Labs
Optimization for Machine Learning
Spring 2020

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

Problem Set 3 — Solutions (Gradient Descent, cont.)

Gradient Descent

Exercise 12. *Consider the function $f(x) = |x|^{3/2}$ for $x \in \mathbb{R}$.*

(i) Prove that f is strictly convex and differentiable, with a unique global minimum $x^\star = 0$.

(ii) Prove that for every fixed stepsize γ in gradient descent (2.11) applied to f, there exists x_0 for which $f(x_1) > f(x_0)$.

(iii) Prove that f is not smooth.

(iv) Let $X \subseteq \mathbb{R}$ be a closed convex set such that $0 \in X$ and $X \neq \{0\}$. Prove that f is not smooth over X.

Solution:

(i) Since for all $x > 0$, $f(x) = x^{3/2}$, and for all $x < 0$, $f(x) = (-x)^{3/2}$, f is (infinitely) differentiable at every point $x \neq 0$. So we need to show that f is differentiable at the point $x = 0$. Indeed, by definition of derivative

$$f'(0) = \lim_{h\to 0} \frac{f(h) - f(0)}{h} = \lim_{h\to 0} \frac{|h|^{3/2}}{h} = \lim_{h\to 0} \operatorname{sign}(h)|h|^{1/2} = 0.$$

To prove that f is strictly convex, we will at first show that the function $x^{3/2}$ (with domain $x > 0$) is strictly convex. Its second derivative $\frac{3}{4}x^{-1/2}$ is positive for all $x > 0$. If some function has positive second derivative at every point of its domain and the domain is open and convex, then this function is strictly convex (see the discussion after definition 1.18). Hence $x^{3/2}$ is strictly convex.

We need to show that

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

holds for all $x, y \in \mathbb{R}$ such that $x \neq y$ and for all $\lambda \in (0, 1)$.

At first assume that both x and y are nonzero. Then $|x| > 0$, $|y| > 0$ and we get

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= |\lambda x + (1-\lambda)y|^{3/2} \\ &\leq (|\lambda x| + |(1-\lambda)y|)^{3/2} \quad \text{(triangle inequality)} \\ &= (\lambda|x| + (1-\lambda)|y|)^{3/2} \\ &< \lambda|x|^{3/2} + (1-\lambda)|y|^{3/2} \quad \text{(strict convexity of } x^{3/2}\text{)} \\ &= \lambda f(x) + (1-\lambda)f(y). \end{aligned}$$

It remains to show that (1) holds when $x = 0$ or $y = 0$. Without loss of generality, assume that $y = 0$. Then for all $x \neq 0$ and $\lambda \in (0, 1)$

$$f(\lambda x + (1-\lambda)y) = \lambda^{3/2}|x|^{3/2} < \lambda|x|^{3/2} = \lambda f(x) + (1-\lambda)f(y).$$

Since f is strictly convex and nonnegative, it has a unique global minimum $x^\star = 0$.

(ii) We need to find x_0 such that

$$|x_1|^{3/2} = |x_0 - \gamma f'(x_0)|^{3/2} > |x_0|^{3/2}.$$

We may assume that $x_0 > 0$. Then $f'(x_0) = \frac{3}{2}x_0^{1/2}$. We get

$$|x_1|^{3/2} = |x_0 - \frac{3}{2}\gamma x_0^{1/2}|^{3/2} = |x_0|^{3/2}|\frac{3}{2}\gamma x_0^{-1/2} - 1|^{3/2}.$$

If $0 < x_0 < \frac{9}{4}\gamma^2$, then $|x_1|^{3/2} > |x_0|^{3/2}$.

(iii) Suppose that f is smooth. Then by Lemma 2.6 there exists a stepsize in gradient descent (2.11) applied to f such that for all points x_0, $f(x_1) \leq f(x_0)$, which is a contradiction to point (ii).

(iv) Suppose that f is smooth with some parameter L. Since $X \neq \{0\}$, it contains some point $a \neq 0$. Then by convexity of X, the closed interval with endpoints a and 0 is a subset of X. Take $y \neq 0$ from this interval such that $|y| < \frac{4}{L^2}$ and $x = 0$. By definition of smoothness

$$f(y) = |y|^{3/2} \leq f(x) + f'(x)(y - x) + \frac{L}{2}|x - y|^2 = \frac{L}{2}|y|^2.$$

We get $|y|^{1/2} \geq \frac{2}{L}$, a contradiction to $|y| < \frac{4}{L^2}$.

Exercise 14. *Prove Lemma 2.5! (Operations which preserve smoothness)*

Solution: For (i), we sum up the weighted smoothness conditions for all the f_i to obtain

$$\sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \leq \sum_{i=1}^m \lambda_i f_i(\mathbf{y}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \sum_{i=1}^m \lambda_i \frac{L_i}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

As the gradient is a linear operator, this equivalently reads as

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\sum_{i=1}^m \lambda_i L_i}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$

and the statement follows. For (ii), we apply smoothness of f at $\mathbf{x}' = A\mathbf{x} + \mathbf{b}$ and $\mathbf{y}' = A\mathbf{y} + \mathbf{b}$ to obtain

$$f(A\mathbf{x} + \mathbf{b}) \leq f(A\mathbf{y} + \mathbf{b}) + \nabla f(A\mathbf{x} + \mathbf{b})^\top(A(\mathbf{y} - \mathbf{x})) + \frac{L}{2}\|A(\mathbf{x} - \mathbf{y})\|^2.$$

As $\nabla(f \circ g)(\mathbf{x})^\top = \nabla f(A\mathbf{x} + \mathbf{b})^\top A$ (chain rule (Lemma 1.8), using that $Dg(\mathbf{x}) = A$, an easy consequence of Definition 1.7). This equivalently reads as

$$(f \circ g)(\mathbf{x}) \leq (f \circ g)(\mathbf{y}) + \nabla(f \circ g)(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|A(\mathbf{x} - \mathbf{y})\|^2.$$

The statement now follows from $\|A(\mathbf{x} - \mathbf{y})\| \leq \|A\|\|\mathbf{x} - \mathbf{y}\|$.

Exercise 16. *Let $a \in \mathbb{R}$. Prove that $f(x) = x^4$ is smooth over $X = (-a, a)$ and determine a concrete smoothness parameter L.*

Solution: The required inequality reads as

$$y^4 \leq x^4 + 4x^3(y - x) + \frac{L}{2}(x - y)^2 = -3x^4 + 4x^3 y + \frac{L}{2}(x^2 - 2xy + y^2) =: r_y(x).$$

We therefore want to ensure that $r_y(x) \geq y^4$ for all $x, y \in (-a, a)$. This is the case if and only if

$$\min\{r_y(x) : x \in [-a, a]\} \geq y^4, \quad \forall y \in [-a, a].$$

To minimize $r_y(x)$, we compute derivatives and get

$$\begin{aligned} r_y'(x) &= -12x^3 + 12x^2 y + Lx - Ly, \\ r_y''(x) &= -36x^2 + 24xy + L. \end{aligned}$$

We have $r_y'(y) = 0$. Moreover, for $L = 60a^2$, we get

$$r_y''(x) \geq -36a^2 - 24a^2 + L \geq 0,$$

so r_y is convex over $(-a, a)$ as a consequence of the second-order characterization Lemma 1.12. For $y \in (-a, a)$, $x = y$ is therefore indeed a minimum of r_y over $(-a, a)$ by Lemma 1.16. As we have

$$r_y(y) = y^4,$$

smoothness follows with $L = 60a^2$.

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 $f'' \leq \frac{1}{\gamma}$.

   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

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

   This condition is already satisfied for any $\gamma > 0$ if $g(x)$ is 1-Lipschitz continuous.

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!