f(x) \ge \sup_{y\in\mathbb{R}^d}(\langle y, x\rangle - f^*(y)) = f^{**}(x)$

The Fenchel-Moreau theorem (which we will not prove here) states that $f = f^{**}$ if and only if f is convex and closed. It will turn out to be useful to show the following property.

4. Assume that f is closed and convex. Then show that for any $x, y \in \mathbb{R}^d$,

   $$\begin{aligned} y \in \partial f(x) &\Leftrightarrow x \in \partial f^*(y) \\ &\Leftrightarrow f(x) + f^*(y) = \langle x, y\rangle \end{aligned}$$

   Proof that $y \in \partial f(x) \Rightarrow f(x) + f^*(y) = \langle x, y\rangle$: Assume that $y \in \partial f(x)$, then we have that for all $z \in \mathbb{R}^d$, $f(z) \ge f(x) + \langle y, z - x\rangle$. Therefore for all $z \in \mathbb{R}^d$, $\langle y, x\rangle - f(x) \ge \langle z, y\rangle - f(z)$. We can therefore take the supremum of the left hand size which gives that $\langle y, x\rangle - f(x) \ge \sup_z(\langle z, y\rangle - f(z))$ which also means that $\langle y, x\rangle - f(x) = \sup_z\langle z, y\rangle - f(z) = f^*(y)$ which proves the first part of the result.

   Proof that $f(x) + f^*(y) = \langle x, y\rangle \Rightarrow y \in \partial f(x)$: We basically do the previous reasoning the other way round. Let $x, y \in \mathbb{R}^d$ such that $f(x) + f^*(y) = \langle x, y\rangle$. Therefore $\langle x, y\rangle - f(x) = f^*(y) = \sup_z(\langle z, y\rangle - f(z)) \ge \langle z, y\rangle - f(z)$ for all $z \in \mathbb{R}^d$. Rearranging we get that for all $z \in \mathbb{R}^d$, $f(z) \ge f(x) + \langle y, z - x\rangle$ which means that $y \in \partial f(x)$.

   Hence we have shown that $y \in \partial f(x) \Leftrightarrow f(x) + f^*(y) = \langle x, y\rangle$. Now we can apply this same result to f^*: $x \in \partial f^*(y) \Leftrightarrow f^*(y) + f^{**}(x) = \langle y, x\rangle$. Since f is closed and convex, by the Fenchel-Moreau theorem we have that $f = f^{**}$, hence $x \in \partial f^*(y) \Leftrightarrow f^*(y) + f(x) = \langle y, x\rangle$. Therefore all the implications are proven.