V. Schulz
M. Klar

SS 2022
Submission: 21.06.21, until 12:15

Numerical Optimization - Sheet 10

If you are a student in mathematics please solve the exercises with no tag and the ones with the tag **Mathematics**. If you are a data science student please solve the problems with no tag and those with the tag **Data Science**. Submissions with tags other than your subject count as bonus points. The tag **Programming** marks programming exercises.

Ex 1 (3 Points)

Let $f, g : \mathbb{R} \to \mathbb{R}$ be smooth functions. Consider the problem

$$\min_{\alpha \in \mathbb{R}} g(f(\alpha)) =: \min_{\alpha \in \mathbb{R}} F(\alpha).$$

(i) The concatenation $g(f(\alpha))$ can be made implicit by the constraints $g(c)$ s.t. $f(\alpha) - c = 0$. Rewrite the given problem as constrained optimzation problem without explicit concatenation (Don't forget the additional optimization parameter c).

(ii) Compute the first order necessary optimality conditions (Theorem 2.24) of the constrained problem.

(iii) Draw a connection between the adjoint λ in the optimal point and the derivative of F.

Ex 2 Programming (2+2 Points)

Consider the example *Warm-up: numpy* on the pytorch tutorials page. The tutorial solves

$$\min_{a,b,c,d \in \mathbb{R}} \sum_{i=1}^{N} \|a + bx_i + cx_i^2 + dx_i^3 - \sin(x_i)\|_2^2, \tag{1}$$

for samples $x_i \in [-\pi, \pi]$ using a gradient descent method.

(i) Propose a well known, classical method to solve the above problem in one step (you don't need to implement it).

(ii) Change the tutorial's code such that it solves

$$\min_{a,b,c,d \in \mathbb{R}} \sum_{i=1}^{N} \|a + b\,\sigma(c + dx_i) - \text{sign}(x_i + 0.5)\|_2^2, \tag{2}$$

where $\sigma(x) = 1/(1 + exp(-x))$ with $\sigma' = \sigma(1 - \sigma)$, and

$$\text{sign}(x) = \begin{cases} \frac{x}{|x|}, \, x \neq 0 \\ 0, \text{ otherwise,} \end{cases}$$

on the same interval $x_i \in [-\pi, \pi]$.

Ex 3 (2+1+1 Points)

This exercise is a simple example for a minimization problem over a random variable where the stochastic gradient (SG) method with constant step size does not converge. To that end, let $f_r : [-1, 1] \to \mathbb{R}$ be defined as $x \mapsto (x + r)^2$ for $r \in [-1, 1]$. Consider the optimization problem

$$\min_{x\in[-1,1]} \int_{-1}^{1} f_r(x)\,\frac{1}{2}dr. \tag{3}$$

(i) Compute the exact solution of problem (3) by hand.

(ii) Show that for a constant step size $0 < \alpha \leq 0.5$ the iterates x^k of the SG-method do not leave the domain $[-1, 1]$.

(iii) Show that for a constant step size $0 < \alpha \leq 0.5$ the algorithm does not converge to the solution.

Remark: The above problem can be interpreted as

$$\min_{x\in[-1,1]} \mathbb{E}f(x)$$

if we assume that $r \sim \mathcal{U}[-1, 1]$ is uniformly distributed on $[-1, 1]$. In order to show, that the algorithm does not converge it suffices to consider the distribution of $x^+ = x^* - \alpha\nabla f_r(x^*)$ where x^* already is the solution.

Ex 4 (2* Bonus- Points)

Assume you are given an input sample $x^{(i)} \in \mathbb{R}^n$ which could be an image or sound etc. which you feed into a trained neural network $N : \mathbb{R}^n \to \mathbb{R}^d$. So we interpret the network as fuction which maps an input to a probability distribution on $\{1, \dots, d\}$ and forget its parameterization for now. The input sample $x^{(i)} \in \mathbb{R}^n$ has a true class, namley $i \in \{1, \dots, d\}$. Assume for now that the network maps the sample to its correct class, i.e. assigns the highest probability to class i. The following optimization problem aims to find a distortion of $x^{(i)}$ by a small vector $\Delta x \in \mathbb{R}^n$.

Let $0 < \varepsilon < 1$, and $k, i \in \{1, \dots, d\}$, $k \neq i$. Let moreover $x^{(i)} \in \mathbb{R}^n$ be a fixed input sample of class i and $N_j(x)$ be the j-th component of $N(x)$ for $j \in \{1, \dots, d\}$. Consider the problem

$$\begin{aligned}
&\min_{\Delta x\in\mathbb{R}^n} \|\Delta x\|_2^2 \\
&\text{s.t. for all } j \in \{1, \dots, d\},\ j \neq k \\
&N_k(x^{(i)} + \Delta x) \geq (1 + \varepsilon)N_j(x^{(i)} + \Delta x).
\end{aligned}$$

What is the purpose of the given problem? What would be the consequence of the accessibility of very good solutions to it?