

Machine Learning I: Foundations

Exercise Sheet 6

Prof. Marius Kloft

TA: Billy Joe Franks

16.06.2021

Deadline: 15.06.2021

- 1) In this exercise we will try to understand gradient descent, its initialization value, and its learning rate schedule. Consider only functions

$$f : \mathbb{R} \rightarrow \mathbb{R}.$$

- a) Find a constant learning rate schedule ($\lambda_i = c$), a convex function f (with a global minimum), and an initialization value x_0 such that the global minimum is never reached if you apply gradient descent with λ_i on f starting at x_0 . Prove that the function is convex. Prove that the global minimum is never reached.

Let $\lambda_i = 1$, $f(x) = x^2$, $x_0 = 1$. The gradient is just the derivative in this case so $\frac{\partial f(x)}{\partial x} = 2x$, now consider the update step

$$x_{i+1} \leftarrow x_i - \lambda_{i+1} \frac{\partial f(x)}{\partial x} \Big|_{x=x_i} \quad (1)$$

$$x_1 \leftarrow 1 - 1 \cdot 2x|_{x=1} = 1 - 2 = -1 \quad (2)$$

$$x_2 \leftarrow -1 - 1 \cdot 2x|_{x=-1} = -1 - 2 = -3 \quad (3)$$

As is evident (2) and (3) will alternate infinitely often, which means the minimum at $x = 0$ is never reached. As for the convexity of $f(x) = x^2$, we consider the second derivative

$$\frac{\partial^2 f(x)}{\partial x^2} = 2.$$

Since the second derivative is strictly positive $f(x) = x^2$ is strictly convex.

- b) For your choice of convex function and initialization value, is it possible to choose a learning rate schedule (constant or otherwise) such that the global minimum is reached? Prove your claim.

Let $f(x) = x^2$, $x_0 = 1$. If we chose $\lambda_1 = \frac{1}{2}$, then

$$x_1 \leftarrow 1 - \frac{1}{2} 2x|_{x=1} = 1 - 1 = 0.$$

The minimum is instantly reached. Alternatively we can choose $\lambda_t = \frac{1}{t}$, which is a typical choice, however convergence is harder to show and will not be done here due to space. This however might be a good practice exercise.

- c) Give an example of an unbounded function ($f(\mathbb{R}) = \mathbb{R}$), a learning rate schedule, and an initialization value for which gradient descent converges to a plateau (a critical point that is neither a minimum nor a maximum). The initialization value can not be chosen as this plateau.

Let $\lambda_i = \frac{1}{3}$, $f(x) = x^3$, $x_0 = 1$. Then

$$x_1 \leftarrow 1 - \frac{1}{3} 3x^2|_{x=1} = 1 - \frac{3}{3} = 0.$$

It instantly reaches the critical point. Obviously $f(x) = x^3$ is unbounded, if a certain value y should be reached the corresponding value $x = \sqrt[3]{y}$.

- d) Consider $f(x) = x^3$, is it possible to find a learning rate schedule which converges to the plateau at $x = 0$ for any initialization value? Prove your claim.

Interestingly it is possible to find such a schedule, by simply making the learning rate schedule depend on the derivative or the current x_i , however as λ_t only contains t as a variable, we will restrict ourselves to considering only learning rate schedules that depend on the current timestep. Originally this answer was supposed to be answered with no it is not possible to give such a learning rate schedule, however it is possible if we also consider negative values for λ_t . For example $\lambda_t = (-1)^t \frac{1}{t}$ should suffice, however proving this is beyond the scope of this lecture. As such we will further assume only positive values for λ_t , otherwise we are not strictly considering gradient descent (a negative λ_t would mean a gradient ascent step).

Now let $\forall t \lambda_t \in \mathbb{R}^+$, then $\lambda_1 = c \in \mathbb{R}^+$ is some constant. We will now find a starting value for which the first step takes us to the left of $x = 0$. We want to solve the following

$$x_0 - c3x_0^2 < 0.$$

Consider $x_0 = \frac{1}{c}$, then

$$x_0 - c3x_0^2 = \frac{1}{c} - \frac{3c}{c^2} = \frac{1-3}{c} = \frac{-2}{c} < 0.$$

This limits us to assuming $c > 0$, however as long as $\lambda_t = 0$ nothing is happening during the optimization, so we can simply remove all 0s from the schedule which results in $\forall t : \tilde{\lambda}_t \in \mathbb{R}_0^+$. As such the question can be answered with no making some assumptions, or yes otherwise.

2) (MANDATORY) 10 Points + 10 Points

Solve programming task 6. The two last programming exercises are mandatory.