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 $f(w)=g(<w,x>+y)$ 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

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

Exercise 2. ($\star$)Let functions g_1 be ρ_1-Lipschitz and g_2 be ρ_2-Lipschitz. Then, show that, f with $f(x)=g_1(g_2(x))$ is $\rho_1\rho_2$-Lipschitz.

Solution.

$$\begin{aligned}|f(w_1)-f(w_2)|=&|g_1(g_2(w_1))-g_1(g_2(w_2))|\\\leq&\rho_1|g_2(w_1)-g_2(w_2)|\\\leq&\rho_1\rho_2|w_1-w_2|\end{aligned}$$

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

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

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

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

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

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

- Since S is open we get that there exist $\epsilon>0$ such as $u:=w+\epsilon\frac{v}{\left\Vert v\right\Vert }$ where $u\in S$. So $\left\langle u-w,v\right\rangle =\epsilon\left\Vert v\right\Vert$ and $\left\Vert u-w\right\Vert =\epsilon$.
- From the subgradient definition we get
$$f\left(u\right)-f\left(w\right)\geq\left\langle u-w,v\right\rangle =\epsilon\left\Vert v\right\Vert$$
- From the Lipschitzness of $f\left(\cdot\right)$ we get
$$f\left(u\right)-f\left(w\right)\leq\rho\left\Vert u-w\right\Vert =\rho\epsilon$$

  Therefore $\left\Vert v\right\Vert \leq\rho$.

  $\Longleftarrow$ It is for all $w\in S$ and $v\in\partial f\left(w\right)$ it is $\left\Vert v\right\Vert \leq\rho$.
- For any $u\in S$, it is
$$\begin{aligned}
f\left(w\right)-f\left(u\right)\leq & \left\langle v,w-u\right\rangle && (\text{ because }v\in\partial f\left(w\right)\text{ })\\
\leq & \left\Vert v\right\Vert \left\Vert w-u\right\Vert && \text{by Cauchy-Schwarz inequality}\\
\leq & \rho\left\Vert w-u\right\Vert && \text{because }\left\Vert v\right\Vert \leq\rho
\end{aligned}\tag{1}$$
- Similarly it results $u,w\in S$
$$f\left(w\right)-f\left(u\right)\leq\left\langle v,u-w\right\rangle \left\Vert v\right\Vert \leq\left\Vert v\right\Vert \left\Vert u-w\right\Vert \leq\rho\left\Vert u-w\right\Vert$$
  from (1) because w,u can be swaped in (1) as they both are any values in S.

Exercise 5. ($\star$)Let $g_{1}\left(w\right),...,g_{r}\left(w\right)$ be r convex functions, and let $f\left(\cdot\right)=\max_{\forall j}\left(g_{j}\left(\cdot\right)\right)$. Show that for some w it is $\nabla g_{k}\left(w\right)\in\partial f\left(w\right)$ where $k=\arg\max_{j}\left(g_{j}\left(w\right)\right)$ is the index of function $g_{j}\left(\cdot\right)$ 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

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

If it is convex then it is

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

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

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

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 $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}_+$ 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

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

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,

$$\begin{aligned}
\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right) &= \underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)} + \underbrace{\mathrm{E}_{\mathcal{S}\sim g}\left(R_g\left(\mathfrak{A}\left(\mathcal{S}\right)\right)-\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)}\\
&\le \mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)+\frac{48\beta}{\lambda m}\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)\\
&= \left(1+\frac{48\beta}{\lambda m}\right)\mathrm{E}_{\mathcal{S}\sim g}\left(\hat{R}_{\mathcal{S}}\left(\mathfrak{A}\left(\mathcal{S}\right)\right)\right)\\
&\le \left(1+\frac{48\beta}{\lambda m}\right)\left(R_g\left(w\right)+\lambda\left\Vert w\right\Vert_2^2\right),\ \forall w\in\mathcal{H}
\end{aligned}$$

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

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

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

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

(3) (continue) it is

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

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

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

(2) So

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

(3) So

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

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

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

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

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

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

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 $y = \text{sign}\left(w^\top x\right) \in \mathcal{Y} := \{\pm 1\}$ where

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

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

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

± 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

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

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.

(1) $f_1(x) = 0$ is convex, $f_2(x) = 1 - x$ is convex, hence from the example in the lecture notes, $f(x) = \max(f_1(x), f_2(x))$ is convex as well. Regarding the loss function, we just have $f_2(w) = 1 - yx^\top w$ which is convex as a composition due to linearity.

(2) Given a fixed example $(x, y) \in \left\{x \in \mathbb{R}^d : \|x'\|_2 \leq R\right\} \times \{-1, 1\}$.

Assume $w_1, w_2 \in \mathbb{R}^d$. Let $\ell_i = \max\left\{0, 1 - yx^\top w_i\right\}$, for $i = 1, 2$. It suffices to show that $|\ell_1 - \ell_2|_2 \leq R\,|w_1 - w_2|_2$. I take cases

Case-1: Assume $yx^\top w_1 \geq 1$ and $yx^\top w_2 \geq 1$ then $|\ell_1 - \ell_2|_2 = 0 \leq R\,|w_1 - w_2|_2$

Case-2: Assume that at least one of $yx^\top w_1 < 1$ or $yx^\top w_2 < 1$ but not both is true. Assume without loss of generality that $1 - yx^\top w_1 < 1 - yx^\top w_2$. Then

$$\begin{aligned}
|\ell_1 - \ell_2|_2 =& \ell_1 - \ell_2 \\
=& 1 - yx^\top w_1 - \max\left(0, 1 - yx^\top w_2\right) \\
\leq& 1 - yx^\top w_1 - \left(1 - yx^\top w_2\right) \\
=& yx^\top \left(w_2 - w_1\right) \\
\leq& y\left\|x^\top\right\|_2 \|w_1 - w_2\|_2 \quad \text{because} \quad a^\top b \leq \|a\| \|b\|
\end{aligned}$$

(3) It is

$$f(x) = \max(0, 1 - x) = \begin{cases} 0 & x > 1 \\ 0 & x = 1 \\ 1 - x & x < 1 \end{cases}$$

- For $x > 1$, f is differentiable so $\partial f(x) = \{f'(x)\} = \{0\}$.
- For $x < 1$, f is differentiable so $\partial f(x) = \{f'(x)\} = \{-1\}$.
- For $x = 1$, f is not differentiable. By definition I have that v is subgradient of $f(x)$ at $x = 0 \in S$ if

  $$\forall u \in \mathbb{R}, \ \ f(u) \geq f(x) + \langle u - x, v \rangle$$

  So, for $u \geq 1$, it is $0 \geq (u - 1)v \implies v \leq 0$, and for $u < 1$ it is $(1 - u) \geq (u - 1)v \implies v \geq -1$. Hence the common space is $v \in [0, 1]$ So $\partial f(x) = [0, 1]$. Hence,

  $$\partial f(x) = \begin{cases} 0, & x > 1 \\ [-1, 0], & x = 1 \\ -1, & x < 1 \end{cases}$$

Now regarding the loss $\partial_w \ell(w, z = (x, y))$

- for $yw^\top x > 1$ it is differentiable so $\nabla_w \ell(w, z = (x, y)) = \nabla_w \left(0 + \lambda \sum_{j=1}^d w_j^2\right) = 2\lambda w$; as

  $$\frac{\mathrm{d}}{\mathrm{d}w_j} \sum_{j'=1}^d w_{j'}^2 = 2\lambda w_j$$

- for $yw^\top x > 1$ it is differentiable so $\nabla_w \ell(w, z = (x, y)) = \nabla_w \left(1 - yw^\top x + \lambda \sum_{j=1}^d w_j^2\right) = yx + 2\lambda w$ as

  $$\frac{\mathrm{d}}{\mathrm{d}w_j}\left(1 - yw^\top x\right) = \frac{\mathrm{d}}{\mathrm{d}w_j}\left(1 - y\sum_{j'=1}^d w_{j'} x_{j'}\right) = -yx_j$$

- for $yw^\top x = 1$, $v = 0$ satisfies the definition of the sub-gradient

$$\forall u, \;\; f\left(u\right) \geq \cancelto{0}{f\left(w\right)} + \langle u - w, v \rangle$$
$$\max\left(0, 1 - yu^\top x\right) \geq 0 + \left(u - w\right)^\top 0$$

So

$$\begin{aligned}
\partial \ell\left(w, z = \left(x, y\right)\right) &= \partial\left(\max\left(0, 1 - yw^\top x\right) + \lambda \left\|w\right\|_2^2\right) \\
&= \partial\left(\max\left(0, 1 - yw^\top x\right)\right) + \partial\left(\lambda \left\|w\right\|_2^2\right) \\
&= \partial\left(\max\left(0, 1 - yw^\top x\right)\right) + \nabla\left(\lambda \left\|w\right\|_2^2\right) \\
& 0 + 2\lambda w
\end{aligned}$$

but $\partial\left(\lambda \left\|w\right\|_2^2\right) = \left\{\nabla\left(\lambda \left\|w\right\|_2^2\right)\right\}$ because $\lambda \left\|w\right\|_2^2$ is differentiable. Hence

$$\partial \ell\left(w, z = \left(x, y\right)\right) = 0 + 2\lambda w$$

Hence

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

(4)

Algorithm. *For $t = 1, 2, 3, \ldots$ iterate:*

(a) *Get a random sub-sample $\left\{\tilde{z}_i^{(t)} = \left(\tilde{x}_i^{(t)}, \tilde{y}_i^{(t)}\right); i = 1, \ldots, m\right\}$ of size m with or without replacement from the complete data-set $\mathcal{S}_n$.*

(b) *For $j = 1, \ldots, d$ (index j indicates the dimension of w) compute*

$$w_j^{(t+1)} = w_j^{(t)} - \eta_t \frac{1}{\sqrt{\left[G_t\right]_{j,j} + \epsilon}} \bar{v}_{t,j}$$

$\left[G_t\right]_{j,j} = \left[G_{t-1}\right]_{j,j} + \left(\bar{v}_{t,j}\right)^2$ *where* $\bar{v}_t = \frac{1}{m}\sum_{i=1}^m \tilde{v}_{t,i}$ *and*

$$\tilde{v}_{t,i} = \begin{cases} 2\lambda w^{(t)}, & \tilde{y}_i^{(t)} \left(w^{(t)}\right)^\top \tilde{x}_i^{(t)} > 1 \\ 2\lambda w^{(t)}, & \tilde{y}_i^{(t)} \left(w^{(t)}\right)^\top \tilde{x}_i^{(t)} = 1 \\ -\frac{1}{m}\tilde{y}_i^{(t)}\tilde{x}_i^{(t)} + 2\lambda w^{(t)}, & \tilde{y}_i^{(t)} \left(w^{(t)}\right)^\top \tilde{x}_i^{(t)} < 1 \end{cases}$$

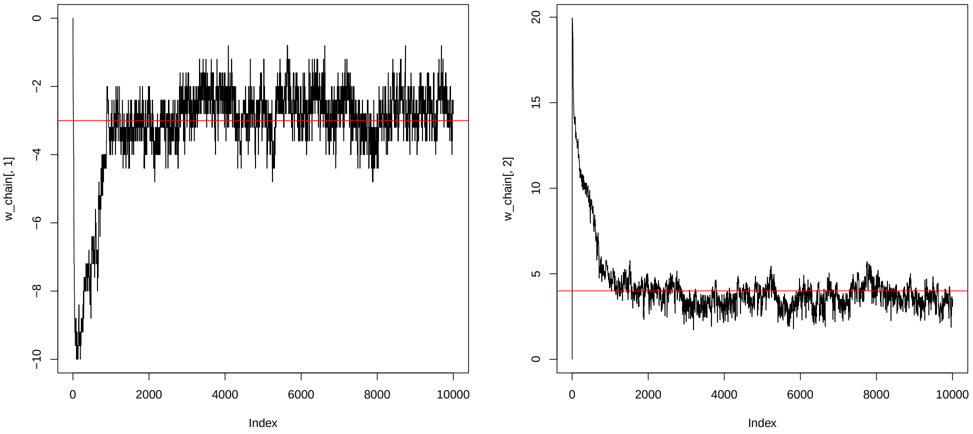
where index i indicates the sub-sample, and $\epsilon > 0$ small.

(c) *Terminate if a termination criterion is satisfied*

(5)

(a) The R code can be found in the link `https://raw.githubusercontent.com/georgios-stats/Machine_Learning_and_Neural_Networks_III_Epiphany_2025/main/Exercise_sheets/supplementary/q6_adagrad.R`

(b) The figures are presented below



(A) $w_1^{(t)}$

(B) $w_2^{(t)}$

FIGURE 0.1. trace plots

(c) I found $w = (-2.674615, 3.205785)$

(d) It belongs to -1

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

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

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

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

It is $\mathrm{E}_{\mathcal{J}^{(t)}\sim\mathrm{SRS}}\left(\nabla_w\log\left(f\left(z_j|w^{(t)}\right)\right)\right)=\frac{1}{n}\sum_{i=1}^{n}\nabla_w\log\left(f\left(z_i|w^{(t)}\right)\right)$ 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 $\left|\mathcal{J}^{(t)}\right|=m$.

Exercise 16. ($\star\star$) Consider a training data set $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^{m}$. Consider the Soft-SVM Algorithm that requires the solution of the following quadratic minimization problem (in a slightly modified but equivalent form to what we have discussed)

Primal problem:

$$(w^*, b^*, \xi^*) = \arg\min_{(w,b,\xi)} \left(\frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{m} \xi_i \right) \tag{8}$$

$$\text{subject to: } y_i \left(\langle w, x_i \rangle + b \right) \geq 1 - \xi_i, \ \forall i = 1, ..., m \tag{9}$$

$$\xi_i \geq 0, \ \forall i = 1, ..., m \tag{10}$$

for some user-specified fixed parameter $C > 0$.

(1) Specify the Lagrangian function L associated to the above primal quadratic minimization problem, where $\{\alpha_i\}$ are the Lagrange coefficients wrt (9), and $\{\beta_i\}$ are the Lagrange coefficients wrt (10). Write down any possible restrictions on the Lagrange coefficients.

(2) Compute the dual Lagrangian function denoted as $\tilde{L}$ as a function of the Lagrange coefficients and the data points $\mathcal{S}$.

(3) Apply the Karush–Kuhn–Tucker (KKT) conditions to the above problem, and write them down.

(4) Derive and write down the dual Lagrangian quadratic maximization problem, along with the inequality and equality constraints, where you seek to find $\{\alpha_i\}$.

(5) Justify why the i-th point x_i lies on the margin boundary when $\alpha_i \in (0, C)$ (beware it is $\alpha_i \neq C$), and why the i-th point x_i can lie inside the margin when $\alpha_i = C$.

(6) Given optimal values $\{\alpha_i^*\}$ for Lagrangian coefficients $\{\alpha_i\}$ as they are derived by solving the dual Lagrangian maximization problem in part 4, derive the optimal values w^* and b^* for the parameters w and b as function of the support vectors. Regarding parameter b it should be in the derived in the form

$$b^* = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(y_i - \sum_{j \in \mathcal{S}} \alpha_j^* y_j \langle x_j, x_i \rangle \right)$$

where you determine the sets $\mathcal{M}$ and $\mathcal{S}$.

(7) Report the halfspace predictive rule $h_{w,b}(x)$ of the above problem as a function of α^* and b^*.

Solution.

(1) It is

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{m} C\xi_i + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(\langle w, x_i \rangle + b \right) - \xi_i \right) - \sum_{i=1}^{m} \beta_i \xi_i \tag{11}$$

(2) Let α, β be fixed. We minimize (11) wrt w, b and we get

$$0 = \frac{\partial L}{\partial w}(w, b, \xi, \alpha, \beta) \implies w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{12}$$

$$0 = \frac{\partial L}{\partial b}(w, b, \xi, \alpha, \beta) \implies 0 = \sum_{i=1}^{m} \alpha_i y_i$$

$$0 = \frac{\partial L}{\partial \xi_i}(w, b, \xi, \alpha, \beta) \implies \alpha_i = C - \beta_i \tag{13}$$

and we substitute (12)-(13) in (11) and we get

$$\tilde{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle$$

(3) The Karush–Kuhn–Tucker (KKT) conditions applied to the above problem are

$$0 = \nabla \frac{1}{2} \|w\|_2^2 \nabla \sum_{i=1}^{m} C\xi_i + \nabla \sum_{i=1}^{m} \alpha_i (1 - y_i (\langle w, x_i \rangle + b) - \xi_i) - \nabla \sum_{i=1}^{m} \beta_i \xi_i \qquad \text{Stationarity}$$

$$1 - y_i (\langle w, x_i \rangle + b) - \xi_i \leq 0, \ \forall i = 1, ..., m \qquad \text{Primal feasibility}$$

$$\xi_i \geq 0$$

(14)

$$\alpha_i \geq 0 \ \forall i = 1, ..., m \qquad \text{Dual feasibility}$$

(15)

$$\beta_i \geq 0 \ \forall i = 1, ..., m$$

(16)

$$\alpha_i (1 - y_i (\langle w, x_i \rangle + b) - \xi_i) = 0, \ \forall i = 1, ..., m \qquad \text{Complementary slackness}$$

(17)

$$\beta_i \xi_i = 0, \ \forall i = 1, ..., m$$

(4) It is

$$\alpha^* = \arg \max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle \right) \tag{18}$$

$$\text{subject to } 0 = \sum_{i=1}^{m} \alpha_i y_i$$

$$\alpha_i \in [0, C] \ \ \forall i = 1, ..., m \tag{19}$$

constrain (19) results from (13), (15), and (14).

(5)

- By (12), if $\alpha_i = 0$ then x_i does not contribute to the computation of the weights.
- By (12), if $\alpha_i \neq 0$ (aka $\alpha_i > 0$), then x_i is a support vector and contributes.

- If $\alpha_i \in (0, C)$ (where $\alpha_i \neq C$) then (13) implies that $\beta_i > 0$. By (17) if $\beta_i > 0$ then $\xi_i = 0$. Hence, given these, from (16), it is $1 = y_i (\langle w, x_i \rangle + b)$ i.e. x_i lies on the boundary.
- If $\alpha_i = C$, then x_i can lie inside the boundary. This is because, from (13) $\beta_i = 0$. Hence
  - from (17) it may be $\xi_i = 0$ hence lie inside the margin, or
  - it may be $\xi_i > 0$ and hence be correctly classified when $\xi_i \in (0, 1]$, or
  - it may be $\xi_i > 0$ and hence be misclassified when $\xi_i \in (1, \infty)$.

(6) From (16), it is either $\alpha_i = 0$ or $(1 - y_i (\langle w, x_i \rangle + b) - \xi_i) = 0$. Let $\mathcal{S} = \{i : y_i (\langle w, x_i \rangle + b) = 1 - \xi_i\}$. From (12) , it is

$$w^* = \sum_{i \in \mathcal{S}} \alpha_i^* y_i x_i \tag{20}$$

Using (16) and summing up indexes in $\mathcal{M} = \{i : \alpha_i \in (0, C)\}$ for which $\xi_i = 0$ it is

$$b^* = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(y_i - \sum_{j \in \mathcal{S}} \alpha_j^* y_j \langle x_j, x_i \rangle \right)$$

(7) The formula is

$$\begin{aligned} h_{w,b}(x) &= \text{sign} \left(\langle w^*, x \rangle + b^* \right) \\ &= \text{sign} \left(\sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x \rangle + b^* \right) \qquad (21) \end{aligned}$$

Exercise 17. ($\star\star\star$) *[This is the Relevance Vector Machine. The Exercise is taken from "Exercise Sheet: Bayesian Statistics" of the module "Bayesian Statistics III/IV (MATH3361/4071)" taught in "Michaelmas term 2021". The supplementary material in the box was mainly provided for the students who had not been introduced to the SVM ideas or the Kernel trick -so it can be skipped. Also, the supplementary material in the box is presented with a statistical (geostatistical modeling) motivation. The exercise requires basic knowledge of Bayesian statistical inference and in particular the use of Bayes theorem for the computation of the posterior as well as basis probability density calculus. However, the exercise is a useful example of extending the SVM ideas to the Bayesian learning setting.]*

Regarding the statistical model: Long story, short (supplementary material)

Consider that we are interested in recovering the mapping

$$x \overset{\eta}{\longmapsto} \eta(x)$$

in the sense that $y \in \mathbb{R}$ is the response (output quantity) that depends on $x = (x_1, ..., x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$ which is the independent variable (input quantity) in a procedure; E.g.:,

- y: precipitation in log scale
- $x = (\text{longitude}, \text{latitude})$: geographical coordinates.

Consider a set of observed data $\{(y_i, x_i)\}_{i=1}^{n}$, which may be contaminated by additive noise of unknown variance; i.e.

$$y_i = \eta(x_i) + \epsilon_i,$$

where $\epsilon_i \overset{\text{IID}}{\sim} \text{N}\left(0, \sigma^2\right)$ and $\sigma^2 > 0$ is unknown. We wish to recover $\eta(x)$ by using the Tikhonov regularization on the functional space $\mathcal{H}$ such that

$$\eta = \arg\min_{\forall \tilde{\eta} \in \mathcal{H}} \left\{ \sum_{i=1}^{n} L\left(y_i - \tilde{\eta}\left(x_i\right)\right) + \lambda \left\| \tilde{\eta} \right\|_{\mathcal{H}}^2 \right\} \tag{22}$$

By assuming that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS), the solution to the above Ridge regularizes loss minimization problem is such that

$$\eta\left(x\right) = \beta_0 + \sum_{j=1}^{n} k(x, x_j)\beta_j = k(x)^\top \beta$$

where $k(x) = \left(1, k\left(x, x_1\right), ..., k\left(x, x_n\right)\right)^\top$, $k\left(x, x_j\right)$ is the reproducing kernel (such as $k_\phi\left(x, x_j\right) = \exp\left(-\phi \left\|x - x_j\right\|^2\right)$ for some known parameter $\phi > 0$), and $\beta \in \mathbb{R}^{n+1}$ is an unknown vector.

Consider the following Bayesian model[2]

$$\begin{cases} y|\beta, \sigma^2 & \sim \text{N}\left(K\beta, I\sigma^2\right) \\ \beta|\lambda & \sim \text{N}\left(0, D^{-1}\right), \quad D = \left(\lambda_0, \lambda_1, ..., \lambda_n\right) \\ \lambda_i & \overset{\text{iid}}{\sim} \text{d}\Pi\left(\lambda_i\right) \propto \lambda_i^{a-1} \exp\left(-b\lambda_i\right) \text{d}\lambda_i, \ \ \forall i = 1, ..., n \\ \sigma^2 & \sim \text{d}\Pi\left(\sigma^2\right) \propto \left(\sigma^2\right)^{c-1} \exp\left(-\frac{1}{\sigma^2} d\right) \text{d}\sigma^2 \\ \beta, \sigma^2 & \text{a priori independent} \end{cases}$$

where K is a known matrix with size $n \times (n+1)$ such that

$$K = \begin{bmatrix} 1 & k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}.$$

The quantities $a > 0$, $b > 0$, $c > 0$, $d > 0$, and $\phi > 0$ are considered as fixed.

[2] Dixit, A., & Roy, V. (2021). Posterior impropriety of some sparse Bayesian learning models. Statistics & Probability Letters, 171, 109039.

(1) When $b = 0$, show that a necessary condition for a valid posterior inference is $a \in (-1/2, 0)$ for any choice of prior for τ (i.e. any choice of (c, d)).

(2) Let $P = K\left(K^\top K\right)^{-1} K^\top$. Show that (2a) and (2b) are sufficient conditions for the Bayesian model to lead to a valid posterior inference

  (a) if $a > 0$ and $b > 0$, or

  (b) if $y^\top (I - P)\, y + 2d > 0$ and $c > -\frac{n}{2}$

(3) Does the the improper Uniform prior on the joint $\log(\lambda_i)$ and $\log\left(\sigma^2\right)$, i.e. $\pi\left(\log(\lambda_i), \log\left(\sigma^2\right)\right) \propto 1$, lead to a valid inference?

(4) Does the Jeffreys' prior $\pi(\lambda_i) \propto 1/\lambda_i$ lead to a valid inference?

Hint-1::

$$(y - K\beta)^\top (y - K\beta) + (\beta - \mu)^\top V^{-1} (\beta - \mu) = (\beta - \mu^*)^\top (V^*)^{-1} (\beta - \mu^*) + S^*;$$

$$S^* = \mu^\top V^{-1} \mu - (\mu^*)^\top (V^*)^{-1} (\mu^*) + y^\top y; \qquad V^* = \left(V^{-1} + K^\top K\right)^{-1}; \qquad \mu^* = V^*\left(V^{-1}\mu + K^\top y\right)$$

Hint-2:: Sherman-Morrison-Woodbury formula:

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U \left(C^{-1} + V A^{-1} U\right)^{-1} V A^{-1}$$

Hint-3::

$$-\frac{y^\top y}{2\sigma^2} \le -\frac{y^\top \left(I\sigma^2 + K D^{-1} K^\top\right)^{-1} y}{2} \le -\frac{1}{2\sigma^2} y^\top (I - P)\, y$$

where $P = K\left(K^\top K\right)^{-1} K$.

Hint-4:: It is given that $\int_{(0,\infty)} \frac{t^{-(a+1)}}{(\xi + t)^{1/2}} \mathrm{d}t < \infty$ if and only if $a \in (-1/2, 0)$.

Solution.

The posterior pdf is given by

$$\pi\left(\beta, \sigma^2, \lambda | y\right) = \frac{f\left(y|\beta, \sigma^2\right) \pi\left(\beta, \sigma^2, \lambda\right)}{f(y)}$$

and is proper iff $f(y) < \infty$ where

$$f(y) = \int \left(\underbrace{\int \left(\underbrace{\int f\left(y|\beta, \sigma^2\right) \pi(\beta|\lambda)\, \mathrm{d}\beta}_{= f(y|\lambda, \sigma^2)} \right) \pi(\lambda)\, \mathrm{d}\lambda}_{= f(y|\sigma^2)} \right) \pi\left(\sigma^2\right) \mathrm{d}\sigma^2$$

It is

$$\begin{aligned} f\left(y|\lambda,\sigma^2\right) &= \int f\left(y|\beta,\sigma^2\right)\pi\left(\beta,\sigma^2\right)\mathrm{d}\beta \\ &= (2\pi)^{-\frac{n+n+1}{2}}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\det\left(D\right)^{\frac{1}{2}}\int\exp\left(-\frac{1}{2\sigma^2}\left(\left(y-K\beta\right)^\top\left(y-K\beta\right)+\beta^\top\left(D\sigma^2\right)\beta\right)\right)\mathrm{d}\beta \\ &= (2\pi)^{-\frac{n+n+1}{2}}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\det\left(D\right)^{\frac{1}{2}}\left[\int\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mu^*\right)^\top V^*\left(\beta-\mu^*\right)\right)\mathrm{d}\beta\right]\left[\exp\left(-\frac{1}{2\sigma^2}S^*\right)\right] \end{aligned}$$

Because

$$\begin{aligned} \int\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mu^*\right)^\top V^*\left(\beta-\mu^*\right)\right)\mathrm{d}\beta &= (2\pi)^{\frac{n+1}{2}}\det\left(V^*/\sigma^2\right)^{-\frac{1}{2}} \\ &= (2\pi)^{\frac{n+1}{2}}\det\left(K^\top K+D\sigma^2\right)^{-\frac{1}{2}} \end{aligned}$$

$$\begin{aligned} \exp\left(-\frac{1}{2\sigma^2}S^*\right) &= \exp\left(-\frac{1}{2\sigma^2}\mu^\top\left(D\sigma^2\right)\mu-\left(\mu^*\right)^\top\left(V^*\right)^{-1}\left(\mu^*\right)+y^\top y\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(y^\top y-y^\top K\left(K^\top K+D\sigma^2\right)^{-1}K^\top y\right)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(y^\top\left(I-K\left(K^\top K+D\sigma^2\right)^{-1}K^\top\right)y\right)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(y^\top\left(K^\top D^{-1}K+I\sigma^2\right)^{-1}y\right)\right) \end{aligned}$$

So

$$\begin{aligned} f\left(y|\lambda,\sigma^2\right) = (2\pi)^{-\frac{n}{2}}&\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\det\left(D\right)^{\frac{1}{2}}\det\left(K^\top K+\sigma^2 D\right)^{-\frac{1}{2}} \\ &\times\exp\left(-\frac{1}{2\sigma^2}\left(y^\top\left(I\sigma^2+K^\top D^{-1}K\right)^{-1}y\right)\right) \end{aligned}$$

(1) I have

$$\begin{aligned} f\left(y|\sigma^2\right) &= \int f\left(y|\lambda,\sigma^2\right)\pi(\lambda)\mathrm{d}\lambda \\ &= (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{1}{2}}\int\left[\det\left(D\right)^{\frac{1}{2}}\right]\left[\det\left(K^\top K+D\sigma^2\right)^{-\frac{1}{2}}\right] \\ &\qquad\times\exp\left(-\frac{1}{2}\left(y^\top\left(I\sigma^2+K^\top D^{-1}K\right)^{-1}y\right)\right)\left[\prod_{i=0}^{n}\lambda_i^{a-1}\right]\mathrm{d}\lambda_0\ldots\mathrm{d}\lambda_n \end{aligned}$$

because $b=0$.

- It is $\exp\left(-\frac{y^\top y}{2\sigma^2}\right)\leq\exp\left(-\frac{y^\top\left(I\sigma^2+K^\top D^{-1}K\right)^{-1}y}{2}\right)$
- It is $\det\left(D\right)^{\frac{1}{2}}=\prod_{i=0}^{n}\lambda_i^{\frac{1}{2}}$.
- If $\{e_j\}_{j=0}^{n=1}$ are eigenvalues of $K^\top K$ and $e_{\max}=\max\left(\{e_j\}\right)$, then $K^\top K+D\sigma^2\leq Ie_{\max}+D\sigma^2$, consequently $\det\left(K^\top K+D\sigma^2\right)^{-\frac{1}{2}}\geq\prod_{j=0}^{n}\left(\lambda_j\sigma^2+e_{\max}\right)^{-\frac{1}{2}}$.

Then

$$f\left(y|\sigma^2\right) \geq (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{1}{2}} \int \prod_{j=0}^{n} \lambda_j^{\frac{1}{2}} \prod_{j=0}^{n} \left(\lambda_j \sigma^2 + e_{\max}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} y^\top y\right) \prod_{j=0}^{n} \lambda_j^{a-1} \mathrm{d}\lambda_0 \ldots \mathrm{d}\lambda_n$$

$$= (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} y^\top y\right) \int \ldots \int \prod_{j=0}^{n}\left[\lambda_j^{\frac{1}{2}}\right]\left[\prod_{j=0}^{n}\left(\lambda_j \sigma^2 + e_{\max}\right)^{-\frac{1}{2}}\right]\left[\prod_{j=0}^{n} \lambda_j^{a-1}\right] \mathrm{d}\lambda_0 \ldots \mathrm{d}\lambda_n$$

$$= (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} y^\top y\right) \prod_{j=0}^{n} \int \frac{\lambda_j^{a-\frac{1}{2}}}{\left(\lambda_j \sigma^2 + e_{\max}\right)^{\frac{1}{2}}} \mathrm{d}\lambda_j$$

Let $t_i = 1/\lambda_i$, then

$$f\left(y|\sigma^2\right) \geq (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} y^\top y\right) \prod_{j=0}^{n} \int \frac{t_{jj}^{-a-1}}{\left(t_j + \frac{\sigma^2}{e_{\max}}\right)^{\frac{1}{2}}} \mathrm{d}\lambda_j$$

which is finite if and only if $a \in (-1/2, 0)$.

(2)

(a) If $a > 0$, $b > 0$ then $\lambda_i \overset{\text{iid}}{\sim} \mathrm{Ga}\left(a, b\right)$ for all $i = 1, ..., n$, and if $c > 0$, $d > 0$ then $\tau \overset{\text{iid}}{\sim} \mathrm{Ga}\left(c, d\right)$ which are proper. So $\Pi\left(\beta, \sigma^2, \lambda, \tau\right)$ is a proper prior, and hence it leads to proper posterior.

(b) I have

$$f\left(y|\sigma^2\right) = (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{1}{2}} \int \left[\det\left(D\right)^{\frac{1}{2}}\right]\left[\det\left(K^\top K + D\sigma^2\right)^{-\frac{1}{2}}\right]$$
$$\times \exp\left(-\frac{1}{2}\left(y^\top\left(I\sigma^2 + K^\top D^{-1} K\right)^{-1} y\right)\right) \pi\left(\lambda\right) \mathrm{d}\lambda$$

It is $\det\left(D\right)^{\frac{1}{2}} = \prod_{i=0}^{n} \lambda_i^{\frac{1}{2}}$. Also, it is $K^\top K + D\sigma^2 \geq D\sigma^2$ then $\det\left(K^\top K + D\sigma^2\right)^{-\frac{1}{2}} \leq \prod_{j=0}^{n}\left(\lambda_j \sigma^2\right)^{-\frac{1}{2}}$. Hence

$$f\left(y|\sigma^2\right) \leq (2\pi)^{-\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} y^\top\left(I - P\right) y\right) \int \pi\left(\lambda\right) \mathrm{d}\lambda$$

which implies that $f\left(y|\sigma^2\right) < \infty$ if $\pi\left(\lambda\right)$ is proper. Yet,

$$f\left(y\right) = \int f\left(y|\sigma^2\right) \pi\left(\sigma^2\right) \mathrm{d}\sigma^2$$
$$\leq (2\pi)^{-\frac{n}{2}} \int \left(\sigma^2\right)^{-\frac{n}{2}+c+1} \exp\left(-\frac{1}{\sigma^2}\left(\frac{y^\top\left(I - P\right) y}{2} + d\right)\right) \mathrm{d}\sigma^2$$

which is finite if $y^\top\left(I - P\right) y + 2d > 0$ and $c > -\frac{n}{2}$.

(c) No. This implies $\pi\left(\lambda, \sigma^2\right) \propto \sigma^2 \prod_{j=0}^{n} \lambda_j^{-1}$. It is improper prior as $\int \pi\left(\lambda, \sigma^2\right) \mathrm{d}\left(\lambda, \sigma^2\right) = \infty$, and $(a, b, c, d) = (0, 0, 0, 0)$ which violates the necessary conditions.

(d) No, it violates the necessary conditions.

Exercise 18. ($\star$) Students are encouraged to practice on the Exercises 6.1-6.19 from the textbook *"Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.*" available from

- https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

The solutions are available from

- https://blackboard.durham.ac.uk/ultra/courses/_44662_1/outline/create/document?id=_1396738_1

Exercise 19. ($\star\star$) Show that the function K defined over $\mathbb{R}\times\mathbb{R}$ with expression

$$K(x,y)=\frac{\sin\left(2\pi\left(N+\frac{1}{2}\right)(x-y)\right)}{\sin\left(\pi(x-y)\right)}$$

is a valid kernel.

Hint-1: You may use that $\sum_{n=0}^{r}z^n=\frac{1-z^{r+1}}{1-z}$
Hint-2: You may use that $e^{ix}=\cos(x)+i\sin(x)$

Solution. It is

$$\begin{aligned}
K(x,y)=&\frac{\sin\left(2\pi\left(N+\frac{1}{2}\right)(x-y)\right)}{\sin\left(\pi(x-y)\right)}=\frac{-2i\sin\left(2\pi\left(N+\frac{1}{2}\right)(x-y)\right)}{-2i\sin\left(\pi(x-y)\right)}\\
=&\frac{e^{-2\pi\left(N+\frac{1}{2}\right)i(x-y)}-e^{2\pi\left(N+\frac{1}{2}\right)i(x-y)}}{e^{-\pi i(x-y)}-e^{\pi i(x-y)}}\\
=&\frac{e^{\pi i(x-y)}}{e^{\pi i(x-y)}}\frac{e^{-2\pi\left(N+\frac{1}{2}\right)i(x-y)}-e^{2\pi\left(N+\frac{1}{2}\right)i(x-y)}}{e^{-\pi i(x-y)}-e^{\pi i(x-y)}}\\
=&e^{-2\pi iN(x-y)}\frac{1-\left(e^{2\pi i(x-y)}\right)^{2N+1}}{1-e^{2\pi i(x-y)}}\\
=&e^{-2\pi iN(x-y)}\sum_{n=0}^{2N}\left(e^{2\pi i(x-y)}\right)^n=\sum_{n=-N}^{N}e^{2\pi in(x-y)}=\sum_{n=-N}^{N}e^{2\pi inx}e^{-2\pi iny}\\
=&\sum_{n=-N}^{N}e^{2\pi inx}\overline{e^{2\pi iny}}=\langle\psi(x),\psi(y)\rangle
\end{aligned}$$

with $\psi(x)=\left(e^{-2\pi iNx},e^{-2\pi i(N-1)x},...,1,...,e^{2\pi i(N-1)x},e^{2\pi iNx}\right)^\top$. Based on the theorem in the Handout, the Kernel can be expressed as an inner product of a vector of bases, hence it is a valid kernel.

Note that given Hint 2 it is

$$\begin{aligned}
e^{-2\pi iny}=&\cos(-2\pi ny)+i\sin(-2\pi ny)\\
=&\cos(2\pi ny)-i\sin(2\pi ny)\\
=&\overline{\cos(2\pi ny)+i\sin(2\pi ny)}\\
=&\overline{e^{2\pi iny}}
\end{aligned}$$

Exercise 20. ($\star\star$) (Kernel ridge regression) Consider the standard linear regression problem with learning rule $h_\theta(x)=\theta^\top x$. Consider a training data set $\mathcal{S}=\{z_i=(x_i,y_i)\}_{i=1}^m$. The ridge regression cost function is then

$$\hat{R}_{\mathcal{S}}(\theta)=\frac{1}{2}\sum_{i=1}^{m}\left(\theta^\top x_i-y_i\right)^2+\frac{\lambda}{2}\|\theta\|_2^2$$

(1) Use a vector notation to find a closed-form expression for the value of θ which minimizes the ridge regression cost function.

(2) Suppose that we want to use kernels to implicitly represent our feature vectors in a high-dimensional (possibly infinite dimensional) space. Using a feature mapping ψ, the ridge regression cost function becomes

$$\hat{R}_{\mathcal{S}}^{\psi}\left(\theta\right)=\frac{1}{2}\sum_{i=1}^{m}\left(\theta^{\top}\psi\left(x_{i}\right)-y_{i}\right)^{2}+\frac{\lambda}{2}\left\|\theta\right\|_{2}^{2}$$

Making a prediction on a new input x_{new} would now be done by computing $\theta^{\top}\psi\left(x_{\text{new}}\right)$. Show how we can use the "kernel trick" to obtain a closed form for the prediction on the new input x_{new} without ever explicitly computing $\psi\left(x_{\text{new}}\right)$. You may assume that the parameter vector θ can be expressed as a linear combination of the input feature vectors; i.e., $\theta=\sum_{i=1}^{m}\alpha_{i}\psi\left(x_{i}\right)$ for some set of parameters $\{\alpha_{i}\}$.

Hint:: You may need to use the identity

$$(\lambda I+BA)^{-1}B=B(\lambda I+AB)^{-1}.$$

Solution.

(1) It is

$$\hat{R}_{\mathcal{S}}\left(\theta\right)=\frac{1}{2}\left(X\theta-y\right)^{\top}\left(X\theta-y\right)+\frac{\lambda}{2}\theta^{\top}\theta$$

The gradient is

$$\nabla_{\theta}\hat{R}_{\mathcal{S}}\left(\theta\right)=X^{\top}X\theta-X^{\top}y+\lambda\theta$$

and by setting it equal to zero at $\hat{\theta}$, I get

$$\begin{aligned}0=\nabla_{\theta}\hat{R}_{\mathcal{S}}\left(\theta\right)\Big|_{\theta=\hat{\theta}}=&X^{\top}X\theta-X^{\top}y+\lambda\hat{\theta}\\ \Longrightarrow\ \hat{\theta}=&\left(X^{\top}X-\lambda I\right)^{-1}X^{\top}y\end{aligned}$$

(2) Now the design matrix is Ψ associated with the feature vectors $\psi\left(x_{i}\right)$ i.e. $\left[\Psi\right]_{i,j}=\psi_{j}\left(x_{i}\right)$. Then from the previous part, it is

$$\begin{aligned}\hat{\theta}=&\left(\Psi^{\top}\Psi-\lambda I\right)^{-1}\Psi^{\top}y\\ =&\Psi^{\top}\left(\Psi^{\top}\Psi-\lambda I\right)^{-1}y\\ =&\Psi^{\top}\left(K-\lambda I\right)^{-1}y\end{aligned}$$

where K is the Gram for the training set i.e. $[K]_{i,j} = k(x_i, x_j)$ there $k(\cdot,\cdot)$ is a kernel such as $k(a,b) = \psi(a)^\top \psi(b)$. To predict a new value $y_{\text{new}} = h_\theta(x_{\text{new}})$, we can compute

$$\begin{aligned}
y_{\text{new}} = h_{\hat{\theta}}(x_{\text{new}}) =& \hat{\theta}^\top \psi(x_{\text{new}}) \\
=& \left(\Psi^\top (K - \lambda I)^{-1} y \right)^\top \psi(x_{\text{new}}) \\
=& y^\top (K - \lambda I)^{-1} \Psi \psi(x_{\text{new}}) \\
=& \sum_{i=1}^{m} \alpha_i K \psi(x_i, x_{\text{new}})
\end{aligned}$$

where $\alpha_i = (K - \lambda I)^{-1} y$

Exercise 21. ($\star\star$) (On the SVM with Gaussian kernel) Consider the task of training a support vector machine using the Gaussian kernel

$$k(a,b) = \exp\left(-\frac{1}{\phi^2} \|a - b\|_2^2 \right)$$

We will show that as long as there are no two identical points in the training set, we can always find a value for the scale parameter $\phi > 0$ such that the SVM achieves zero training error. Consider a training data set $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^m$.

(1) Recall from class that the decision function learned by the support vector machine can be written as

$$h_w(x) = \sum_{i=1}^{m} \alpha_i y_i k(x_i, x) + b$$

Assume that the training data $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i=1}^m$ consists of points which are separated by at least a distance of ϵ that is $\|x_i - x_j\|_2 \geq \epsilon$ for any $i \neq j$. Find values for the set of parameters $\{\alpha_i\}$ and b and Gaussian kernel scale parameter ϕ such that x_i is correctly classified, for all $i = 1, ..., m$.

Hint:: Let $\alpha_i = 1$ for all i and $b = 0$. Now notice that for $y \in \{-1, +1\}$ the prediction on x_i will be correct if $|f(x_i) - y_i| < 1$, so find a value of ϕ that satisfies this inequality for all i.

(2) (Requires additional reading) Suppose we run a SVM with slack variables using the parameter $\phi > 0$ you found in the previous part. Consider that you are using the "Sequential minimal optimization" to solve the quadratic programming (QP) problem that arises during the training of support-vector machines (SVM) –for details see:

`https://en.wikipedia.org/wiki/Sequential_minimal_optimization`

Will the resulting classifier necessarily obtain zero training error? Why or why not? A short explanation (without proof) will suffice.

Solution.

(1) First we set $\alpha_i = 1$ for all $i = 1, ..., m$ and $b = 0$. Then, for a training example $\{x_i, y_i\}$, we get

$$
\begin{aligned}
|h_w(x_i) - y_i| &= \left| \sum_{j=1}^{m} y_j k(x_j, x_i) - y_i \right| \\
&= \left| \sum_{j=1}^{m} y_j \exp\left(-\frac{1}{\phi^2} \|x_j - x_i\|_2^2 \right) - y_i \right| \\
&= \left| y_i + \sum_{j \neq i} y_j \exp\left(-\frac{1}{\phi^2} \|x_j - x_i\|_2^2 \right) - y_i \right| \\
&\leq \sum_{j \neq i} \left| y_j \exp\left(-\frac{1}{\phi^2} \|x_j - x_i\|_2^2 \right) \right| \\
&= \sum_{j \neq i} |y_j| \exp\left(-\frac{1}{\phi^2} \|x_j - x_i\|_2^2 \right) \\
&\leq \sum_{j \neq i} |y_j| \exp\left(-\frac{1}{\phi^2} \epsilon^2 \right) \\
&= (m-1) \exp\left(-\frac{1}{\phi^2} \epsilon^2 \right)
\end{aligned}
$$

Thus we need to choose a γ such that

$$
(m-1) \exp\left(-\frac{1}{\phi^2} \epsilon^2 \right) < 1 \iff \phi < \frac{\epsilon}{\log(m-1)}
$$

E.g. $\phi = \frac{\epsilon}{\log(m)}$

(2) The classifier will obtain zero training error. The SVM without slack variables will always return zero training error if it is able to find a solution, so all that remains to be shown is that there exists at least one feasible point. Consider the constraint $y_i(w^\top x_i + b)$ for some i, and let $b = 0$. Then

$$
y_i \left(w^\top x_i + b \right) = y_i f(x_i) > 0
$$

since $f(x_i)$ and y_i have the same sign, and shown above. Therefore, as we choose all the α_i's large enough, $y_i \left(w^\top x_i + b \right) > 1$, so the optimization problem is feasible.

Exercise 22. ($\star\star$) (Kernel Principal Component Analysis) We will kernelize the classical PCA.

Classical Principal Component Analysis (PCA)

If x is a random vector with mean μ and covariance matrix Σ then the principal component transformation is the transformation

$$x \mapsto y = \Gamma^{\top} (x - \mu),$$

where Γ is orthogonal, $\Gamma^{\top}\Sigma\Gamma = \Lambda$ is diagonal $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$, $\lambda_1 \geq ... \geq \lambda_j \geq \lambda_{j+1} \geq ... \geq \lambda_p = 0$. The i-th principal component of x may be defined as the i-th element of the vector y, i.e.

$$y_i = \gamma_{(i)}^{\top} (x - \mu)$$

where $\gamma_{(i)}$ is the i-th column of Γ and it is called the i-th vector of principal component loadings. In other words, Γ and Λ contain the eigenvectors and eigenvalues from the eigen-decomposition of Σ.

Consider the classical PCA. Consider a given dataset in the form of a matrix X such as $[X]_{i,j} = x_{i,j}$ for $i = 1, ..., m$ and $j = 1, ..., p$. Let $x_i = (x_{i,1}, ..., x_{i,p})^{\top}$ for $i = 1, ..., m$. Assume the data are centered around zero as $\sum_{i=1}^{m} x_i = 0$. Hence, $\mu = \text{E}(x) = 0$. The sample covariance matrix is

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^{\top}$$

the ith PC is $y_i = \gamma_{(i)} x$ where $\gamma_{(i)}$ is the i-th column of Γ where $\Gamma^{\top}\Sigma\Gamma = \Lambda$ is diagonal $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$, $\lambda_1 \geq ... \geq \lambda_j \geq \lambda_{j+1} \geq ... \geq \lambda_p = 0$.

Suppose that we want to use kernels to implicitly represent our feature vectors in a high-dimensional (possibly infinite dimensional) space. Using a feature mapping $x \in \mathbb{R}^p \to \psi(x) \in \mathbb{H}$, the sample covariance matrix of the images, after mapping into $\mathbb{H}$ and assuming centered data, is given by

$$C = \frac{1}{m} \sum_{i=1}^{m} \psi(x_i) (\psi(x_i))^{\top}$$

Keep on assuming that the transformation keeps our data centered around zero; $\text{E}(\psi(x)) = 0$. Show how we can use the "kernel trick" to obtain a closed form for the principal component loadings, and the principal components without ever explicitly computing $\psi(\cdot)$.

Hint:: You may follow the steps:

(1) Show that for $n = 1, ..., m$ it is

$$\frac{1}{m} \sum_{j} \alpha_j \sum_{i=1}^{m} \left[(\psi(x_n))^{\top} \psi(x_i) \right] \left[(\psi(x_i))^{\top} \psi(x_j) \right] = \lambda \sum_{j=1}^{m} \alpha_j \left[(\psi(x_n))^{\top} \psi(x_j) \right]$$

(2) Show that this implies $K\alpha_k = m\lambda_k\alpha_k$ for some function $k(a, b) = (\psi(a))^{\top} \psi(b)$ and certain vectors α_k.

(3) Normalise α_k's properly

(4) Compute the principal components y_k

Solution.

(1) The Sample covariance matrix is

$$C=\frac{1}{m}\sum_{i=1}^{m}\psi\left(x_i\right)\left(\psi\left(x_i\right)\right)^{\top}$$

The goal is to perform the eigen-decomposition of C^{ψ} i.e.

$$Cv=\lambda v$$

so

$$\begin{aligned}C\gamma&=\lambda\gamma\\ \Longleftrightarrow\frac{1}{m}\sum_{i=1}^{m}\psi\left(x_i\right)\left(\psi\left(x_i\right)\right)^{\top}\gamma&=\lambda\gamma\\ \Longleftrightarrow\frac{1}{m}\sum_{i=1}^{m}\left[\left(\psi\left(x_i\right)\right)^{\top}\gamma\right]\psi\left(x_i\right)&=\lambda\gamma\end{aligned}$$

Since $\lambda\neq 0$, γ must be the span of $\psi\left(x_i\right)$ i.e. v_1 can be written as

$$\gamma=\sum_{j=1}^{m}\alpha_j\psi\left(x_j\right)$$

So by substituting, I get

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left(\psi\left(x_i\right)\right)^{\top}\left[\sum_{j=1}^{m}\alpha_j\psi\left(x_j\right)\right]\psi\left(x_i\right)&=\lambda\left[\sum_{j=1}^{m}\alpha_j\psi\left(x_j\right)\right]\\ \frac{1}{m}\sum_{j=1}^{m}\alpha_j\sum_{i=1}^{m}\left(\psi\left(x_j\right)\right)^{\top}\psi\left(x_i\right)\psi\left(x_i\right)&=\lambda\left[\sum_{j=1}^{m}\alpha_j\psi\left(x_j\right)\right]\end{aligned}$$

To get only inner products, I multiply both sides with $\psi\left(x_n\right)$ for any $n=1,...,m$, and I get

$$\frac{1}{m}\sum_{j}\alpha_j\sum_{i=1}^{m}\left[\left(\psi\left(x_n\right)\right)^{\top}\psi\left(x_i\right)\right]\left[\left(\psi\left(x_i\right)\right)^{\top}\psi\left(x_j\right)\right]=\lambda\sum_{j=1}^{m}\alpha_j\left[\left(\psi\left(x_n\right)\right)^{\top}\psi\left(x_j\right)\right]$$

(2) Now we can set/specify a kernel functions as

$$k\left(a,b\right)=\left(\psi\left(a\right)\right)^{\top}\psi\left(b\right)$$

and substitute the inner product with the above kernel as

$$\frac{1}{m}\sum_{j=1}^{m}\alpha_j\sum_{i=1}^{m}k\left(x_n,x_i\right)k\left(x_i,x_j\right)=\lambda\sum_{j=1}^{m}\alpha_jk\left(x_n,x_j\right)$$

which allow as to solve the problem without working directly in a high-dimensional space. Let K be a Gram matrix such as $\left[K\right]_{i,j}=k\left(x_i,x_j\right)$. Hence I get

$$\frac{1}{m}K^{\top}K\alpha=\lambda K\alpha\Longleftrightarrow KK\alpha=m\lambda K\alpha\overset{K^{-1}}{\Longleftrightarrow}K\alpha=m\lambda\alpha$$

Hence we have to simply solve the eigenvalue problem on the Gram matrix (Kernel matrix) K. Hence α and $m\lambda$ are eigen-vector and eigen-value of the Kernel matrix K. In a similar way we can find the rest eigen-vectors $\gamma_{(k)}$ and eigen-values $m\lambda_k$.

(3) If we consider v_k are normalised according to PCA, then for $k = 1, ..., m$ it is

$$\begin{aligned}1 = \left\langle\gamma_{(k)}, \gamma_{(k)}\right\rangle &= \left\langle\sum_{j=1}^{m} \alpha_{k,j}\psi\left(x_j\right), \sum_{i=1}^{m} \alpha_{k,i}\psi\left(x_i\right)\right\rangle \\ &= \sum_{j=1}^{m}\sum_{i=1}^{m} \alpha_{k,j}\alpha_{k,i}\left\langle\psi\left(x_j\right), \psi\left(x_i\right)\right\rangle \\ &= \sum_{j=1}^{m}\sum_{i=1}^{m} \alpha_{k,j}\alpha_{k,i}k\left(x_i, x_j\right) = a_k^\top K\alpha_k \\ &= m\lambda a_k^\top \alpha_k\end{aligned}$$

(4) Hence to extract the k-th principal component we simply take

$$y_k = \gamma_{(k)}^\top\psi\left(x\right) = \sum_{j=1}^{m} \alpha_{k,j}\left(\psi\left(x_j\right)\right)^\top\psi\left(x\right) = \sum_{j=1}^{m} \alpha_{k,j}k\left(x_j, x\right)$$

for each i.

Part 5. **Multi-class classification**

Part 6. Artificial Neural Networks

Part 7. Gaussian process regression

Part 8. Revision