MATH3431 Machine Learning and Neural Networks III — Epiphany term 2025

Exercise sheet

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

Part 1. Convex learning problems

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 $x\in\mathbb{R}^{d}$, $y\in\mathbb{R}$. Show that: 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

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

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

Solution.

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

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 show that f is a $\left(\beta\left\Vert x\right\Vert ^{2}\right)$-smooth.

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

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

Exercise 4. ($\star$)Show that $f : S \to \mathbb{R}$ 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

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

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

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

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

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

- 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 $f(\cdot) = \max_{\forall j}(g_j(\cdot))$. Show that for some w it is $\nabla g_k(w) \in \partial f(w)$ where $k = \arg\max_j(g_j(w))$ is the index of function $g_j(\cdot)$ presenting the greatest value at w.

Solution. Since g_k is convex, for all u

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

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

$$\begin{aligned} f(u) \geq& g_k(u) \\ \geq& g_k(w) + \langle u - w, \nabla g_k(w) \rangle \\ =& f(w) + \langle u - w, \nabla g_k(w) \rangle \end{aligned}$$

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

Exercise 6. ($\star$)Consider the regression learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with predictor rule $h(x) = \langle w, x \rangle$ labeled by some unknown parameter $w \in \mathcal{W}$, loss function $\ell(w, (x, y)) = (\langle w, x \rangle - y)^2$, feature $x \in \mathcal{X}$, and target $y \in \mathbb{R}$. Let $\mathcal{W} = \mathcal{X} = \left\{\omega \in \mathbb{R}^d : |\omega| \leq \rho\right\}$ for some $\rho > 0$.

(1) Show that the resulting learning problem is Convex-Lipschitz-Bounded learning problem.
(2) Specify the parameters of Lipschitnzess.

Solution. According to the definitions given in the lecture:

- Convex-Lipschitz-Bounded Learning Problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with parameters ρ, and B, is called the learning problem whose the hypothesis class $\mathcal{H}$ is a convex set, for all $w \in \mathcal{H}$ it is $\|w\| \leq B$, and the loss function $\ell(\cdot, z)$ is convex and ρ-Lipschitz function for all $z \in \mathcal{Z}$.

I have:

Convexity: The function $g : \mathbb{R} \to \mathbb{R}$, defined by $g(a) = a^2$ is convex convex. Eg. $\frac{\mathrm{d}^2}{\mathrm{d}a^2} g(a) = 1 \geq 0$ is non-negative. The convexity of $\ell(w, z = (x, y))$ for all z follows as a composition of g with a linear function.

Lipschitzness: The function $g(a) = a^2$ is 1-Lipschitz since It is

$$|g(a_2) - g(a_1)| = \left|a_2^2 - a_1^2\right| = |(a_2 + a_1)(a_2 - a_1)| \leq 2\rho(a_2 - a_1) = 2\rho|a_2 - a_1|$$

Hence because $|x| \leq \rho$, $g(a)$ is $2\rho^2$-Lipschitz as a composition.

Boundness: The norm of each hypothesis w is bounded by ρ according to the assumptions.

Therefore,

(1) the learning problem under consideration is a Convex-Lipschitz-Bounded learning problem.
(2) the parameter of Lipschitzness is $2\rho^2$.

Exercise 7. ($\star$)If $f$is λ-strongly convex and u is a minimizer of f then for any w

$$f(w) - f(u) \geq \frac{\lambda}{2}\|w - u\|^2$$

Hint:: Use the definition, and set $\alpha \to 0$.

Solution.

The following is given as a homework (Formative assessment 1)

Exercise 8. $(\star)$ Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and β-smooth function.

(1) Show that for $v, w \in \mathbb{R}^d$

$$f(v) - f(w) \in \left(\langle \nabla f(w), v - w \rangle, \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2 \right)$$

(2) Show that for $v, w \in \mathbb{R}^d$ such that $v = w - \frac{1}{\beta} \nabla f(w)$, it is

$$\frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v)$$

(3) Additionally assume that $f(x) > 0$ for all $x \in \mathbb{R}^d$. Show that for $w \in \mathbb{R}^d$,

$$\|\nabla f(w)\| \leq \sqrt{2\beta f(w)}$$

Solution.

(1) If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth then it is

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2$$
$$f(v) - f(w) \leq \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2$$

If it is convex then it is

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle$$
$$f(v) - f(w) \geq \langle \nabla f(w), v - w \rangle$$

Together these conditions imply upper and lower bounds

$$f(v) - f(w) \in \left(\langle \nabla f(w), v - w \rangle, \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2 \right)$$

(2) For $v, w \in \mathbb{R}^d$ such that $v = w - \frac{1}{\beta} \nabla f(w)$, it is

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|_2^2 \quad \text{(due to smoothness)}$$
$$\iff f(w) - f(v) \leq f(w) - f(v)$$
$$\iff \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|_2^2 \leq f(w) - f(v)$$
$$\iff \left\langle \nabla f(w), \frac{1}{\beta} \nabla f(w) \right\rangle + \frac{\beta}{2} \left\| \frac{1}{\beta} \nabla f(w) \right\|_2^2 \leq f(w) - f(v)$$
$$\iff \frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v)$$
$$\|\nabla f(w)\|^2 \leq 2\beta (f(w) - f(v))$$

as $f(\cdot) \geq 0$

$$\|\nabla f(w)\|^2 \leq 2\beta f(w)$$

(3) From part 2, this is obvious because $f(x) > 0$ for all $x \in \mathbb{R}^d$, as

$$\|\nabla f(w)\|^2 \leq 2\beta f(w) \Leftrightarrow \|\nabla f(w)\| \leq \sqrt{2\beta f(w)}$$

The following is given as a homework (Formative assessment 1)

Exercise 9. ($\star$)Let $f : \mathbb{R}^d \to \mathbb{R}$ be a λ-strongly convex function. Assume that w^* is a minimizer of f i.e.

$$w^* = \arg\min_{w} \{f(w)\}$$

Show that for any $w \in \mathbb{R}^d$ it holds

$$f(w) - f(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

Hint: Use the definition of λ-strongly convex function, properly rearrange it, and ...

Solution. We use the definition of λ-strongly convex function; i.e. for all w, u, and $\alpha \in (0, 1)$ we have

$$f(\alpha w + (1-\alpha) u) \leq \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2} \alpha (1-\alpha) \|w - u\|^2 \Leftrightarrow$$
$$\frac{f(\alpha w + (1-\alpha) u) - f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2} (1-\alpha) \|w - u\|^2$$

For $u = w^*$ it is

$$\frac{f(\alpha w + (1-\alpha) w^*) - f(w^*)}{\alpha} \leq f(w) + f(w^*) - \frac{\lambda}{2} (1-\alpha) \|w - w^*\|^2$$

When $a \to 0$

$$\frac{\lambda}{2} \alpha (1-\alpha) \|w - w^*\|^2 \to 0$$

I know that w^* is the minimizer of f. So 0 is the minimizer of g with $g(a) = f(aw + (1-\alpha) w^*)$ hence when $a \to 0$

$$\frac{f(\alpha w + (1-\alpha) w^*) - f(w^*)}{\alpha} \to \left. \frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) \right|_{\alpha=0}$$

So

$$0 \leq f(w) + f(w^*) - \frac{\lambda}{2} \|w - w^*\|^2$$

which concludes the proof.

Exercise 10. ($\star$)Show that the function $J(x; \lambda) = \lambda \|x\|^2$ is 2λ-strongly convex

Solution. We just need to check that for all w, u, and $\alpha \in (0, 1)$ we have

$$J(aw + (1-\alpha) u; \lambda) \leq aJ(w; \lambda) + (1-\alpha) J(u; \lambda) - \frac{2\lambda}{2} \alpha (1-\alpha) \|w - u\|^2 \Longleftrightarrow$$
$$\|aw + (1-\alpha) u\|_2^2 \leq a \|w\|_2^2 + (1-\alpha) \|u\|_2^2 - a(1-\alpha) \|w - u\|_2^2 \Longleftrightarrow 0 \leq 0$$

Exercise 11. ($\star\star\star$)Consider a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with $\mathcal{H} \subset \mathbb{R}^d$, $d > 0$, and loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ which is convex, β-smooth and non-negative. Let $\mathfrak{A}$ be a learning algorithm with output $\mathfrak{A}(\mathcal{S})$ trained against training dataset $\mathcal{S} = \{z_1, ..., z_m\}$ of IID samples $z_1, ..., z_m \sim g$ where g is a data generating distribution. In particular, consider that $\mathfrak{A}(\mathcal{S})$ is the Regularized Loss Minimization learning rule that outputs a hypothesis in

$$\min_w \left\{ \hat{R}_{\mathcal{S}}(w) + \lambda \|w\|_2^2 \right\}$$

for $\lambda \geq \frac{2\beta}{m}$ where $\hat{R}_{\mathcal{S}}(w) = \frac{1}{m}\sum_{i=1}^m \ell(w, z_i)$ for all $w \in \mathcal{H}$.

(1) Prove that

$$\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S}))\right) \leq R_g(w) + \lambda \|w\|_2^2$$

for all $w \in \mathcal{H}$. $R_g(\cdot)$ denotes the risk function under the real data generating distribution g.

(2) Prove that

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g(\mathfrak{A}(\mathcal{S})) - \hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S}))\right) \leq \frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S}))\right).$$

Hint:: If needed you can use the following:

Let $\mathcal{S}^{(i)} = \{z_1, ..., z_{i-1}, z', z_{i+1}, ..., z_m\}$ be a set resulting from $\mathcal{S}$ by replacing its i-th element z_i with an independently drawn $z' \sim g$. Then

$$24\beta\ell(\mathfrak{A}(\mathcal{S}), z_i) + \lambda m \ell(\mathfrak{A}(\mathcal{S}), z_i) + 24\beta\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z'\right) - \lambda m \ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right), z_i\right) \geq 0$$

(3) Show that the learning algorithm $\mathfrak{A}$ is on-average-replace-one-stable with rate ε. Specify that rate ε as a function of β, λ, m and possibly any other user specified constants if needed. Explain how the shrinkage parameter λ, the training dataset size m, and the smoothness parameter β affect the stability of the learning algorithm $\mathfrak{A}$.

(4) Show that the expected risk is bounded as follows

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g(\mathfrak{A}(\mathcal{S}))\right) \leq \left(1 + \frac{48\beta}{\lambda m}\right)\left(R_g(w) + \lambda \|w\|_2^2\right)$$

for all $w \in \mathcal{H}$.

Solution.

(1) We have

$$\begin{aligned}\hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S})) &\leq \hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S})) + \lambda \|\mathfrak{A}(\mathcal{S})\|_2^2 \\ &\leq \hat{R}_{\mathcal{S}}(w) + \lambda \|w\|_2^2; \quad \forall w \in \mathcal{W}\end{aligned}$$

and by taking expectations w.r.t. $\mathcal{S}$, it is

$$\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}(\mathfrak{A}(\mathcal{S}))\right) \leq R_g(w) + \lambda \|w\|_2^2; \quad \forall w \in \mathcal{W} \tag{2}$$

because $\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}(\cdot)\right) = \frac{1}{m}\sum_{i=1}^m \mathrm{E}_{\mathcal{S}\sim g}(\ell(\cdot, z_i)) = R_g(\cdot)$.

(2) From a well known theorem to us, it is

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)$$

Now I ll gonna work on the second term as this is what I ve given in the hint...

It is

$$24\beta\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)+\lambda m\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)+24\beta\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z'\right)-\lambda m\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)\geq 0\Leftrightarrow$$

$$\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\leq\frac{24\beta}{\lambda m}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)+\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z'\right)\right)$$

Taking expectations

$$\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)\leq\frac{24\beta}{\lambda m}\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)+\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z'\right)\right)$$

Due to the sampling it is

$$\mathrm{E}_{\mathcal{S}}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)=\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z'\right)\right)=\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)$$

So I get

$$\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right).$$

So I get

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right).$$

(3) Okay, let's say, I did not do the previous part. I see it is

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right).$$

From a well known theorem to us, it is

$$\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)=\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)$$

So

$$\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)$$

but the expectation depends on m... so

$$\begin{aligned}\mathrm{E}_{\mathcal{S},z',i}\left(\ell\left(\mathfrak{A}\left(\mathcal{S}^{(i)}\right),z_i\right)-\ell\left(\mathfrak{A}\left(\mathcal{S}\right),z_i\right)\right)&\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)\\&\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(0\right)\right)\\&\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\max\hat{R}_{\mathcal{S}}\left(0\right)\right)\\&\leq\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\frac{1}{m}\sum_{i=1}^{n}\underbrace{\max\ell\left(0,z_i\right)}_{=C}\right)\\&\leq\frac{48\beta}{\lambda m}C\end{aligned}$$

so it is the learning algorithm $\mathfrak{A}$ is on-average-replace-one-stable with rate

$$\varepsilon = \frac{48\beta}{\lambda m} C$$

and $C = \max \ell\left(0, z_i\right)$ ...or whatever constant they pick.

Larger training sample size m, and larger regularization parameter λ (eg more parsimonious model) lead to a more stable learning algorithm. Smaller smoothness parameter (the gradient changes less wrt the argument) leads to more stable learning algorithm.

(4) We use the decomposition discussed in the lectures,

$$\begin{aligned}
\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) &= \underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)} + \underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right) - \hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)} \\
&\leq \mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) + \frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \\
&= \left(1 + \frac{48\beta}{\lambda m}\right)\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) \\
&\leq \left(1 + \frac{48\beta}{\lambda m}\right)\left(R_g\left(w\right) + \lambda\left\|w\right\|_2^2\right),\ \forall w \in \mathcal{H}
\end{aligned}$$

Exercise 12. ($\star$) Let $\{v_t; t=1,...,T\}$ be a sequence of vectors with $v_t \in \mathbb{R}^d$ and $d \in \mathbb{N}-\{0\}$. Consider an algorithm producing $\left\{w^{(t)}; t=1,2,3,...\right\}$ with

$$w^{(1)} = 0$$
$$w^{(t+1)} = w^{(t)} - \eta v_t$$

$w_t \in \mathbb{R}^d$ and $d \in \mathbb{N}-\{0\}$. Show that

(1) it is

$$\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta}\left(-\left\|w^{(t+1)} - w^*\right\|^2 + \left\|w^{(t)} - w^*\right\|^2\right) + \frac{\eta}{2}\|v_t\|^2$$

Hint:: Recall that

$$\|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2\langle x, y\rangle, \ \forall x, y \in \mathbb{R}^d, d \in \mathbb{N}-\{0\}$$

(2) it is

$$\sum_{t=1}^{T}\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta}\sum_{t=1}^{T}\left(-\left\|w^{(t+1)} - w^*\right\|^2 + \left\|w^{(t)} - w^*\right\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

(3) (continue) it is

$$\sum_{t=1}^{T}\langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

Solution.

(1) It is

$$\begin{aligned}\left\langle w^{(t)} - w^*, v_t \right\rangle =& \frac{1}{\eta}\left\langle w^{(t)} - w^*, \eta v_t \right\rangle \\ =& \frac{1}{\eta}\left(-\left\langle w^{(t)} - w^*, -\eta v_t \right\rangle\right)\end{aligned}$$

Then by using the Hint as

$$\langle x, y\rangle = \frac{1}{2}\left(\|x+y\|_2^2 - \|x\|_2^2 - \|y\|_2^2\right)$$

for $x = w^{(t)} - w^* \in \mathbb{R}^d$ and $y = -\eta v_t \in \mathbb{R}^d$, I get

$$\begin{aligned}\left\langle w^{(t)} - w^*, v_t \right\rangle =& \frac{1}{2\eta}\left(-\left\|w^{(t)} - w^* - \eta v_t\right\|^2 + \left\|w^{(t)} - w^*\right\|^2 + \|-\eta v_t\|^2\right) \\ =& \frac{1}{2\eta}\left(-\left\|w^{(t+1)} - w^*\right\|^2 + \left\|w^{(t)} - w^*\right\|^2 + \eta^2\|v_t\|^2\right) \\ =& \frac{1}{2\eta}\left(-\left\|w^{(t+1)} - w^*\right\|^2 + \left\|w^{(t)} - w^*\right\|^2\right) + \frac{\eta}{2}\|v_t\|^2\end{aligned}$$

(2) So

$$\sum_{t=1}^{T}\left\langle w^{(t)}-w^{*},v_t\right\rangle = \frac{1}{2\eta}\sum_{t=1}^{T}\left(-\left\|w^{(t+1)}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
$$=\frac{1}{2\eta}\left(\left\|w^{(1)}-w^{*}\right\|^2-\left\|w^{(T+1)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

(3) So

$$\sum_{t=1}^{T}\left\langle w^{(t)}-w^{*},v_t\right\rangle = \frac{1}{2\eta}\left(\left\|w^{(1)}-w^{*}\right\|^2-\left\|w^{(T+1)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
$$\le \frac{1}{2\eta}\left\|w^{(1)}-w^{*}\right\|^2+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
$$=\frac{1}{2\eta}\|w^{*}\|^2+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

Exercise 13. ($\star$) Let $\{v_t; t=1,...,T\}$ be a sequence of vectors. Consider an algorithm producing $\left\{w^{(t)}; t=1,2,3,...\right\}$ with

$$w^{(1)}=0$$
$$w^{\left(t+\frac{1}{2}\right)}=w^{(t)}-\eta v_t$$
$$w^{(t+1)}=\arg\min_{w\in\mathcal{H}}\left(\left\|w-w^{\left(t+\frac{1}{2}\right)}\right\|\right)$$

for $t=1,...,T$.

Hint: You can use the following Lemma

(Projection Lemma): Let $\mathcal{H}$ be a closed convex set and let v be the projection of w onto $\mathcal{H}$,i.e.

$$v=\arg\min_{x\in\mathcal{H}}\|x-w\|^2$$

then for every $u\in\mathcal{H}$ it is

$$\|v-u\|^2\le\|w-u\|^2$$

Show that

(1) it is

$$\langle w^{(t)}-w^{*},v_t\rangle\le\frac{1}{2\eta}\left(-\left\|w^{(t+1)}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\|v_t\|^2$$

(2) it is

$$\sum_{t=1}^{T}\langle w^{(t)}-w^{*},v_t\rangle\le\frac{1}{2\eta}\sum_{t=1}^{T}\left(-\left\|w^{(t+1)}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

(3) (continue) it is

$$\sum_{t=1}^{T}\langle w^{(t)}-w^{*},v_t\rangle \leq \frac{\|w^{*}\|^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

Comment: Above we show that Lemma 17 from “Handout 4: Gradient descent” holds even when a projection step is included. Hence, even if a projection step is included after the update step of the recursion of GD algorithm or the SGD algorithm the analysis in Section 4 in “Handout 4: Gradient descent” holds. Hence, even if a projection step is included after the update step of the recursion of SGD algorithm or the SGD algorithm the analysis in Section 3 in “Handout 5: Stochastic gradient descent” holds.

Solution.

(1) It is

$$\begin{aligned}
\left\langle w^{(t)}-w^{*},v_t\right\rangle &= \frac{1}{\eta}\left\langle w^{(t)}-w^{*},\eta v_t\right\rangle \\
&= \frac{1}{2\eta}\left(-\left\|w^{(t)}-w^{*}-\eta v_t\right\|^2+\left\|w^{(t)}-w^{*}\right\|+\eta^2\left\|v_t\right\|^2\right) \\
&= \frac{1}{2\eta}\left(-\left\|w^{(t+\frac{1}{2})}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|+\eta^2\left\|v_t\right\|^2\right) \\
&= \frac{1}{2\eta}\left(-\left\|w^{(t+\frac{1}{2})}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|\right)+\frac{\eta}{2}\left\|v_t\right\|^2 \\
&\leq \frac{1}{2\eta}\left(-\left\|w^{(t+1)}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|\right)+\frac{\eta}{2}\left\|v_t\right\|^2
\end{aligned}$$

because from the Projection Lemma

$$\left\|w^{(t+1)}-w^{*}\right\|^2\leq\left\|w^{(t+\frac{1}{2})}-w^{*}\right\|^2$$

(2) So

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle w^{(t)}-w^{*},v_t\right\rangle &\leq \frac{1}{2\eta}\sum_{t=1}^{T}\left(-\left\|w^{(t+1)}-w^{*}\right\|^2+\left\|w^{(t)}-w^{*}\right\|\right)+\frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&= \frac{1}{2\eta}\left(\left\|w^{(1)}-w^{*}\right\|^2-\left\|w^{(T+1)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2
\end{aligned}$$

(3) So

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle w^{(t)}-w^{*},v_t\right\rangle &\leq \frac{1}{2\eta}\left(\left\|w^{(1)}-w^{*}\right\|^2-\left\|w^{(T+1)}-w^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&\leq \frac{1}{2\eta}\left\|w^{(1)}-w^{*}\right\|^2+\frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2 \\
&= \frac{1}{2\eta}\left\|w^{*}\right\|^2+\frac{\eta}{2}\sum_{t=1}^{T}\left\|v_t\right\|^2
\end{aligned}$$

Exercise 14. ($\star$) [1]Consider the binary classification problem with inputs $x \in \mathcal{X}$ where $\mathcal{X} := \left\{x \in \mathbb{R}^d : \|x\|_2 \leq L\right\}$ for some given value $L > 0$, target $y \in \mathcal{Y}$ where $\mathcal{Y} := \{-1, +1\}$, and prediction rule $h_w : \mathbb{R}^d \to \{-1, +1\}$ with

$$h_w(x) = \operatorname{sign}\left(w^\top x\right) \tag{3}$$

$$= \operatorname{sign}\left(\sum_{j=1}^{d} w_j x_j\right) \tag{4}$$

Let the hypothesis class is

$$\mathcal{H} = \left\{x \to w^\top x : \forall w \in \mathbb{R}^d\right\} \tag{5}$$

In other words, the hypothesis $h_w \in \mathcal{H}$ is parametrized by $w \in \mathbb{R}^d$, it receives an input vector $x \in \mathcal{X} := \mathbb{R}^d$ and it returns the label $y = \operatorname{sign}\left(w^\top x\right) \in \mathcal{Y} := \{\pm 1\}$ where

$$\operatorname{sign}(\xi) = \begin{cases} -1, & \text{if } \xi < 0 \\ +1, & \text{if } \xi > 0 \end{cases}$$

Consider a loss function $\ell : \mathbb{R}^d \to \mathbb{R}_+$ with

$$\ell(w, z = (x, y)) = \max\left(0, 1 - y w^\top x\right) + \lambda \|w\|_2^2 \tag{6}$$

for some given value $\lambda > 0$.

Assume there is available a dataset of examples $S_n = \{z_i = (x_i, y_i); i = 1, ..., n\}$ of size n. Do the following:

(1) Show that the function $f : \mathbb{R} \to \mathbb{R}_+$ with $f(x) = \max(0, 1 - x)$ is convex in $\mathbb{R}$; and show that the loss (6) is convex.

   Hint:: You may use Note 11 from Lecture notes 2: Elements of convex learning problems.

(2) Show that the loss $\ell(w, z)$ for $\lambda = 0$ (6) is L-Lipschitz (with respect to w) when $x \in \mathcal{X}$ where $\mathcal{X} := \left\{x \in \mathbb{R}^d : \|x\|_2 \leq L\right\}$.

   Hint:: You may use the definition of Lipschitz function. Without loss of generality, you can consider any $w_1 \in \mathbb{R}^d$ and $w_2 \in \mathbb{R}^d$ such that $1 - y w_2^\top x \leq 1 - y w_1^\top x$, and then take cases $1 - y w_2^\top x >$ or < 0 and $1 - y w_1^\top x >$ or < 0 to deal with the max.

[1]We use standard notation

$$\operatorname{sign}(\xi) = \begin{cases} -1, & \text{if } \xi < 0 \\ +1, & \text{if } \xi > 0 \end{cases}$$

± 1 means either -1 or $+1$, $\mathbb{R}_+ := (0, +\infty)$, and $\|x\|_2 := \sqrt{\sum_{\forall j} (x_j)^2}$ for the Euclidean distance.

(3) Construct the set of sub-gradients $\partial f(x)$ for $x \in \mathbb{R}$ of the function $f : \mathbb{R} \to \mathbb{R}_{+}$ with $f(x) = \max(0, 1 - x)$. Show that the vector v with

$$v = \begin{cases} 2\lambda w, & y w^{\top} x > 1 \\ 2\lambda w, & y w^{\top} x = 1 \\ -yx + 2\lambda w, & y w^{\top} x < 1 \end{cases}$$

is $v \in \partial_{w} \ell(w, z = (x, y))$, aka a sub-gradient of $\ell(w, z = (x, y))$ at w, for any $w \in \mathbb{R}^{d}$.

(4) Write down the algorithm of online AdaGrad (Adaptive Stochastic Gradient Descent) with learning rate $\eta_t > 0$, batch size m, and termination criterion $t > T_{\max}$ for some $T_{\max} > 0$ in order to discover w^{*} such as

$$w^{*} = \arg \min_{\forall w : h_w \in \mathcal{H}} \left(\mathrm{E}_{z \sim g} \left(\ell(w, z = (x, y)) \right) \right) \tag{7}$$

The formulas in your algorithm shoud be implemented for the above learning problem and tailored to 3, 5, and 6.

(5) Use the R code given below in order to generate the dataset of observed examples $S_n = \{z_i = (x_i, y_i)\}_{i=1}^{n}$ that contains $n = 10^{6}$ examples with inputs x of dimension $d = 2$. Consider $\lambda = 0$. Use a seed $w^{(0)} = (0, 0)^{\top}$.

(a) By using appropriate values for m, η_t and $T_{\max}$, code in R the algorithm you designed in part 4, and run it.

(b) Plot the trace plots for each of the dimensions of the generated chain $\left\{ w^{(t)} \right\}$ against the iteration t.

(c) Report the value of the output w^{*}_{adaGrad} (any type) of the algorithm as the solution to (7).

(d) To which cluster y (i.e., -1 or 1) $x_{\text{new}} = (1, 0)^{\top}$ belongs?

```
# R code. Run it before you run anything else
#
data_generating_model <- function(n,w) {
z <- rep( NaN, times=n*3 )
z <- matrix(z, nrow = n, ncol = 3)
z[,1] <- rep(1,times=n)
z[,2] <- runif(n, min = -10, max = 10)
p <- w[1]*z[,1] + w[2]*z[,2] p <- exp(p) / (1+exp(p))
z[,3] <- rbinom(n, size = 1, prob = p)
ind <- (z[,3]==0)
z[ind,3] <- -1
x <- z[,1:2]
y <- z[,3]
return(list(z=z, x=x, y=y))
}
n_obs <- 1000000
w_true <- c(-3,4)
set.seed(2023)
out <- data_generating_model(n = n_obs, w = w_true)
set.seed(0)
z_obs <- out$z #z=(x,y)
x <- out$x
y <- out$y
#z_obs2=z_obs
#z_obs2[z_obs[,3]==-1,3]=0
#w_true <- as.numeric(glm(z_obs2[,3]~ 1+ z_obs2[,2],family = "binomial"
)$coefficients)
```

Solution.

Exercise 15. ($\star$) Assume a Bayesian model

$$\begin{cases} z_i|w & \stackrel{\text{ind}}{\sim} f\left(z_i|w\right),\ i=1,...,n \\ w & \sim f\left(w\right) \end{cases}$$

and consider that our objective is the discovery of MAP estimate w^* i.e.

$$w^* = \arg\min_{\forall w\in\Theta}\left(-\log\left(L_n\left(w\right)\right) - f\left(w\right)\right) = \arg\min_{\forall w\in\Theta}\left(-\sum_{i=1}^{n}\log\left(f\left(z_i|w\right)\right) - \log\left(f\left(w\right)\right)\right)$$

by using SGD with update

$$w^{(t+1)} = w^{(t)} + \eta_t \left(\frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) + \nabla_w \log \left(f \left(w^{(t)} \right) \right) \right)$$

for some randomly selected set $\mathcal{J}^{(t)} \subseteq \{1, ..., n\}^m$ of m integers from 1 to n via simple random sampling (SRS) with replacement. Show that

$$\mathrm{E}_{\mathcal{J}^{(t)} \sim \text{simple-random-sampling}} \left(\frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) \right) = \sum_{i=1}^{n} \nabla_w \log \left(f \left(z_i | w^{(t)} \right) \right)$$

Solution. It is

$$\begin{aligned}
\mathrm{E}_{\mathcal{J}^{(t)} \sim \text{SRS}} \left(\frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) \right) =& \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \mathrm{E}_{\mathcal{J}^{(t)} \sim \text{SRS}} \left(\nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) \right) \\
=& \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \mathrm{E}_{\mathcal{J}^{(t)} \sim \text{SRS}} \left(\nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) \right) \\
=& \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \frac{1}{n} \sum_{i=1}^{n} \nabla_w \log \left(f \left(z_i | w^{(t)} \right) \right) \\
=& \sum_{i=1}^{n} \nabla_w \log \left(f \left(z_i | w^{(t)} \right) \right)
\end{aligned}$$

It is $\mathrm{E}_{\mathcal{J}^{(t)} \sim \text{SRS}} \left(\nabla_w \log \left(f \left(z_j | w^{(t)} \right) \right) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla_w \log \left(f \left(z_i | w^{(t)} \right) \right)$ because the expectation is under the probability I get randomly an integer and for the jth on the probability is $1/n$ due to the random scheme. Also $\left| \mathcal{J}^{(t)} \right| = m$.

Part 3. Support Vector Machines

Part 4. The kernel trick

Part 5. Multi-class classification

Part 6. **Artificial Neural Networks**

Part 7. Gaussian process regression

Part 8. **Revision**