Deep learning and continuous optimization
Submission deadline:
5 March, 14:00

Exercise 1 (2pts)**.** Give an example of a function $f : \mathbb{R}^n \to \mathbb{R}$ such that $\|\nabla f(x)\|_2 \leq 1$ for all $x \in \mathbb{R}^n$ but the Lipschitz constant of its gradient is unbounded.

Exercise 2 (2pts)**.** Let $K \subseteq R$ be the set $K = \{x \in R \mid a^T x \leq b\}$, for $b \in R$ and some non-zero vector $a \in R$. Let $y \in R$. Find an explicit expression for the projection of y on to the set K (w.r.t. Euclidean norm) in terms of y, a and b.

Exercise 3 (2pts)**.** Since the function $x \mapsto \max\{x_1, \ldots, x_n\}$ is not differentiable, one often considers the so-called soft-max function

$$s_\alpha(x) := \frac{1}{\alpha} \log \left(\sum_{i=1}^{n} e^{\alpha x_i} \right)$$

for some $\alpha > 0$, as a replacement for the maximum. Prove that

$$\max\{x_1, \ldots, x_n\} \leq s_\alpha(x) \leq \frac{\log n}{\alpha} + \max\{x_1, \ldots, x_n\},$$

thus the larger α is, the better approximation we obtain. Further, prove that for every $x \in \mathbb{R}^n$,

$$\|\nabla s_\alpha(x)\|_\infty \leq 1.$$

Exercise 4 (3pts)**.** Let $G = (V, E)$ be an undirected graph and let B be the $|E| \times |V|$ matrix such that the ith row corresponds to the ith edge. Let the ith edge be uv. Then the ith row of B is $\chi_u - \chi_v$, where χ_u is the vector in $\mathbb{R}^{|V|}$ with one in the uth position and zero otherwise. The $|V| \times |V|$ matrix L defined by $B^T B$ is called the Laplacian matrix of the graph.

(a) Prove that $L = D - A$, where D is the diagonal degree matrix (i.e. $D_{v,v} = deg(v)$), and A is the adjacency matrix of the graph (i.e. $A_{u,v} = 1$ if $uv \in E$).

(b) Show that the all-1 vector is in the kernel of L.

(c) Prove that when the graph is connected, the kernel is just the one-dimensional space spanned by the all-1 vector.

Exercise 5 (1pt)**.** Prove that for all $p \in \Delta_n$, $D_{KL}(p, p^1) \leq \log n$. Here p^1 is the uniform probability distribution with $p^1_i = \frac{1}{n}$ for $i = 1, \ldots, n$, while Δ_n denotes the probability simplex.