Labs
Optimization for Machine Learning
Spring 2023

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

Problem Set 3 — Solutions (Projected Gradient Descent)

Gradient Descent

Exercise 1. *23 Consider the projected gradient descent algorithm as in (3.1) and (3.2), with a convex differentiable function f. Suppose that for some iteration t, $\mathbf{x}_{t+1} = \mathbf{x}_t$. Prove that in this case, $\mathbf{x}_t$ is a minimizer of f over the closed and convex set X!*

Solution: By Fact 3.1 (i) with $\mathbf{y} = \mathbf{y}_{t+1}$, and using $\mathbf{x}_{t+1} = \mathbf{x}_t$, we have

$$(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{y}_{t+1} - \mathbf{x}_t) \leq 0$$

for all $\mathbf{x} \in X$. On the other hand, by definition of projected gradient descent,

$$\mathbf{y}_{t+1} - \mathbf{x}_t = -\gamma\nabla f(\mathbf{x}_t), \quad \gamma > 0.$$

Substituting this equation into the former inequality yields

$$-\gamma(\mathbf{x} - \mathbf{x}_t)^\top\nabla f(\mathbf{x}_t) \leq 0, \quad \mathbf{x} \in X.$$

Multiplying by -1 and dividing by γ gives

$$(\mathbf{x} - \mathbf{x}_t)^\top\nabla f(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) \geq 0, \quad \mathbf{x} \in X.$$

By Lemma 1.28, this precisely says that $\mathbf{x}_t$ minimizes f over X.

Exercise 2. *24 Prove that in Theorem 3.4 (i),*

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t).$$

Solution: By definition of projected gradient descent we have

$$\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\| \leq \|\mathbf{y}_{t+1} - \mathbf{x}_t\| = \gamma\|\nabla f(\mathbf{x}_t)\|.$$

The inequality holds because of (3.1) (by definition, $\mathbf{x}_{t+1}$ is the point closest to $\mathbf{y}_{t+1}$ in X). The equality holds because of (3.2) (by definition, $\mathbf{y}_{t+1} = \mathbf{x}_t - \gamma\nabla f(\mathbf{x}_t)$). Combining the above inequality with the step size $\gamma = 1/L$ and squaring yields

$$\|\nabla f(\mathbf{x}_t)\|^2 \geq L^2\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.$$

The desired inequality now easily follows from Lemma 3.3.

Exercise 3. *26 Prove Lemma 3.12!*

Hint: *It is useful to prove that with $\mathbf{x}^\star(p)$ as in (3.12) and satisfying (3.13),*

$$\mathbf{x}^\star(p) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^{d} x_i = 1, x_{p+1} = \cdots = x_d = 0\}.$$

Solution: We claim that

$$\mathbf{x}^\star(p) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^{d} x_i = 1, x_{p+1} = \cdots = x_d = 0\}.$$

Assume for the moment that this claim is true. By Lemmas 3.10 and 3.11 we know that there exists $1 \le p \le d$ such that $\Pi_X(\mathbf{v}) = \mathbf{x}^\star(p)$. Which means that $\mathbf{x}^\star(p) = \operatorname{argmin}_{\mathbf{x}\in X} \|\mathbf{x} - \mathbf{v}\|^2$. Now suppose Lemma 3.12 is wrong, which means that we can find $p' > p$, $(p' \ge p+1)$ with $\mathbf{x}^\star(p')$ as in (3.12) and satisfying (3.13), which means that we also get

$$\mathbf{x}^\star(p') = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^d x_i = 1, x_{p'+1} = \cdots = x_d = 0\}.$$

Here we are minimizing $\|\mathbf{x} - \mathbf{v}\|$ with less constraint than in the previous case with $\mathbf{x}^\star(p)$ (components $p+1$ to p' do not have to be equal to 0), which implies that $\|\mathbf{x}^\star(p') - \mathbf{v}\| \le \|\mathbf{x}^\star(p) - \mathbf{v}\|$. Combining this with the previous assumption of $\mathbf{x}^\star(p) = \Pi_X(\mathbf{v})$ we get $\|\mathbf{x}^\star(p') - \mathbf{v}\| = \|\mathbf{x}^\star(p) - \mathbf{v}\|$. And since we are projecting on a convex set we know that the projection is unique, and thus $\mathbf{x}^\star(p') = \mathbf{x}^\star(p)$. However, from the way $\mathbf{x}^\star(p)$ and $\mathbf{x}^\star(p')$ are defined using (3.12), we know that the $p+1$ component of $\mathbf{x}^\star(p)$ is equal to 0, and that of $\mathbf{x}^\star(p')$ is strictly positive which leads to a contradiction.

It remains only to prove our claim. That is, to show that for a given $1 \le p \le d$ indeed

$$\mathbf{x}^\star(p) = \operatorname{argmin}\{\|\mathbf{x} - \mathbf{v}\| : \sum_{i=1}^d x_i = 1, x_{p+1} = \cdots = x_d = 0\},$$

provided that $\mathbf{x}^\star(p)$ satisfies conditions (3.12) and (3.13).

Let $Y = \{\mathbf{x} \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_{p+1} = \cdots = x_d = 0\}$, and let $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(x) = \|\mathbf{v} - \mathbf{x}\|^2$. To prove our claim, it suffices to show that $\mathbf{x}^\star(p) \in Y$ is a minimizer of f over Y. By the optimality condition of Lemma 1.28, it suffices to show that $\nabla f(\mathbf{x}^\star(p))^\top(\mathbf{x} - \mathbf{x}^\star(p)) \ge 0$ for all $\mathbf{x} \in Y$. Because $\nabla f(\mathbf{x}) = 2(\mathbf{v} - \mathbf{x})$, we want to show that

$$-2(\mathbf{v} - \mathbf{x}^\star(p))^\top(\mathbf{x} - \mathbf{x}^\star(p)) \ge 0. \tag{1}$$

Notice that the first p coordinates of $(\mathbf{v} - \mathbf{x}^\star(p))$ are all equal to Θ_p. Moreover, the last $(d-p)$ coordinates of both $\mathbf{x} \in Y$ and $\mathbf{x}^\star(p)$ are all equal to 0. Therefore, we get that $(\mathbf{v} - \mathbf{x}^\star(p))^\top(\mathbf{x} - \mathbf{x}^\star(p))$ equals

$$(\Theta_p, \ldots, \Theta_p, v_{p+1}, \ldots, v_d)^\top(x_1 - v_1 + \Theta_p, \ldots, x_p - v_p + \Theta_p, 0, \ldots, 0)$$

Expanding this product, we get

$$(\mathbf{v} - \mathbf{x}^\star(p))^\top(\mathbf{x} - \mathbf{x}^\star(p)) = \Theta_p \sum_{i=1}^p (x_i - v_i + \Theta_p) = \Theta_p \left(\sum_{i=1}^p x_i - \sum_{i=1}^p v_i + p\Theta_p \right).$$

Because $\mathbf{x} \in Y$, we know that $\sum_{i=1}^p x_i = 1$, and since $\Theta_p = \frac{1}{p}(\sum_{i=1}^p v_i - 1)$, we get that

$$(\mathbf{v} - \mathbf{x}^\star(p))^\top(\mathbf{x} - \mathbf{x}^\star(p)) = \Theta_p \left(1 - \sum_{i=1}^p v_i + p\frac{1}{p} \left(\sum_{i=1}^p v_i - 1 \right) \right) = 0.$$

That is, equation (1) holds, and by Lemma 1.28 we conclude that $\mathbf{x}^\star(p)$ is a minimizer of f over Y proving our claim.

Exercise 27. *Prove Theorem 3.14!*

Solution: From (3.17), the proximal step could be written as

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d}\{g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 + h(\mathbf{y})\} = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^d}\{\psi(\mathbf{y})\},$$

where the function $\psi(\mathbf{y}) = g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 + h(\mathbf{y})$ is strongly convex with the parameter L. This means that $\psi(\mathbf{y}) \ge \psi(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_{t+1}\|^2$. This is equivalent to

$$\nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 + h(\mathbf{y}) \ge \nabla g(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + h(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_{t+1}\|^2,$$

Rearranging terms and subtracting $h(\mathbf{x}_t)$ from both sides,

$$\nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 - \frac{L}{2}\|\mathbf{y} - \mathbf{x}_{t+1}\|^2 + h(\mathbf{y}) - h(\mathbf{x}_t) \ge \nabla g(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + h(\mathbf{x}_{t+1}) - h(\mathbf{x}_t)$$

As the function g is L-smooth, we can estimate the right side as $g(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \geq g(\mathbf{x}_{t+1}) - g(\mathbf{x}_t)$, and because g is convex, on the left side we estimate $\nabla g(\mathbf{x}_t)^\top(\mathbf{y} - \mathbf{x}_t) \leq g(\mathbf{y}) - g(\mathbf{x}_t)$. Putting this together

$$f(\mathbf{y}) - f(\mathbf{x}_t) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 - \frac{L}{2}\|\mathbf{y} - \mathbf{x}_{t+1}\|^2 \geq f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)$$

This holds for any $\mathbf{y} \in \mathbb{R}^d$. Lets take $\mathbf{y} = \mathbf{x}^*$ and sum up the inequation above from $t = 0$ to $t = T - 1$

$$\sum_{t=0}^{T-1} (f(\mathbf{x}^*) - f(\mathbf{x}_t)) + \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}_0\|^2 - \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}_T\|^2 \geq f(\mathbf{x}_T) - f(\mathbf{x}_0)$$

or equivalently,

$$\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}_0\|^2 - \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}_T\|^2 \leq \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}_0\|^2$$

Because $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ for each $0 \leq t \leq T$

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{L}{2T}\|\mathbf{x}^* - \mathbf{x}_0\|^2.$$

Computing Fixed Points

Gradient descent turns up in a surprising number of situations which apriori have nothing to do with optimization. In this exercise we will see how computing the fixed point of functions can be seen as a form of gradient descent. Suppose that we have a 1-Lipschitz continuous function $g : \mathbb{R} \to \mathbb{R}$ such that we want to solve for

$$g(x) = x\,.$$

A simple strategy for finding such a fixed point is to run the following algorithm: starting from an arbitary x_0, we iteratively set

$$x_{t+1} = g(x_t)\,. \tag{2}$$

Practical exercise. We will try solve for x starting from $x_0 = 1$ in the following two equations:

$$x = \log(1 + x)\,, \text{ and} \tag{3}$$
$$x = \log(2 + x)\,. \tag{4}$$

Follow the Python notebook provided here:

github.com/epfml/OptML_course/tree/master/labs/ex03/

What difference do you observe in the rate of convergence between the two problems? Let's understand why this occurs.

Theoretical questions.

1. We want to re-write the update (2) as a step of gradient descent. To do this, we need to find a function f such that the gradient descent update is identical to (2):

   $$x_{t+1} = x_t - \gamma f'(x_t) = g(x_t)\,.$$

   Derive such a function f.

   Solution: We need $\gamma f'(x) = x - g(x)$. Thus upto additional linear terms, f is

   $$f = \frac{1}{2\gamma}x^2 - \frac{1}{\gamma}\int g(x)dx\,.$$

2. Give sufficient conditions on g to ensure convergence of procedure (2). What γ would you need to pick? *Hint: We know that gradient descent on f with fixed step-size converges if f is convex and smooth. What does this mean in terms of g?*

   Solution: If f is convex and $1/\gamma$-smooth, Theorem 2.1 guarantees convergence of (2). For this we need to show that $f'' \geq 0$ and $f'' \leq \frac{1}{\gamma}$.

   Firstly, we assume that g is differentiable in order for f'' to exist.

   We will use the relation derived in the previous question

$$\begin{aligned}(f'(x))' &= \frac{1}{\gamma}(x - g(x))' \\ &= \frac{1}{\gamma}(1 - g'(x))\,.\end{aligned}$$

   For $f'' \in [0, \frac{1}{\gamma}]$, we need

$$g'(x) \in [0, 1]\,.$$

   The condition $g'(x) \leq 1$ is already satisfied for any $\gamma > 0$ if $g(x)$ is 1-Lipschitz continuous. Hence, we only additionally require $g'(x) \geq 0$, i.e. g is non-decreasing.

3. What condition does g need to satisfy to ensure *linear* convergence? Are these satisfied for problems (3) and (4) in the exercise?

   Solution: To get linear convergence, we need that there exists a constant $\mu > 0$ such that $f''(x) \geq \mu$. In terms of g, this translates to the existence of $\mu > 0$ such that

$$f''(x) = \frac{1}{\gamma}(1 - g'(x)) \geq \mu \Rightarrow g'(x) \leq (1 - \gamma\mu) < 1\,.$$

   Thus we only need that $g'(x) < 1$.

   For $g(x) = \log(1 + x)$, $g'(x) = \frac{1}{1+x}$. Over the domain $[0, 2]$ which we consider, $g'(x) \in [0, 1]$ and so our procedure converges. However for $x = 0$, $g'(0) = 1$ and so we will not get linear convergence. This explains why (2) was slow.

   For $g(x) = \log(2 + x)$, $g'(x) = \frac{1}{2+x}$. Over the domain $[0, 2]$ which we consider, $g'(x) \in [0, 0.5]$. This shows that not only does (2) converge, but it converges at a linear rate!