Continuous Optimization **Spring 2025**

Exercise 9: Adaptive Methods

Lecturer: Aurelien Lucchi

Problem 1 (Polyak's step size):

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex function we aim to minimize. Assume that we have access to a subdifirential $\partial f(x)$ of f at each point $x \in \mathbb{R}^d$. We denote by x^* the minimizer of f. We consider the following iterative process to solve the problem

$$x^{t+1} = \arg\min \frac{1}{2}\|x - x^t\|^2 \quad \text{such that} \quad x \in \mathcal{S}_t \coloneqq \left\{x \mid f(x^t) + \left\langle g^t, x - x^t \right\rangle \leq f(x^*)\right\}, \tag{1}$$

where $g^t \in \partial f(x^t)$ is any subgradient of f at point x^t.

a) Using the definition of a convex function, show that $\mathcal{S}_t$ is non-empty for all iterations t.

b) Write the Lagrangian function $L(x, \lambda)$ for the problem (1).

c) Write KKT conditions for the Lagrangian.

d) Solving KKT conditions find the closed-form solution, i.e., derive the expression for x^{t+1}.

e) Does the iteration process from d) remind you of an algorithm that has already been covered during the course? Write the name of the algorithm.

Problem 2 (Lispchitz Continuity Inequality):

Prove that for a L-smooth function $\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \leq (f(\mathbf{x}_t) - f^*)$

Problem 3 (Gauss-Newton Decomposition):

Let $\ell \colon \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}$ be a loss function (e.g., a squared loss $\ell(\mathbf{z}, \mathbf{y}) = \frac{1}{2}\|\mathbf{z} - \mathbf{y}\|^2$). Here $\mathbf{z}$ are predictions and $\mathbf{y}$ is the ground truth. Let us consider some ML model with parameters $\mathbf{w}$: $\mathbf{f} \colon \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}^m$ which takes features $\mathbf{x}$ as input and outputs a prediction $\mathbf{f}(\mathbf{w}, \mathbf{x})$. The quality of the model given an input $\mathbf{x}$ is $\ell(\mathbf{f}(\mathbf{w}, \mathbf{x}), \mathbf{y})$. Show that the Hessian of the function $g(\mathbf{w}) \coloneqq \ell(\mathbf{f}(\mathbf{w}, \mathbf{x}), \mathbf{y})$ can be written as

$$\nabla^2 g(\mathbf{w}) = J_f(\mathbf{w}, \mathbf{x})^\top \nabla^2_{\mathbf{z}} \ell(\mathbf{z}, \mathbf{y}) J_f(\mathbf{w}, \mathbf{x}) + \sum_{l=1}^{m} \frac{\partial \ell(\mathbf{z}, \mathbf{y})}{\partial \mathbf{z}_l} \nabla^2_{\mathbf{w}} \mathbf{f}_l(\mathbf{w}, \mathbf{x}), \tag{2}$$

where $J_f(\mathbf{w}, \mathbf{x})$ is the Jacobian of $\mathbf{f}$, i.e. $[J_f(\mathbf{w}, \mathbf{x})]_{i,j} = \frac{\partial \mathbf{f}_i(\mathbf{w}, \mathbf{x})}{\partial \mathbf{w}_j}$.