Winter term 2024/25
U. von Luxburg
E. Günther/ K. Frohnapfel

Assignment 08

Mathematics for Machine Learning

Submission due Friday **13.12.2024, 23:59** via Ilias

Justify all your claims.

Exercise 1 (**Online gradient descent**, 3+2+2+1 points).
Online learning is a machine learning method in which not all of the training data is available from the beginning but sequentially revealed. The challenge lies in choosing parameters such that the model fits the (unknown) data revealed in the next step well. Online learning is used, for example, in scenarios where the data is dynamic, like the prediction of prices in financial markets. One possibility of modelling this is by a loss function that changes with time.

Our setup is the following: At each time step t

- The learner chooses parameters w_t
- An adversary (or nature) chooses a convex loss function f_t
- The loss is $f_t(w_t)$.

We fix a convex set $U \subseteq \mathbb{R}^d$ containing 0 as our parameter space. Our objective after T rounds is

$$R_T = \max_{u \in U} \sum_{t=1}^{T} (f_t(w_t) - f_t(u)),$$

which is the cumulative difference between the algorithm's loss and the loss w.r.t. the optimal parameter in hindsight. R_T is also called *regret.*

We define *online gradient descent* by

$$w_0 = 0, \quad w_{t+1} = P_U(w_t - \eta \nabla f_t(w_t)),$$

where $P_U(v) = \arg\min_{u \in U} \|u - v\|$ is the projection onto U and $\eta \in \mathbb{R}$.

In the following, we will prove that online gradient descent is a robust algorithm suited for our scenario, as it yields a regret that only scales linearly in $\sqrt{T}$.

a) Prove the *Pythagorean inequality,* which states

$$\|y - \hat{y}\|^2 + \|\hat{y} - u\|^2 \leq \|y - u\|^2$$

for any $u \in U$, $y \in \mathbb{R}^d$ and $\hat{y} = P_U(y)$.

b) Show that for any $u \in U$ it holds

$$\langle w_t - u, \nabla f_t(w_t) \rangle \leq \frac{\|w_t - u\|^2 - \|w_{t+1} - u\|^2}{2\eta} + \frac{\eta}{2} \|\nabla f_t(w_t)\|^2.$$

c) Given $\|u\| \leq D$ and $\nabla|f_t(u)| \leq G$ for all $u \in U$, prove that it holds

$$R_T = \max_{u \in U} \sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} T G^2.$$

d) What choice of η minimizes the regret and which worst-case regret does it produce according to the upper bound?

Exercise 2 (**Formal proof of the existence of Langrange multipliers**, 4 points)**.**
For some differentiable functions $f, g : \mathbb{R}^d \to \mathbb{R}$, consider the optimization problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} f(x) \\ &\text{subject to } g(x) = 0. \end{aligned}$$

Prove that if x^* is a minimum, it holds $\nabla f(x^*) + \nu \nabla g(x^*) = 0$ for some $\nu \in \mathbb{R}$.

You may assume that there exists a parametrization of the constraint boundary, that is, there exists an injective differentiable function $h : \mathbb{R}^{d-1} \to \mathbb{R}^d$ whose Jacobian has full rank, such that the image of h equals $\left\{x \in \mathbb{R}^d \mid g(x) = 0\right\}$.
Hint: Remember the chain rule from assignment 6, exercise 4c) and prove that $\nabla f(x^*)$ *and* $\nabla g(x^*)$ *are both perpendicular to* $\operatorname{range}(\mathrm{D}h(p))$, *where* $h(p) = x^*$.

Exercise 3 (**Differentiable approximation of $\mathcal{L}^1$-approximation**, 3+4+1 points)**.**
The function $\varphi(u) = \sqrt{u^2 + \varepsilon}$, with parameter $\varepsilon > 0$, is sometimes used as a differentiable approximation of the absolute value function $|u|$. To approximately solve the $\mathcal{L}^1$-norm approximation problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_1, \tag{1}$$

where $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, we instead solve the (differentiable) problem

$$\min_{x \in \mathbb{R}^m} \sum_{i=1}^{m} \varphi(a_i \cdot x - b_i), \tag{2}$$

where a_i is the i-th row of A. We assume $\operatorname{rk}(A) = n$. In the following, we will derive an error bound for the relaxed problem.

a) Use the equivalent form

$$\begin{aligned} &\min_{y \in \mathbb{R}^m, x \in \mathbb{R}^n} \|y\|_1 \\ &\text{subject to } y - Ax + b = 0, \end{aligned}$$

of the $\mathcal{L}^1$-norm approximation problem to prove that its dual problem is given by

$$\begin{aligned} &\max_{\lambda \in \mathbb{R}^m} \lambda^t b \\ &\text{subject to } |\lambda_i| \leq 1 \quad \forall i = 1, ..., m \\ &\qquad \lambda^t A = 0. \end{aligned}$$

b) Let $p^* \in \mathbb{R}$ denote the optimal value of the $\mathcal{L}^1$-norm approximation problem (1). Let $\hat{x}$ denote the optimal solution of the approximate problem (2), and let $\hat{r} = A\hat{x} - b$ denote the associated residual. Show that

$$p^* \geq \sum_{i=1}^{m} \frac{\hat{r}_i^2}{\sqrt{\hat{r}_i^2 + \varepsilon}}.$$

Hint: Use the dual problem derived in a) for some suitable λ.

c) Conclude that

$$\|A\hat{x} - b\|_1 \leq p^* + \sum_{i=1}^{m} |\hat{r}_i| \left(1 - \frac{|\hat{r}_i|}{\sqrt{\hat{r}_i^2 + \varepsilon}} \right),$$

which gives an error bound for the relaxed problem.