[SD-TSIA 211] Optimization for Machine Learning

“TD-02 Corrections”

Olivier FERCOQ — Iyad WALWIL

December 2, 2022

Recalls

Definition 1. *(Strong Convexity)*
A function f is μ-strongly convex if $f - \frac{\mu}{2}\|.\|^2$ is convex.

Proposition. ***2.1.4***
A function $f\colon \mathcal{X} \to [-\infty, +\infty]$ is μ-strongly convex if, and only if
$\forall (x, y) \in \mathcal{X}^2, \forall t \in (0, 1)$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\mu}{2}t(1-t)\|x - y\|^2 \tag{1}$$

Proposition. ***2.2.2***
Let f be a l.s.c. function such that $\lim_{\|x\|\to+\infty} f(x) = +\infty$, then there exists x^* such that $f(x^*) = \inf_{x\in\mathcal{X}} f(x)$.

Definition 2. *(Sub-differential)*
*Let $f\colon \mathcal{X} \to [-\infty, +\infty]$ and $x \in \operatorname{dom} f$. A vector $\phi \in \mathcal{X}$ is called a **sub-gradient** of f at x if:*

$$\forall y \in \mathcal{X}, \quad f(y) - f(x) \geq \langle \phi, y - x\rangle \tag{2}$$

Lemma 1. *Let $g\colon \mathcal{X} \to (-\infty, +\infty]$ be a convex function, $f\colon \mathcal{Y} \to \mathbb{R}$ a convex differentiable function, and $M\colon \mathcal{X} \to \mathcal{Y}$ a linear operator.*

$$\forall x \in \mathcal{X}, \quad \partial(g + f \circ M)(x) = \partial g(x) + \{M^*\nabla g(Mx)\} \tag{3}$$

Theorem 1. *(Fermat’s rule)*

$$x \in \arg\min f \iff 0 \in \partial f(x) \tag{4}$$

Exercise 1. *(Proximal operator).*

Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex lower-semicontinuous function such that $\operatorname{dom} f \neq \emptyset$

1. Recall the definition of the domain of a convex function.
2. It is possible to prove (but we do not ask you to do it) that $\exists x_0 \in \operatorname{dom} f$ such that $\exists q_0 \in \partial f(x_0)$. Using this information, show that there exists $\alpha \in \mathbb{R}$ and $w \in \mathbb{R}^n$ such that for all x, $f(x) \geq \alpha + \langle w, x\rangle$.
3. Let us fix $x \in \mathbb{R}^n$. Let us define $g\colon y \mapsto f(y) + \frac{1}{2}\|x-y\|^2$. Show that g is strongly convex.
4. Show that $\lim_{\|y\| \to +\infty} g(y) = +\infty$
5. Show that g has a minimizer and that it is unique.

We will denote this minimizer as $\operatorname{Prox}_f(x)$. The function $\operatorname{Prox}_f\colon \mathbb{R}^n \to \mathbb{R}^n$ is called the proximal operator of f.

Answers

1. $\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$ (Lecture notes, page 8).
2. Assume $\exists x_0$ s.t. $\exists q_0 \in \partial f(x_0)$, then by Definition 2 (Sub-differential),

$$\begin{aligned} q_0 \in \partial f(x_0) &\iff \forall x \in \mathbb{R}^n, \quad f(x) \geq f(x_0) + \langle q_0, x - x_0\rangle \\ &\iff \forall x \in \mathbb{R}^n, \quad f(x) \geq f(x_0) - \langle q_0, x_0\rangle + \langle q_0, x\rangle \\ &\iff \forall x \in \mathbb{R}^n, \quad f(x) \geq \alpha + \langle w, x\rangle \end{aligned} \tag{5}$$

by taking $\alpha = f(x_0) - \langle q_0, x_0\rangle$ and $w = q_0$ $\square$

3. Fix $x \in \mathbb{R}^n$, and define $g(y) = f(y) + \frac{1}{2}\|x - y\|^2$

   To show that $g(y)$ is strongly convex, define $h(y) = g(y) - \frac{1}{2}\|y\|^2$

   If $h(y)$ is convex, then by Definition 1 $g(y)$ is strongly convex.

   Now, let's prove that $h(y)$ is, indeed, convex:

$$\begin{aligned} h(y) &= g(y) - \frac{1}{2}\|y\|^2 \\ &= f(y) + \frac{1}{2}\|x - y\|^2 - \frac{1}{2}\|y\|^2 \\ &= f(y) + \frac{1}{2}\|x\|^2 - \langle x, y\rangle + \frac{1}{2}\|y\|^2 - \frac{1}{2}\|y\|^2 \\ &= f(y) + \frac{1}{2}\|x\|^2 - \langle x, y\rangle \\ &= f(y) + p(y) \end{aligned}$$

   Thus, $h(y)$ is convex as it's a sum of two convex function (Lecture notes - Exercise 2.1.2. (2)). $f(y)$ is convex by assumption, and $p(y) = \frac{1}{2}\|x\|^2 - \langle x, y\rangle$ is convex since it's affine. Hence, $g(y)$ is strongly convex. □

4. $\lim_{\|y\| \to +\infty} g(y) = +\infty$ means that g is coercive.

$$\begin{aligned} g(y) &= f(y) + \frac{1}{2}\|x - y\|^2 \\ &\overset{(5)}{\geq} \alpha + \langle w, y\rangle + \frac{1}{2}\|x - y\|^2 \\ &= \alpha + \langle w - x, y\rangle + \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 \\ &\geq \alpha + \frac{1}{2}\|x\|^2 - \|w - x\|\|y\| + \frac{1}{2}\|y\|^2 \quad \text{(by Cauchy-Schwarz inequality)} \\ &= \|y\|\Big[\frac{1}{2}\|y\| - \|w - x\|\Big] + \frac{1}{2}\|x\|^2 + \alpha \end{aligned}$$

Thus, $\lim_{\|y\|\to+\infty} g(y) \geq \lim_{\|y\|\to+\infty} \|y\|\left[\frac{1}{2}\|y\| - \|w-x\|\right] + \frac{1}{2}\|x\|^2 + \alpha = +\infty$. Hence,

$$\lim_{\|y\|\to+\infty} g(y) = +\infty$$

$\square$

5. Show that g has a unique minimizer.

- g is l.s.c. (as a sum of two l.s.c. functions) and coercive, then by proposition **2.2.2** g has a minimizer.
- The uniqueness is implied by the strong convexity of g.

  Suppose that x_1 and x_2 are two minimizers of g, then by proposition **2.1.4**, one has:

$$\begin{aligned} g\Big(\frac{1}{2}(x_1+x_2)\Big) &\leq \frac{1}{2}g(x_1) + \frac{1}{2}g(x_2) - \frac{1}{8}\|x_1-x_2\|^2 \\ &= \frac{1}{2}\min g + \frac{1}{2}\min g - \frac{1}{8}\|x_1-x_2\|^2 \\ &= \min g - \frac{1}{8}\|x_1-x_2\|^2 \end{aligned}$$

  Thus,

$$0 \leq g\Big(\frac{1}{2}(x_1+x_2)\Big) - \min g \leq -\frac{1}{8}\|x_1-x_2\|^2 \leq 0 \Rightarrow \|x_1-x_2\|^2 = 0 \Rightarrow x_1 = x_2$$

$\square$

Exercise 2.

Let us denote

$$\operatorname{Prox}_{\gamma g}(y) = \arg\min_{x \in \mathcal{X}} g(x) + \frac{1}{2\gamma}\|x - y\|^2 \tag{6}$$

the proximal operator of g at y.

Fix $\gamma > 0$. Show that if f and g are convex, then the fixed points of the non-linear equation

$$x = \operatorname{Prox}_{\gamma g}(x - \gamma \nabla f(x))$$

are the minimizers of the function $F = f + g$.

Proof:

$$\begin{aligned}
x &= \operatorname{Prox}_{\gamma g}(x - \gamma \nabla f(x)) \\
&\overset{(6)}{=} \arg\min_{u \in \mathcal{X}} g(u) + \frac{1}{2\gamma}\|u - (x - \gamma \nabla f(x))\|^2 \\
&\overset{(4)}{\iff} 0 \in \partial\Big(g + \frac{1}{2\gamma}\|. - (x - \gamma \nabla f(x))\|^2\Big)(x) \\
&\overset{(3)}{\iff} 0 \in \partial g(x) + \frac{1}{\gamma}(x - x + \gamma \nabla f(x)) \\
&\iff 0 \in \partial g(x) + \nabla f(x) \\
&\iff 0 \in \partial F(x)
\end{aligned}$$

Thus, x is a minimizer of $F = f + g$. $\square$

Exercise 3. *(Taylor-Lagrange inequality).*

The goal of this exercise is to prove Taylor- Lagrange inequality. This is a fundamental inequality for the study of gradient descent and related methods. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, *i.e.*

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \quad \|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\| \tag{7}$$

1. Prove that, $\forall (x, y) \in \mathbb{R}^{2n}, \quad \langle \nabla f(y) - \nabla f(x), y - x\rangle \leq L\|y - x\|^2$
2. Set $\varphi(t) = f(x + t(y - x))$ for all $t \in [0, 1]$. Prove that
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \varphi(1) - \varphi(0) - \varphi'(0)$$
3. Deduce that
$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle dt$$
4. Using the 1^{st} question, conclude that
$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$$

Answers

1. Prove that $\forall (x, y) \in \mathbb{R}^{2n}, \quad \langle \nabla f(y) - \nabla f(x), y - x\rangle \leq L\|y - x\|^2$

   Proof: for any $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$

$$\begin{aligned}\langle \nabla f(y) - \nabla f(x), y - x\rangle &\leq \|\nabla f(y) - \nabla f(x)\|\|y - x\| && \text{(Cauchy–Schwarz inequality)}\\ &\overset{(7)}{\leq} L\|y - x\|^2 && (\nabla f \text{ is Lipschitz}) \quad \square\end{aligned}$$

2. Set $\varphi(t) = f(x + t(y - x))$ for all $t \in [0, 1]$. Prove that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \varphi(1) - \varphi(0) - \varphi'(0)$$

Proof:

$$\varphi(t) = f(x + t(y - x)) \Rightarrow \varphi'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle \tag{8}$$

Thus, $\varphi(0) = f(x), \varphi(1) = f(y)$, and $\varphi'(0) = \langle \nabla f(x), y - x \rangle$.

Putting everything together, we get:

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \varphi(1) - \varphi(0) - \varphi'(0)$$

$\square$

3. Deduce that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt$$

Proof: As φ is a primitive (anti-derivative) of φ', then

$$\int_0^1 \varphi'(t)dt = \varphi(1) - \varphi(0) \tag{9}$$

Now, from (2), we know that:

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \varphi(1) - \varphi(0) - \varphi'(0) \\ &\overset{(9)}{=} \int_0^1 \varphi'(t)dt - \varphi'(0) \\ &\overset{(8)}{=} \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt - \langle \nabla f(x), y - x \rangle \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \end{aligned}$$

$\square$

4. Using the 1^{st} question, conclude that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Proof:

- From (1), we know that:

$$\forall (z, x) \in \mathbb{R}^n \times \mathbb{R}^n, \quad \langle \nabla f(z) - \nabla f(x), z - x \rangle \leq L\|z - x\|^2$$

  So, for $z = x + t(y - x)$, we get:

$$\begin{aligned} &\langle \nabla f(x + t(y - x)) - \nabla f(x), t(y - x) \rangle \leq L\|t(y - x)\|^2 \\ \Rightarrow &\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \leq Lt\|(y - x)\|^2 \end{aligned} \tag{10}$$

From (3), we know that (equivalently):

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &\overset{(10)}{\leq} f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 Lt\|y - x\|^2 dt \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 t^2 \Big|_0^1 \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \end{aligned}$$

$\square$