# *Machine Learning B*

*2021-2022*

## Home Assignment 1

**Christian Igel** **Sadegh Talebi**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **November 30, 2021, 22:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your full source code in the PDF file, only selected lines if you are asked to do so.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.
- Do NOT zip the PDF file, since zipped files cannot be opened in speed grader. Zipped PDF submissions will not be graded.
- Your PDF report should be self-sufficient. It should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted, please use the provided latex template to write your report.

# 1 Convex Functions

For each function below, determine whether it is convex, strictly convex, strongly convex or none of the above.

(a) $f(x) = x\log(x) + (1-x)\log(1-x), \quad x \in (0,1)$

(b) $f(x) = -\sqrt{x+1}, \quad x \geq 1$

(c) $f(x) = (x_1 - 3x_2)^2 + (x_1 - 2x_2)^2, \quad x = (x_1, x_2) \in \mathbb{R}^2$

(d) $f(x) = \|x\|_1, \quad x \in \mathbb{R}^d$

(e) $f(x) = a^\top x + b + \lambda\|x\|_2^2, \quad x \in \mathbb{R}^d$

# 2 KKT Conditions

Consider the following optimization problem:

$$\mathsf{P}: \qquad \min \sum_{i=1}^{d} \frac{1}{x_i + \alpha_i} \tag{1}$$

$$\text{subject to: } \sum_{i=1}^{d} x_i \leq C, \tag{2}$$

$$x_i \in [m_i, M_i], \quad \forall i = 1, \ldots, d, \tag{3}$$

where $\alpha_i, m_i, M_i, i = 1, \ldots, d$ and $C$ are positive parameters.

(a) Rewrite $\mathsf{P}$ using the standard form (see Definition 3.1 in Igel-LN).

(b) Show that $\mathsf{P}$ is a convex optimization problem.

(c) Write down the Lagrangian function for $\mathsf{P}$.

(d) Write down the KKT conditions (Sufficiency Theorem) for $\mathsf{P}$.

(e) Form the dual function and dual problem associated to $\mathsf{P}$.

# 3 Perceptron as a Subgradient Descent Algorithm

Let $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathbb{R}^d \times \{+1, -1\})^m$. Assume that there exists $w \in \mathbb{R}^d$ such that for every $i \in \{1, \ldots, m\}$, we have $y_i \langle w, x_i \rangle \geq 1$, and let $w^\star$ be a vector that has the minimal norm among all vectors that satisfy the preceding requirement. Let $R = \max_i \|x_i\|$. Consider the function

$$f(w) = \max_{i \in \{1,\ldots,m\}} (1 - y_i \langle w, x_i \rangle).$$

(a) Show that $\min_{w:\|w\| \leq \|w^\star\|} f(w) = 0$ and show that any $w$ for which $f(w) < 1$ separates the examples in $S$.

(b) Show how to calculate a subgradient of $f$.

(c) Describe and analyze the subgradient descent algorithm for this case.

# 4 Kernels

The first question should improve the understanding of the geometry of the kernel-induced feature space. You can directly use the result to implement a kernel nearest-neighbor algorithm. The second question should make you more familiar with the basic definition of the important concept of positive definiteness. The third question is important to understand the real dimensionality of learning problems using a linear kernel – one reason why linear kernels are often treated differently in efficient implementations.

## 4.1 Distance in feature space

Given a kernel $k$ on input space $\mathcal{X}$ defining RKHS $\mathcal{H}$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ denote the corresponding feature map (think of $\Phi(x) = k(x, .)$). Let $x, z \in \mathcal{X}$. Show that the distance of $\Phi(x)$ and $\Phi(z)$ in $\mathcal{H}$ is given by

$$\|\Phi(x) - \Phi(z)\| = \sqrt{k(x, x) - 2k(x, z) + k(z, z)}$$

(if distance is measured by the canonical metric induced by $k$).

## 4.2 Sum of kernels

Let $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive-definite kernels.

Prove that $k(x, z) = a \cdot k_1(x, z) + b \cdot k_2(x, z)$ for $a, b \in \mathbb{R}^+$ is also positive-definite.

## 4.3 Rank of Gram matrix

Let the input space be $\mathcal{X} = \mathbb{R}^d$. Assume a linear kernel, $k(x, z) = x^{\mathrm{T}} z$ for $x, z \in \mathbb{R}^d$ (i.e., the feature map $\Phi$ is the identity) and $m$ input patterns $x_1, \ldots, x_m \in \mathbb{R}^d$.

Prove a non-trivial upper bound on the rank of the Gram matrix from the $m$ input patterns in terms of $d$ and $m$.