MATH3431 Machine Learning and Neural Networks III — Epiphany term 2023

Exercise sheet

Lecturer: Georgios P. Karagiannis — georgios.karagiannis@durham.ac.uk

Part 1. Stochastic learning

Exercise 1. ($\star$)Let $f:\mathbb{R}^{d}\rightarrow\mathbb{R}$ such that $f\left(w\right)=g\left(<w,x>+y\right)$ or some $x\in\mathbb{R}^{d}$, $y\in\mathbb{R}$. If $g$is convex function then $f$is convex function.

Solution. Let $u,v\in\mathbb{R}^{d}$ and $a\in[0,1]$. It is

$$\begin{aligned}
f\left(\alpha u+\left(1-\alpha\right)v\right)= & g\left(<\alpha u+\left(1-\alpha\right)v,x>+y\right)\\
= & g\left(<\alpha u,x>+<\left(1-\alpha\right)v,x>+y\right)\\
= & g\left(\alpha\left(<u,x>+y\right)+\left(1-\alpha\right)\left(<v,x>+y\right)\right) && y=\alpha y+\left(1-\alpha\right)y\\
\le & \alpha g\left(<u,x>+y\right)+\left(1-\alpha\right)g\left(<v,x>+y\right) && (g\text{ is convex})\\
= & \alpha f\left(u\right)+\left(1-\alpha\right)f\left(v\right)
\end{aligned}$$

Exercise 2. ($\star$)Let functions g_{1} be ρ_{1}-Lipschitz and g_{2} be ρ_{2}-Lipschitz. Then f with $f\left(x\right)=g_{1}\left(g_{2}\left(x\right)\right)$ is $\rho_{1}\rho_{2}$-Lipschitz.

Solution.

$$\begin{aligned}
\left|f\left(w_{1}\right)-f\left(w_{2}\right)\right|= & \left|g_{1}\left(g_{2}\left(w_{1}\right)\right)-g_{1}\left(g_{2}\left(w_{2}\right)\right)\right|\\
\le & \rho_{1}\left|g_{2}\left(w_{1}\right)-g_{2}\left(w_{2}\right)\right|\\
\le & \rho_{1}\rho_{2}\left|w_{1}-w_{2}\right|
\end{aligned}$$

Exercise 3. ($\star$)Let $f:\mathbb{R}^{d}\rightarrow\mathbb{R}$ with $f\left(w\right)=g\left(\langle w,x\rangle+y\right)$ $x\in\mathbb{R}^{d}$ and $y\in\mathbb{R}$. Let $g:\mathbb{R}\rightarrow\mathbb{R}$ be a β-smooth function. Then f is a $\left(\beta\left\Vert x\right\Vert ^{2}\right)$-smooth.

Hint:: You may use Cauchy-Schwarz inequality $\langle y,x\rangle\le\left\Vert y\right\Vert \left\Vert x\right\Vert$

$$
\begin{aligned}
f(v) =& g(\langle w, x\rangle + y) \\
&\leq g(\langle w, x\rangle + y) + g'(\langle w, x\rangle + y)\langle v - w, x\rangle + \frac{\beta}{2}(\langle v - w, x\rangle)^2 && (g \text{ is smooth}) \\
&\leq g(\langle w, x\rangle + y) + g'(\langle w, x\rangle + y)\langle v - w, x\rangle + \frac{\beta}{2}(\|v - w\|\,\|x\|)^2 && (\text{Cauchy-Schwatz inequality}) \\
&= f(w) + \langle \nabla f(w), v - w\rangle + \frac{\beta\|x\|^2}{2}\|v - w\|^2
\end{aligned}
$$

Exercise 4. $(\star) f : S \to \mathbb{R}$ is ρ-Lipschitz over an open convex set S if and only if for all $w \in S$ and $v \in \partial f(w)$ it is $\|v\| \leq \rho$.

Hint:: You may use Cauchy-Schwarz inequality $\langle y, x\rangle \leq \|y\|\,\|x\|$

Solution. $\Longrightarrow$ Let $f : S \to \mathbb{R}$ be ρ-Lipschitz over convex set S, $w \in S$ and $v \in \partial f(w)$.

- Since S is open we get that there exist $\epsilon > 0$ such as $u := w + \epsilon\frac{v}{\|v\|}$ where $u \in S$. So $\langle u - w, v\rangle = \epsilon\|v\|$ and $\|u - w\| = \epsilon$.
- From the subgradient definition we get

$$f(u) - f(w) \geq \langle u - w, v\rangle = \epsilon\|v\|$$

- From the Lipschitzness of $f(\cdot)$ we get

$$f(u) - f(w) \geq \rho\|u - w\| = \rho\epsilon$$

Therefore $\|v\| \leq \rho$.

Proof. $\Longleftarrow$ It is for all $w \in S$ and $v \in \partial f(w)$ it is $\|v\| \leq \rho$. $\square$

- For any $u \in S$, it is

$$
\begin{aligned}
f(w) - f(u) \leq& \langle v, w - u\rangle && (\text{ because } v \in \partial f(w)\ ) \\
\leq& \|v\|\,\|w - u\| && \text{by Cauchy-Schwarz inequality} \\
\leq& \rho\|w - u\| && \text{because } \|v\| \leq \rho
\end{aligned}
\tag{1}
$$

- Similarly it results $u, w \in S$

$$f(w) - f(u) \leq \langle v, u - w\rangle\|v\| \leq \|v\|\,\|u - w\| \leq \rho\|u - w\|$$

from (1) because w, u can be swaped in (1) as they both are any values in S.

Exercise 5. $(\star)$Let $g_1(w), ..., g_r(w)$ be r convex functions, and let $g(\cdot) = \max_{\forall j}(g_j(\cdot))$. Show that for some w it is $\nabla g_k(w) \in \partial g(w)$ where $k = \arg\max_j(g_j(w))$ is the index of function $g_j(\cdot)$ presenting the greatest value at w.

Since g_j is convex, for all u

$$g_j(u) \geq g_j(w) + \langle u - w, \nabla g_j(w) \rangle$$

However $g(u) = \max_{\forall j}(g_j(u)) \geq g_j(u)$ for any j, and $g(w) = g_j(w)$ at w. Then

$$g(u) \geq g(w) + \langle u - w, \nabla g_j(w) \rangle$$

Then by the definition of the sub-gradient $\nabla g_j(w) \in \partial g(w)$
