# Machine Learning I: Foundations
# Exercise Sheet 5

Prof. Marius Kloft TA:Billy Joe Franks

02.06.2020
Deadline: 02.06.2020

**1) (MANDATORY) 10 Points**

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

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

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 $\lambda_i$ 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

$$x_{i+1} \leftarrow x_i - \lambda_{i+1} \left.\frac{\partial f(x)}{\partial x}\right|_{x=x_i} \tag{1}$$

$$x_1 \leftarrow 1 - 1\ 2x|_{x=1} \qquad =1-2 \qquad =-1 \tag{2}$$

$$x_2 \leftarrow -1 - 1\ 2x|_{x=-1} \qquad =-1+2 \qquad =1 \tag{3}$$

As is evident (2) and (3) will alternate infintely 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 stricly 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

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

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

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

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 $\lambda_t$ 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 $\lambda_t$. 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 $\lambda_t$, otherwise we are not strictly considering gradient descent (a negative $\lambda_t$ 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)** Let $k(\cdot,\cdot)$ be a kernel on $\mathbb{R}^d$. Let $\phi(\cdot)$ be the kernel mapping, i.e. $\langle\phi(x),\phi(y)\rangle = k(x,y)$. Let $x_1,\ldots,x_n \in \mathbb{R}^d$, $a = [a_1,\ldots,a_n]^T \in \mathbb{R}^n$ and $b = [b_1,\ldots,b_n]^T \in \mathbb{R}^n$. Let $K \in \mathbb{R}^{n\times n} = [k(x_i,x_j)]_{i,j}$ be the kernel matrix. Prove that

$$\left\langle \sum_{i=1}^n a_i\phi(x_i), \sum_{j=1}^n b_j\phi(x_j) \right\rangle = a^T K b$$

We have that

$$\begin{aligned}
\left\langle \sum_{i=1}^n a_i\phi(x_i), \sum_{j=1}^n b_j\phi(x_j) \right\rangle &= \sum_{i=1}^n \left\langle a_i\phi(x_i), \sum_{j=1}^n b_j\phi(x_j) \right\rangle \\
&= \sum_{i=1}^n a_i \left\langle \phi(x_i), \sum_{j=1}^n b_j\phi(x_j) \right\rangle \\
&= \sum_{i=1}^n a_i \sum_{j=1}^n b_j \left\langle \phi(x_i), \phi(x_j) \right\rangle \\
a^T K b &= \sum_{i=1,j=1}^n a_i b_j k(x_i,x_j)
\end{aligned}$$

**3)** For a matrix $X \in \mathbb{R}^{m\times n}$ let $X_{i,:} = [X_{i,1},\ldots,X_{i,n}]$ be the $i$-th row vector and $X_{:,i} = [X_{1,i},\ldots,X_{m,i}]^T$ be the $i$-th column vector. For $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times q}$ show that

$$XY = [X_{i,:}Y_{:,j}]_{i,j}$$

and

$$XY = \sum_{i=1}^n X_{:,i}Y_{i,:}.$$

Be sure to note the orientations of the vectors, some of these are row vectors and others are column vectors.

The first equality is the defnition of matrix product in an elementwise fashion. One way to interpret the multiplication of two matrices is as a way to concisely represent every inner product of the rows of $X$ with the colmns of $Y$ as another grid of number. So we will just prove this by looking at the $i, j$th entry of $XY$:

$$(XY)_{i,j} \quad = \quad \sum_{k=1}^{n} X_{i,k} Y_{k,j}$$

and since, for some vectors $x$ and $y$, we have that

$$x^T y = \sum_{i=1}^{d} x_i y_i$$

it follows that

$$\sum_{k=1}^{n} X_{i,k} Y_{k,j} = X_{i,:} Y_{:,j}.$$

Please convince yourself of every step presented here.

For the second equality we begin with the observation that for two vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ we have that

$$xy^T = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

From this we have that

$$\sum_{i=1}^{n} X_{:,i} Y_{i,:} = \sum_{i=1}^{n} \begin{bmatrix} X_{1,i} Y_{i,1} & X_{1,i} Y_{i,2} & \cdots & X_{1,i} Y_{i,q} \\ X_{2,i} Y_{i,1} & X_{1,i} Y_{i,2} & \cdots & X_{2,i} Y_{i,q} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m,i} Y_{i,1} & X_{m,i} Y_{i,2} & \cdots & X_{m,i} Y_{i,q} \end{bmatrix}.$$

Looking at some entry of this, say the $(k, l)$th entry, we get the summation

$$\sum_{i=1}^{n} X_{k,i} Y_{i,l} = X_{k,:} Y_{:,l}$$

which is equal to the $(k, l)$th entry of $XY$ by the first equality.

**4)** Solve programming task 5.