Mathematics of Data Science

Xiaochuan Gong

October 16, 2022

Contents

1 Solutions to Homework 1 (February 28, 2022)

1.1 Problem 1 (Courant-Fischer theorem for singular values)

Suppose $A \in \mathbb{R}^{m\times n}$ has sigular values $\sigma_1, \ldots, \sigma_n$ ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. Then for $k = 1, \ldots, n$:

(a) $\sigma_k = \max\limits_{S\subseteq R^n, \dim(S)=k} \min\limits_{\mathbf{v}\in S, \|\mathbf{v}\|=1} \|A\mathbf{v}\|$.

(b) $\sigma_k = \min\limits_{S\subseteq R^n, \dim(S)=n-k+1} \max\limits_{\mathbf{v}\in S, \|\mathbf{v}\|=1} \|A\mathbf{v}\|$.

Proof. (a). Before starting the proof, we denote $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ orthonormal as the eigenvectors of $A^\top A$ corresponding to ${\sigma_1}^2, {\sigma_2}^2, \ldots, {\sigma_n}^2$ respectively.

First, let $W = \text{span}\{\mathbf{u}_k, \ldots, \mathbf{u}_n\}$, then $\dim(W) = n - k + 1$. So for any k-dimensional subspace S, we should have $\dim(W \cap S) \geq 1$. This is because of the equality

$$\dim(W \cup S) = \dim(W) + \dim(S) - \dim(W \cap S),$$

and of course $\dim(W \cup S) \leq n$, thus $\dim(W \cap S) \geq 1$. Now, choose any $\mathbf{v} \in W \cap S$ and $\|\mathbf{v}\| = 1$, note that $\mathbf{v} = \sum\limits_{j=k}^{n} \langle \mathbf{v}, \mathbf{u}_j\rangle \mathbf{u}_j$, $A^\top A\mathbf{u}_j = \sigma_j^2 \mathbf{u}_j$, and $\|A\mathbf{v}\|^2 = \langle A\mathbf{v}, A\mathbf{v}\rangle = \langle \mathbf{v}, A^\top A\mathbf{v}\rangle = \langle A^\top A\mathbf{v}, \mathbf{v}\rangle$, then it follows that

$$\begin{aligned}
\|A\mathbf{v}\|^2 &= \langle A^\top A\mathbf{v}, \mathbf{v}\rangle \\
&= \langle \sum_{j=k}^{n} \langle \mathbf{v}, \mathbf{u}_j\rangle \sigma_j^2 \mathbf{u}_j, \mathbf{v}\rangle \\
&= \sum_{j=k}^{n} \sigma_j^2 \langle \mathbf{v}, \mathbf{u}_j\rangle \langle \mathbf{u}_j, \mathbf{v}\rangle \\
&= \sum_{j=k}^{n} \sigma_j^2 |\langle \mathbf{v}, \mathbf{u}_j\rangle|^2 \\
&\overset{(i)}{\leq} \sigma_k^2.
\end{aligned}$$

Here in (i) we use the fact that $\sigma_k \geq \cdots \geq \sigma_n$, and that $\sum\limits_{j=k}^{n} |\langle \mathbf{v}, \mathbf{u}_j\rangle|^2 = \|\mathbf{v}\|^2 = 1$. Since $\|A\mathbf{v}\|^2 \leq \sigma_k^2$, then $\|A\mathbf{v}\| \leq \sigma_k$. Thus for any k-dimensional subspace S,

$$\inf_{\mathbf{v}\in S, \|\mathbf{v}\|=1} \|A\mathbf{v}\| \leq \sigma_k,$$

and since $\{\mathbf{v} \in S : \|\mathbf{v}\| = 1\}$ is compact, infimum is attained. So we have

$$\min_{\mathbf{v}\in S, \|\mathbf{v}\|=1} \|A\mathbf{v}\| \leq \sigma_k.$$

Since it holds for any k-dimensional subspace S, thus

$$\sup_{S\subseteq R^n, \dim(S)=k} \min_{\mathbf{v}\in S, \|\mathbf{v}\|=1} \|A\mathbf{v}\| \leq \sigma_k.$$

On the other hand, consider a particular k-dimensional subspace $S^* = \text{span}\{\mathbf{u}_1, \ldots \mathbf{u}_k\}$, then for any $\mathbf{v} \in S^*, \|\mathbf{v}\| = 1$,

$$\begin{aligned}
\|A\mathbf{v}\|^2 &= \sum_{j=1}^{k} \sigma_j^2 |\langle \mathbf{v}, \mathbf{u}_j\rangle|^2 \\
&\geq \sigma_k^2.
\end{aligned}$$

Hence, we obtain $\min_{\mathbf{v}\in S^*,\|\mathbf{v}\|=1} \|A\mathbf{v}\| \geq \sigma_k$. In particular, if we choose $\mathbf{v} = \mathbf{u}_k$, then $\|A\mathbf{v}\| = \|A\mathbf{u}_k\| = \sigma_k$, thus $\min_{\mathbf{v}\in S^*,\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \sigma_k$. Therefore, we have

$$\sup_{S\subseteq R^n,\dim(S)=k} \min_{\mathbf{v}\in S,\|\mathbf{v}\|=1} \|A\mathbf{v}\| \geq \sigma_k,$$

so we get

$$\sup_{S\subseteq R^n,\dim(S)=k} \min_{\mathbf{v}\in S,\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \sigma_k.$$

Note that for $S = S^*$, the supremum over all k-dimensional subspace S is attained, and finally we conclude that

$$\sigma_k = \max_{S\subseteq R^n,\dim(S)=k} \min_{\mathbf{v}\in S,\|\mathbf{v}\|=1} \|A\mathbf{v}\|.$$

(b). The proof of this formula follows the same idea as above and we shall omit the similar details. We first choose $W = \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, so $\dim(W) = k$. Then for any $(n-k+1)$-dimensional subspace S, $\dim(W \cap S) \geq 1$. Next, choose any $\mathbf{v} \in W \cap S$ and $\|\mathbf{v}\| = 1$, we obtain $\|A\mathbf{v}\| \geq \sigma_k$, and therefore we have

$$\max_{\mathbf{v}\in S,\|\mathbf{v}\|=1} \|A\mathbf{v}\| \geq \sigma_k.$$

Now, we again choose a particular $S^* = \text{span}\{\mathbf{u}_k, \ldots, \mathbf{u}_n\}$, and this gives

$$\inf_{S\subseteq R^n,\dim(S)=n-k+1} \max_{\mathbf{v}\in S,\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \sigma_k.$$

Note that for $S = S^*$, infimun is attained, and finally we conclude that

$$\sigma_k = \min_{S\subseteq R^n,\dim(S)=n-k+1} \max_{\mathbf{v}\in S,\|\mathbf{v}\|=1} \|A\mathbf{v}\|.$$

□

1.2 Problem 2 (Principal component analysis)

Suppose we have data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with each $\mathbf{x}_i \in \mathbb{R}^d$ and d is huge, find the optimal linear encoder $E \in \mathbb{R}^{s\times d}$ and linear decoder $D \in \mathbb{R}^{d\times s}$ with $s \ll d$ that minimize the reconstruction loss

$$\mathcal{L}(E, D) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - DE\mathbf{x}_i\|^2.$$

Proof. We divide the proof into three parts. For the first part, we'll show that the columns of optimal linear decoder D are orthonormal and $E = D^\top$; for the second part, we'll find an equivalent equation to optimize; for the third part, we'll demonstrate that the solution to the PCA problem is to set D to be the matrix whose columns are $\mathbf{u}_1, \ldots, \mathbf{u}_s$ and to set $E = D^\top$, where $\mathbf{u}_1, \ldots, \mathbf{u}_s$ are eigenvectors of the matrix $A \coloneqq \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ corresponding to the largest s eigenvalues of A.

Fix any D, E and denote $R = \{DE\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\}$, then R is an s-dimensional linear subspace of $\mathbb{R}^d$. Let $V \in \mathbb{R}^{d\times s}$ be a matrix whose columns form an orthonormal basis of this subspace, namely, the range of V is R and $V^\top V = I_s$. Therefore, each vector in R can be written as $V\mathbf{y}$ where $\mathbf{y} \in \mathbb{R}^s$. For every $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^s$ we have

$$\begin{aligned}\|\mathbf{x} - V\mathbf{y}\|^2 &= \mathbf{x}^\top\mathbf{x} + \mathbf{y}^\top V^\top V\mathbf{y} - 2\mathbf{y}^\top V^\top\mathbf{x} \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{y}^\top(V^\top\mathbf{x}).\end{aligned}$$

Let $\frac{\partial \|\mathbf{x}-V\mathbf{y}\|^2}{\partial \mathbf{y}} = 0$, we obtain $\mathbf{y} = V^\top \mathbf{x}$. Therefore, for each $\mathbf{x}$ we have that

$$VV^\top \mathbf{z} = \arg\min_{\mathbf{z}\in R} \|\mathbf{x}-\mathbf{z}\|^2.$$

In particular this holds for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and therefore we can replace D, E by $V, V^\top$ and by that do not increase the objective

$$\sum_{i=1}^{n} \|\mathbf{x}_i - DE\mathbf{x}_i\|^2 \geq \sum_{i=1}^{n} \|\mathbf{x}_i - VV^\top \mathbf{x}_i\|^2.$$

Since this holds for every D, E, it follows that the columns of optimal linear decoder D are orthonormal and $E = D^\top$.

On the basis of the preceding proof, we can rewrite the optimization problem as follows:

$$\arg\min_{D\in\mathbb{R}^{d\times s}: D^\top D = I_s} \sum_{i=1}^{n} \|\mathbf{x}_i - DD^\top \mathbf{x}_i\|^2.$$

Then for every $\mathbf{x} \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|\mathbf{x} - DD^\top \mathbf{x}\|^2 &= \mathbf{x}^\top \mathbf{x} - 2\mathbf{x}^\top D^\top D\mathbf{x} + \mathbf{x}^\top DD^\top DD^\top \mathbf{x} \\
&= \|\mathbf{x}\|^2 - \mathbf{x}^\top DD^\top \mathbf{x} \\
&= \|\mathbf{x}\|^2 - \mathrm{tr}(\mathbf{x}^\top DD^\top \mathbf{x}) \\
&= \|\mathbf{x}\|^2 - \mathrm{tr}(D^\top \mathbf{x}\mathbf{x}^\top D).
\end{aligned}$$

Since trace is a linear operator, therefore

$$\sum_{i=1}^{n} \|\mathbf{x}_i - DD^\top \mathbf{x}_i\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \mathrm{tr}(D^\top \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top D).$$

Then it allows us to rewrite the problem as follows:

$$\arg\max_{D\in\mathbb{R}^{d\times s}: D^\top D = I_s} \mathrm{tr}(D^\top \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top D).$$

Let $A = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$, $\mathbf{u}_1, \ldots, \mathbf{u}_s$ be eigenvectors of the matrix A corresponding to the largest s eigenvalues of A, and $P\Lambda P^\top$ be the spectral decomposition of A. In fact, $P\Lambda P^\top$ is also the SVD of A since A is symmetric. Fix some $D \in \mathbb{R}^{d\times s}$ with orthonormal columns and let $B = P^\top D$. Then, $PB = PP^\top D = D$, $D^\top AD = B^\top P^\top P\Lambda P^\top PB = B^\top \Lambda B$, thus we have

$$\mathrm{tr}(D^\top AD) = \sum_{j=1}^{d} \Lambda_{j,j} \sum_{i=1}^{s} {B_{j,i}}^2.$$

Note that $B^\top B = D^\top PP^\top D = I_s$. Therefore, the columns of B are orthonormal, then $\sum_{j=1}^{d}\sum_{i=1}^{s} {B_{j,i}}^2 = s$. In addition, let $\tilde{B} \in \mathbb{R}^{d\times d}$ be a matrix such that its first s columns are the columns of B and that $\tilde{B}^\top \tilde{B} = I_d$. Then, for every j we have $\sum_{i=1}^{d} \tilde{B}_{j,i}^2 = 1$, which implies that $\sum_{i=1}^{s} {B_{j,i}}^2 = \sum_{i=1}^{s} \tilde{B}_{j,i}^2 \leq 1$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, it follows that

$$\mathrm{tr}(D^\top AD) \leq \max_{\boldsymbol{\beta}\in[0,1]^d: \|\boldsymbol{\beta}\|_1 = s} \sum_{j=1}^{d} \Lambda_{j,j}\beta_j.$$

We can verify the right-hand side equals $\sum_{j=1}^{s} \Lambda_{j,j}$. Thus for every $D \in \mathbb{R}^{d\times s}$, $\mathrm{tr}(D^\top AD) \leq \sum_{j=1}^{s} \Lambda_{j,j}$. In particular, if we set D to be the matrix whose columns sre the s leading eigenvectors of A, we obtain that $\mathrm{tr}(D^\top AD) = \sum_{j=1}^{s} \Lambda_{j,j}$. In other words, $D = (\mathbf{u}_1, \ldots, \mathbf{u}_s)$ enables $\mathrm{tr}(D^\top AD)$ to attain the upper bound, and this concludes our proof. $\square$

2 Solutions to Homework 2 (March 7, 2022)

2.1 Problem 1 (Equivalence of closed functions and lower-semicontinous functions)

For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ the following statements are equivalent:

(a) f is closed.

(b) The sublevel set C_α is closed for any α.

(c) f is lower-semicontinous over $\mathbb{R}^n$.

Proof. (a)$\Longrightarrow$(b). Let α be any scalar and consider C_α. If $C_\alpha = \emptyset$, then C_α is closed. Suppose now that $C_\alpha \neq \emptyset$. Pick $\{\mathbf{x}_k\} \subset C_\alpha$ such that $\mathbf{x}_k \to \bar{\mathbf{x}}$ for some $\bar{\mathbf{x}} \in \mathbb{R}^n$. We have $f(\mathbf{x}_k) \leq \alpha$ for all α, implying that $(\mathbf{x}_k, \alpha) \in \text{epi}(f)$ for all α. Since $(\mathbf{x}_k, \alpha) \to (\bar{\mathbf{x}}, \alpha)$ and $\text{epi}(f)$ is closed, it follows that $(\bar{\mathbf{x}}, \alpha) \in \text{epi}(f)$. Consequently $f(\bar{\mathbf{x}}) \leq \alpha$, showing that $\bar{\mathbf{x}} \in C_\alpha$. Hence C_α is closed for any α.

(b)$\Longrightarrow$(c). Let $\mathbf{x}_0 \in \mathbb{R}^n$ be arbitrary and let $\{\mathbf{x}_k\}$ be a sequence such that $\mathbf{x}_k \to \mathbf{x}_0$. To arrive at a contradiction, assume that f is not lower-semicontinous at $\mathbf{x}_0$, i.e.,

$$\liminf_{k\to\infty} f(\mathbf{x}_k) < f(\mathbf{x}_0).$$

Then, there exist a scalar β and a subsequence $\{\mathbf{x}_{k_j}\} \subset \{\mathbf{x}_k\}$ such that

$$f(\mathbf{x}_{k_j}) \leq \beta < f(\mathbf{x}_0),$$

yielding that $\{\mathbf{x}_{k_j}\} \subset C_\beta$. Since $\mathbf{x}_k \to \mathbf{x}_0$, then $\mathbf{x}_{k_j} \to \mathbf{x}_0$. Note that the set C_β is closed, it follows that $\mathbf{x}_0 \in C_\beta$. Hence, $f(\mathbf{x}_0) \leq \beta$, a controdiction arrives. Thus, we must have

$$f(\mathbf{x}_0) \leq \liminf_{k\to\infty} f(\mathbf{x}_k).$$

(c)$\Longrightarrow$(a). To arrive at a controdiction we assume that $\text{epi}(f)$ is not closed. Then, there exists a sequence $\{(\mathbf{x}_k, t_k)\} \subset \text{epi}(f)$ such that

$$(\mathbf{x}_k, t_k) \to (\bar{\mathbf{x}}, \bar{t}) \quad \text{and} \quad (\bar{\mathbf{x}}, \bar{t}) \notin \text{epi}(f).$$

Since $(\mathbf{x}_k, t_k) \in \text{epi}(f)$ for all k, we have

$$f(\mathbf{x}_k) \leq t_k, \forall k.$$

Taking the limit inferior as $k \to \infty$, and using $t_k \to \bar{t}$, we obtain

$$\liminf_{k\to\infty} f(\mathbf{x}_k) \leq \lim_{k\to\infty} t_k = \bar{t}.$$

Since $(\bar{\mathbf{x}}, \bar{t}) \notin \text{epi}(f)$, we have $f(\bar{\mathbf{x}}) > \bar{t}$, implying that

$$\liminf_{k\to\infty} f(\mathbf{x}_k) \leq \bar{t} < f(\bar{\mathbf{x}}).$$

On the other hand, because $\mathbf{x}_k \to \bar{\mathbf{x}}$, and f is lower-semicontinous at $\bar{\mathbf{x}}$, we have

$$\liminf_{k\to\infty} f(\mathbf{x}_k) \geq f(\bar{\mathbf{x}}),$$

a controdiction arrives. Hence, $\text{epi}(f)$ must be closed. □

3 Solutions to Homework 3 (March 14, 2022)

3.1 Problem 1 (Subdifferential of ℓ_1-norm)

Compute the subdifferential ∂f with $f(\mathbf{x}) = \|\mathbf{x}\|_1$.

Proof. The ℓ_1-norm

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

can be expressed as summation of convex functions, i.e., $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$, where $f_i(\mathbf{x}) = |x_i| = \operatorname{sgn}(x_i) \cdot x_i$. Therefore, we use the summation formula to compute the subdifferential. It follows that

$$\begin{aligned}
\partial f(\mathbf{x}) &= \sum_{i=1}^{n} \partial f_i(\mathbf{x}) \\
&= \sum_{i \in \{j \,:\, x_j \neq 0\}} \operatorname{sgn}(x_i) \cdot e_i + \sum_{i \in \{j \,:\, x_j = 0\}} [-e_i, e_i] \\
&= \left\{ g = (g_1, \ldots, g_n)^\top : g_i = \operatorname{sgn}(x_i) \text{ if } x_i \neq 0, \text{ and } [-1, 1] \text{ otherwise} \right\}.
\end{aligned}$$

□

3.2 Problem 2 (Subdifferential of ℓ_2-norm)

Compute the subdifferential ∂f with $f(\mathbf{x}) = \|\mathbf{x}\|_2$.

Proof. The ℓ_2-norm

$$f(\mathbf{x}) = \|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$$

is convex and differentiable away from $\mathbf{0}$, therefore,

$$\partial f(\mathbf{x}) = \nabla f(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \quad \forall \mathbf{x} \neq \mathbf{0}.$$

Now we compute the subdifferential at $\mathbf{x} = \mathbf{0}$. For any $\mathbf{g}$ and $\|\mathbf{g}\|_2 \leq 1$, by Cauchy-Schwarz inequality we have

$$\mathbf{g}^\top (\mathbf{x} - \mathbf{0}) \leq \|\mathbf{g}\|_2 \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2 - \mathbf{0},$$

therefore we obtain

$$\{\mathbf{g} : \|\mathbf{g}\|_2 \leq 1\} \subseteq \partial f(\mathbf{0}).$$

Next we will show that $\mathbf{g} \notin \partial f(\mathbf{0})$ if $\|\mathbf{g}\|_2 > 1$. Let $\mathbf{x} = \mathbf{g}$, if $\mathbf{g}$ is a subgradient, then

$$\|\mathbf{g}\|_2 - \mathbf{0} \geq \mathbf{g}^\top (\mathbf{g} - \mathbf{0}) = \|\mathbf{g}\|_2^2 > \|\mathbf{g}\|_2,$$

thus a controdiction arrives. Therefore $\partial f(\mathbf{0}) = \{\mathbf{g} : \|\mathbf{g}\|_2 \leq 1\}$. Hence,

$$\partial f(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} & \mathbf{x} \neq \mathbf{0}, \\ \{\mathbf{g} : \|\mathbf{g}\|_2 \leq 1\} & \mathbf{x} = \mathbf{0}. \end{cases}$$

□

3.3 Problem 3 (Subdifferential of ℓ_∞-norm)

Compute the subdifferential ∂f with $f(\mathbf{x}) = \|\mathbf{x}\|_\infty$.

Proof. The ℓ_∞-norm

$$f(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$$

can be expressed as maximum of convex functions, i.e., $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$, where $f_i(\mathbf{x}) = |x_i|$. Using

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

and pointwise maximum formula for computing subdifferential, we have

$$\partial f(\mathbf{x}) = \mathbf{conv}\left(\cup_{i \in I(\mathbf{x})}\{\partial f_i(\mathbf{x})\}\right) = \left(\sum_{i \in I(\mathbf{x})} \theta_i \cdot \partial f_i(\mathbf{x}) : \sum_{i \in I(\mathbf{x})} \theta_i = 1,\, \theta_i \geq 0,\, i \in I(\mathbf{x})\right),$$

where $I(\mathbf{x}) = \{i : f_i(\mathbf{x}) = f(\mathbf{x})\}$. □

3.4 Problem 4 (Subdifferential of nuclear norm)

Compute the subdifferential ∂f with $f(A) = \|A\|_*$, where $\|A\|_*$ is the nuclear norm of matrix A.

Before starting the proof, we first give the following two lemmas, which characterize the nuclear norm and the subdifferential of a norm in an inner product space respectively.

Lemma 3.1. *Nuclear norm is the dual norm of the operator norm, i.e.,*

$$\|X\|_* = \sup_{\|Z\|_2 \leq 1} \langle X, Z \rangle. \tag{3.1}$$

Proof of Lemma 3.1. Let $X \in \mathbb{R}^{m \times n}$ be a matrix of rank r, and let $X = U\Sigma V^\top$ denote the SVD of X, where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r) \in \mathbb{R}^{m \times r}, \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$ and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n \times r}$. Let $Z = UV^\top$, note that

$$\|Z\|_2 = \sqrt{\lambda_{\max}\left(Z^\top Z\right)} = \sqrt{\lambda_{\max}\left(VU^\top UV^\top\right)} = \sqrt{\lambda_{\max}\left(I_n\right)} = 1,$$

so $\|Z\|_2 \leq 1$ holds. Therefore,

$$\begin{aligned} \sup_{\|Z\|_2 \leq 1} \langle X, Z \rangle &\geq \langle X, Z \rangle \\ &= \mathrm{tr}\left(Z^\top X\right) \\ &= \mathrm{tr}\left(VU^\top U\Sigma V^\top\right) \\ &= \mathrm{tr}\left(V\Sigma V^\top\right) \\ &= \mathrm{tr}\left(V^\top V\Sigma\right) \\ &= \mathrm{tr}\left(\Sigma\right) \\ &= \sum_{i=1}^{r} \sigma_i(X) = \|X\|_*. \end{aligned}$$

Thus we have

$$\sup_{\|Z\|_2 \leq 1} \langle X, Z \rangle \geq \|X\|_*. \tag{3.2}$$

For another direction, we have

$$\begin{aligned}
\langle X, Z\rangle &= \operatorname{tr}\left(Z^\top X\right) \\
&= \operatorname{tr}\left(Z^\top U\Sigma V^\top\right) \\
&= \operatorname{tr}\left(V^\top Z^\top U\Sigma\right) \\
&= \operatorname{tr}\left(\left(U^\top ZV\right)^\top \Sigma\right) \\
&= \langle U^\top ZV, \Sigma\rangle \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r}\left[U^\top ZV\right]_{ij}\Sigma_{ij} \\
&= \sum_{i=1}^{r}\left[U^\top ZV\right]_{ii}\sigma_i \\
&= \sum_{i=1}^{r}\mathbf{u}_i^\top Z\mathbf{v}_i\sigma_i
\end{aligned} \tag{3.3}$$

To continue, we need

$$\mathbf{u}_i^\top Z\mathbf{v}_i \le \max_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}\mathbf{u}^\top Z\mathbf{v} = \sigma_1(Z) = \|Z\|_2,$$

where we use the equivalent definition for the operator norm. Continue from equality (3.2), take supremum on both sides, this yields

$$\begin{aligned}
\sup_{\|Z\|_2\le 1}\langle X, Z\rangle &= \sup_{\|Z\|_2\le 1}\sum_{i=1}^{r}\mathbf{u}_i^\top Z\mathbf{v}_i\sigma_i \\
&= \sup_{\|Z\|_2\le 1}\sum_{i=1}^{r}\|Z\|_2\cdot\sigma_i \\
&\le \sum_{i=1}^{r}\sigma_i \\
&= \|X\|_*.
\end{aligned}$$

Thus we have

$$\sup_{\|Z\|_2\le 1}\langle X, Z\rangle \le \|X\|_*. \tag{3.4}$$

Combine (3.2) and (3.4), this implies that

$$\|X\|_* = \sup_{\|Z\|_2\le 1}\langle X, Z\rangle.$$

□

Lemma 3.2. *Let $\|\cdot\|$ be an arbitrary norm, then we have*

$$\partial\|\mathbf{x}\| = \{\mathbf{v} \mid \langle\mathbf{v}, \mathbf{x}\rangle = \|\mathbf{x}\|, \|\mathbf{v}\|_* \le 1\}, \tag{3.5}$$

where $\|\mathbf{x}\|_* \coloneqq \sup_{\|\mathbf{z}\|\le 1}\langle\mathbf{x}, \mathbf{z}\rangle$ is the dual norm to $\|\cdot\|$.

Remark. *Note that here $\|\cdot\|_*$ denotes the dual norm to $\|\cdot\|$ instead of the nuclear norm.*

Proof of Lemma 3.2. To do this, we will prove two directions. First, define the set

$$\mathcal{G}(\mathbf{x}) \coloneqq \{\mathbf{v} \mid \langle\mathbf{v}, \mathbf{x}\rangle = \|\mathbf{x}\|, \|\mathbf{v}\|_* \le 1\}.$$

We first show that if $\mathbf{v} \in \mathcal{G}(\mathbf{x})$, then $\mathbf{v} \in \partial\|\mathbf{x}\|$, showing that $\mathcal{G}(\mathbf{x}) \subset \partial\|\mathbf{x}\|$. Let $\mathbf{v} \in \partial\|\mathbf{x}\|$, then for any $\mathbf{y}$, we have

$$\|\mathbf{x}\| + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle = \|\mathbf{x}\| + \langle \mathbf{v}, \mathbf{y} \rangle - \langle \mathbf{v}, \mathbf{x} \rangle = \langle \mathbf{v}, \mathbf{y} \rangle \overset{(i)}{\leq} \|\mathbf{v}\|_* \|\mathbf{y}\| \leq \|\mathbf{y}\|, \tag{3.6}$$

where in (i) we use Holder's inequality which states $\|\langle \mathbf{x}, \mathbf{y} \rangle\| \leq \|\mathbf{x}\| \|\mathbf{y}\|_*$ for any dual pair norms. Since (3.6) holds for all $\mathbf{y}$ and therefore $\mathbf{v} \in \partial\|\mathbf{x}\|$. Hence we obtain $\mathcal{G}(\mathbf{x}) \subset \partial\|\mathbf{x}\|$.

For another direction, we first introduce the definition of the Frenchel conjugate of a function. Given a real valued function $f(\mathbf{x})$ on an inner product space, define the conjugate $f^*(\mathbf{y})$ of f as

$$f^*(\mathbf{y}) = \sup_{\mathbf{z}} \left\{ \langle \mathbf{y}, \mathbf{z} \rangle - f(\mathbf{z}) \right\}.$$

It turns out that the Fenchel conjugate of a norm is just the indicator function on the unit ball of the dual norm, that is,

$$\|\mathbf{y}\|^* = \begin{cases} 0 & \|\mathbf{y}\|_* \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Equipped with this, we are ready to proceed. Let $\mathbf{v} \in \partial\|\mathbf{x}\|$, then for any $\mathbf{y}$, we have

$$\|\mathbf{y}\| \geq \|\mathbf{x}\| + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle \iff \langle \mathbf{v}, \mathbf{y} \rangle - \|\mathbf{y}\| \leq \langle \mathbf{v}, \mathbf{x} \rangle - \|\mathbf{x}\|. \tag{3.7}$$

Since (3.7) holds for all $\mathbf{y}$, we take the supremum over all $\mathbf{y}$'s on both sides,

$$\sup_{\mathbf{y}} \left\{ \langle \mathbf{v}, \mathbf{y} \rangle - \|\mathbf{y}\| \right\} \leq \langle \mathbf{v}, \mathbf{x} \rangle - \|\mathbf{x}\|.$$

Notice the left-hand side is simply $\|\mathbf{v}\|^*$, and therefore

$$\begin{cases} 0 & \|\mathbf{y}\|_* \leq 1, \\ +\infty & \text{otherwise.} \end{cases} \leq \langle \mathbf{v}, \mathbf{x} \rangle - \|\mathbf{x}\|.$$

If $\|\mathbf{v}\|_* > 1$, then this cannot possibly hold since the right-hand side will always be finite for a fixed $\mathbf{v}$. Thus we have $\|\mathbf{v}\|_* \leq 1$, which implies

$$0 \leq \langle \mathbf{v}, \mathbf{x} \rangle - \|\mathbf{x}\| \overset{(i)}{\leq} \|\mathbf{v}\|_* \|\mathbf{x}\| - \|\mathbf{x}\| \leq 0,$$

where in (i) we again use Holder's inequality. Therefore, all inequalities above are strict, which yields $\langle \mathbf{v}, \mathbf{x} \rangle = \|\mathbf{x}\|$. This means $\mathbf{v} \in \mathcal{G}(\mathbf{x})$, which shows that $\partial\|\mathbf{x}\| \subset \mathcal{G}(\mathbf{x})$. Hence we obtain $\partial\|\mathbf{x}\| = \mathcal{G}(\mathbf{x})$, which yields the desired result. □

Now we use the above two lemmas to derive the subdifferential of the nuclear norm.

Proof. Let $A \in \mathbb{R}^{m \times n}$ be a matrix of rank r, and let $X = U\Sigma V^\top$ denote the SVD of X, where $U \in \mathbb{R}^{m \times r}, \Sigma \in \mathbb{R}^{r \times r}$ and $V \in \mathbb{R}^{n \times r}$. By Lemma 3.1, the dual norm of $\|A\|_*$ is $\|A\|_2$, and by Lemma 3.2 we get

$$\partial\|A\|_* = \left\{ Z \mid \langle Z, A \rangle = \|A\|_*, \|Z\|_2 \leq 1 \right\}.$$

□

Remark. *In fact, there are some other equivalent characterizations of the nuclear norm. Please check Dual certificates for nuclear norm minimization and Understanding the uniqueness of the solution of the nuclear norm minimization for more detailed proofs. We list those equivalent characterizations as below:*

(1) $\partial\|A\|_* = \{Z \mid \langle Z, A \rangle = \|A\|_*, \|Z\|_2 \leq 1\}$.

(2) $\partial\|A\|_* = \left\{ UV^\top + W \mid U^\top W = 0, WV = 0, \|W\|_2 \leq 1, W \in \mathbb{R}^{m \times n} \right\}$.

(3) $\partial\|A\|_* = \left\{ Z \mid \Pi_{\mathcal{T}}(Z) = UV^\top, \|\Pi_{\mathcal{T}^\perp}(Z)\|_2 \leq 1 \right\}$, *where* $\mathcal{T}$ *is the linear space defined as* $\mathcal{T} = \left\{ UX^\top + YV^\top, X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{m \times r} \right\}$.

4 Solutions to Homework 4 (April 11, 2022)

4.1 Problem 1 (Norm of matrices with sub-gaussian entries)

Let A be an $m \times n$ random matrix, and $A \sim \text{sub } G_{m\times n}(\sigma^2)$. Then, for any t, we have

$$\|A\|_2 \leq C\sigma(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$.

Before starting the proof, we first prove the following lemma.

Lemma 4.1. *Let A be an $n \times n$ matrix and $\epsilon \in [0, \frac{1}{2})$. Show that for any ϵ-net $\mathcal{N}$ of the sphere S^{n-1} and any ϵ-net $\mathcal{M}$ of the sphere S^{m-1}, we have*

$$\|A\|_2 \leq \frac{1}{1-2\epsilon} \cdot \sup_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

Proof of Lemma 4.1. To prove the upper bound, fix $\mathbf{x} \in S^{n-1}$ and $\mathbf{y} \in S^{m-1}$ such that

$$\|A\|_2 = \sup_{\mathbf{u}\in S^{n-1},\, \mathbf{v}\in S^{m-1}} \langle A\mathbf{u}, \mathbf{v}\rangle = \langle A\mathbf{x}, \mathbf{y}\rangle.$$

Choose $\mathbf{x}_0 \in \mathcal{N}$ and $\mathbf{y}_0 \in \mathcal{M}$ so that

$$\|\mathbf{x} - \mathbf{x}_0\|_2 \leq \epsilon,\ \|\mathbf{y} - \mathbf{y}_0\|_2 \leq \epsilon.$$

By Cauchy-Schwarz inequality and the definition of spectral norm, we have the following:

$$\begin{aligned}
\langle A\mathbf{x}, \mathbf{y}\rangle - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle &= \langle A\mathbf{x}, \mathbf{y} - \mathbf{y}_0\rangle + \langle A(\mathbf{x} - \mathbf{x}_0), \mathbf{y}_0\rangle \\
&\leq \|A\mathbf{x}\|_2\|\mathbf{y} - \mathbf{y}_0\|_2 + \|A(\mathbf{x} - \mathbf{x}_0)\|_2\|\mathbf{y}_0\|_2 \\
&\leq \|A\|_2\|\mathbf{x}\|_2\|\mathbf{y} - \mathbf{y}_0\|_2 + \|A\|_2\|\mathbf{y}_0\|_2\|\mathbf{x} - \mathbf{x}_0\|_2 \\
&\leq 2\epsilon\|A\|_2.
\end{aligned}$$

Note that $\|A\|_2 = \langle A\mathbf{x}, \mathbf{y}\rangle$, then we obtain $(1 - 2\epsilon)\|A\|_2 \leq \langle A\mathbf{x}_0, \mathbf{y}_0\rangle$. Dividing both sides of this inequality by $1 - 2\epsilon$ and take supremum over $\mathbf{x}_0 \in \mathcal{N}$ and $\mathbf{y}_0 \in \mathcal{M}$, we have

$$\|A\|_2 \leq \frac{1}{1-2\epsilon} \cdot \langle A\mathbf{x}_0, \mathbf{y}_0\rangle \leq \frac{1}{1-2\epsilon} \cdot \sup_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

□

Now we start to prove the problem.

Proof. This proof is an example of an ϵ-net argument. We need to control $\langle A\mathbf{x}, \mathbf{y}\rangle$ for all vectors $\mathbf{x}$ and $\mathbf{y}$ on the unit sphere. To this end, we will discretize the sphere using a net (approximation step), establish a tight control of $\langle A\mathbf{x}, \mathbf{y}\rangle$ for fixed vectors $\mathbf{x}$ and $\mathbf{y}$ from the net (concentration step), and finish by taking a union bound over all $\mathbf{x}$ and $\mathbf{y}$ in the net (union bound step).

Approximation step

Choose $\epsilon = \frac{1}{4}$, using the inequality

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(S^{n-1}, \|\cdot\|_2, \epsilon) \leq \left(\frac{2}{\epsilon} + 1\right)^n$$

for covering numbers of the unit Euclidean sphere S^{n-1}, we can find an ϵ-net $\mathcal{N}$ of the sphere S^{n-1} and ϵ-net $\mathcal{M}$ of the sphere S^{m-1} with cardinalities

$$|\mathcal{N}| \leq 9^n \quad \text{and} \quad |\mathcal{M}| \leq 9^m. \tag{4.1}$$

Recall from Lemma 4.1 that the spectral norm can be bounded using these nets as follows:

$$\|A\|_2 \leq 2 \max_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle. \tag{4.2}$$

Concentration step

Fix $\mathbf{x} \in \mathcal{N}$ and $\mathbf{y} \in \mathcal{M}$. Since $A \sim \text{sub } G_{m\times n}(\sigma^2)$, by definition, $\langle A\mathbf{x}, \mathbf{y}\rangle \sim \text{sub } G(\sigma^2)$. Then for any $u \geq 0$, we have the following tail bound:

$$\mathbb{P}\{\langle A\mathbf{x}, \mathbf{y}\rangle \geq u\} \leq \exp(-\frac{u^2}{2\sigma^2}). \tag{4.3}$$

Union bound step

Next, we unfix $\mathbf{x}$ and $\mathbf{y}$ using a union bound. Suppose the event $\left\{ \max_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq u \right\}$ occurs. Then there exist $\mathbf{x} \in \mathcal{N}$ and $\mathbf{y} \in \mathcal{M}$ such that $\langle A\mathbf{x}, \mathbf{y}\rangle \geq u$. Thus the union bound yields

$$\mathbb{P}\left\{ \max_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq u \right\} \leq \sum_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \mathbb{P}\left\{\langle A\mathbf{x}, \mathbf{y}\rangle \geq u\right\}$$

Using the tail bound (4.3) and the estimate (4.1) on the sizes of $\mathcal{N}$ and $\mathcal{M}$, we can bound the above probability by

$$9^{m+n} \cdot \exp(-\frac{u^2}{2\sigma^2}) \tag{4.4}$$

Choose $u = C\sigma(\sqrt{m} + \sqrt{n} + t)/2$, then $u^2 \geq C^2\sigma^2(m + n + t^2)/4$, and if the constant C is chosen sufficiently large, say

$$\frac{u^2}{2\sigma^2} \geq \frac{C^2\sigma^2(m + n + t^2)}{8\sigma^2} \geq 3(m + n) + t^2 - \ln 2.$$

Thus we can obtain

$$\mathbb{P}\left\{ \max_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq u \right\} \leq 9^{m+n} \cdot \exp(-3(m + n) - t^2 + \ln 2) \leq 2\, \exp(-t^2).$$

Finally, combining this with (4.2), we conclude that

$$\mathbb{P}\left\{\|A\|_2 \geq 2u\right\} \leq \mathbb{P}\left\{ \max_{\mathbf{x}\in\mathcal{N},\, \mathbf{y}\in\mathcal{M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq u \right\} \leq 2\, \exp(-t^2).$$

Hence for any $t > 0$, we have $\|A\|_2 \leq C\sigma(\sqrt{m} + \sqrt{n} + t)$ with probability at least $1 - 2\, \exp(-t^2)$. $\square$

4.2 Problem 2 (Johnson-Lindenstrauss lemma)

For any $\epsilon \in (0, \frac{1}{2})$ and integer $m > 4$, let $k = \frac{20 \log m}{\epsilon^2}$. Then for any set $V \subset \mathbb{R}^N$ of m points, there exists a mapping $f : \mathbb{R}^N \to \mathbb{R}^k$ such that for all $\mathbf{u}, \mathbf{v} \in V$,

$$(1 - \epsilon)\|\mathbf{u} - \mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1 + \epsilon)\|\mathbf{u} - \mathbf{v}\|^2.$$

Before starting the proof, we first introduce two lemmas which will be used in the proof of the Johnson-Lindenstrauss lemma.

Lemma 4.2. *Let Q be a random variable following a χ^2-squared distribution with k degrees of freedom. Then for any $\epsilon \in (0, \frac{1}{2})$, the following inequality holds:*

$$\mathbb{P}[(1-\epsilon)k \leq Q \leq (1+\epsilon)k] \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.$$

Proof of Lemma 4.2. By Markov's inequality, we can write

$$\begin{aligned}\mathbb{P}[Q \geq (1+\epsilon)k] = \mathbb{P}[\exp(\lambda Q) \geq \exp(\lambda(1+\epsilon)k)] &\leq \frac{\mathbb{E}[\exp(\lambda Q)]}{\exp(\lambda(1+\epsilon)k)} \\ &= \frac{(1-2\lambda)^{-k/2}}{\exp(\lambda(1+\epsilon)k)},\end{aligned}$$

where we used for the final inequality the expression of the moment-generating function of a χ^2-squared distribution, $\mathbb{E}[\exp(\lambda Q)]$, for $\lambda < \frac{1}{2}$. Choosing $\lambda = \frac{\epsilon}{2(1+\epsilon)} < \frac{1}{2}$, which minimize the right-hand side of the final equality, and using the inequality $1 + \epsilon \leq \exp(\epsilon - (\epsilon^2 - \epsilon^3)/2)$ yield

$$\mathbb{P}[Q \geq (1+\epsilon)k] \leq \left(\frac{1+\epsilon}{\exp(\epsilon)}\right)^{k/2} \leq \left(\frac{\exp(\epsilon - \frac{\epsilon^2-\epsilon^2}{2})}{\exp(\epsilon)}\right)^{k/2} = \exp\left(-\frac{k}{4}(\epsilon^2-\epsilon^3)\right).$$

By using similar techniques we can derive that

$$\mathbb{P}[Q \leq (1-\epsilon)k] \leq \exp\left(-\frac{k}{4}(\epsilon^2-\epsilon^3)\right).$$

Then the statement of the lemma follows by applying the union bound

$$\begin{aligned}\mathbb{P}[(1-\epsilon)k \leq Q \leq (1+\epsilon)k] &= 1 - \mathbb{P}[Q \leq (1-\epsilon)k] - \mathbb{P}[Q \geq (1+\epsilon)k] \\ &\geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.\end{aligned}$$

□

Lemma 4.3. *Let $\mathbf{x} \in \mathbb{R}^N$, define $k < N$ and assume that entries in $A \in \mathbb{R}^{k\times N}$ are sampled independently from the standard normal distribution, $N(0,1)$. Then, for any $\epsilon \in (0, \frac{1}{2})$, we have*

$$\mathbb{P}\left[(1-\epsilon)\|\mathbf{x}\|^2 \leq \|\frac{1}{\sqrt{k}}A\mathbf{x}\|^2 \leq (1+\epsilon)\|\mathbf{x}\|^2\right] \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.$$

Proof of Lemma 4.3. Let $\hat{\mathbf{x}} = A\mathbf{x}$ and observe that

$$\mathbb{E}[\hat{x}_j^2] = \mathbb{E}\left[\left(\sum_{i=1}^N A_{ji}x_i\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^N A_{ji}^2 x_i^2\right] = \sum_{i=1}^N x_i^2 = \|\mathbf{x}\|^2.$$

The second and the third equalities follow from the independence and unit variance, respectively, of the A_{ij}. Now define $T_j = \hat{x_j}/\|\mathbf{x}\|$ and note that the T_js are independently standard normal random variables since the A_{ij} are i.i.d standard normal random variables and $\mathbb{E}[\hat{x}_j^2] = \|\mathbf{x}\|^2$. Thus, the variable Q defined by $Q = \sum_{j=1}^k T_j^2$ follows a χ^2-squared distribution with k degrees of freedom and we have

$$\begin{aligned}\mathbb{P}\left[(1-\epsilon)\|\mathbf{x}\|^2 \leq \frac{\|\hat{\mathbf{x}}\|^2}{k} \leq (1+\epsilon)\|\mathbf{x}\|^2\right] &= \mathbb{P}\left[(1-\epsilon)k \leq \sum_{j=1}^k T_j^2 \leq (1+\epsilon)k\right] \\ &= \mathbb{P}\Big[(1-\epsilon)k \leq Q \leq (1+\epsilon)k\Big] \\ &\geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4},\end{aligned}$$

where the final inequality holds by Lemma 4.2, thus proving the statement of the Lemma 4.3. □

Now we start to prove the Johnson-Lindenstrauss lemma.

Proof. Let $f = \frac{1}{\sqrt{k}}A$ where $k < N$ and entries in $A \in \mathbb{R}^{k\times N}$ are sampled independently from the standard normal distribution, $N(0,1)$. For fixed $\mathbf{u}, \mathbf{v} \in V$, we can apply Lemma 4.2, with $\mathbf{x} = \mathbf{u}-\mathbf{v}$, to lower bound the success probability by $1-2e^{-(\epsilon^2-\epsilon^3)k/4}$. Note that there are $O(m^2)$ pairs of $\mathbf{u}, \mathbf{v} \in V$, applying the union bound over those $O(m^2)$ pairs in V, setting $k = \frac{20\log m}{\epsilon^2}$ and $0 < \epsilon < \frac{1}{2}$, we have

$$\begin{aligned}
&\mathbb{P}\left[\exists \mathbf{u}, \mathbf{v} \text{ s.t. the following event fails: } (1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \le \|f(\mathbf{u})-f(\mathbf{v})\|^2 \le (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2\right] \\
&\le \sum_{\mathbf{u},\mathbf{v}\in V} \mathbb{P}\left[\text{s.t. the following event fails: } (1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \le \|f(\mathbf{u})-f(\mathbf{v})\|^2 \le (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2\right] \\
&\le 2m^2 e^{-(\epsilon^2-\epsilon^3)k/4} = 2m^{(5\epsilon-3)} \le 2m^{-\frac{1}{2}} < 1.
\end{aligned}$$

Therefore, for all $\mathbf{u}, \mathbf{v} \in V$, we have $\mathbb{P}[\text{success}] > 0$. Since the success probability is strictly greater than zero, a mapping that satisfies the desired conditions must exist, thus proving the statement of the lemma. □

5 Solutions to Homework 5 (April 18, 2022)

5.1 Problem 1 (Metric entropy for Lipschitz functions on the unit interval)

Consider the class of Lipschitz functions on the interval $[0,1]$ defined as

$$\mathcal{F}_L \coloneqq \{f : [0,1] \to \mathbb{R} \mid f(0) = 0 \text{ and } |f(x) - f(x')| \le L|x - x'| \quad \forall x, x' \in [0,1]\},$$

equipped with the metric associated with the sup-norm

$$d(f, f') \coloneqq \sup_{x \in [0,1]} |f(x) - f(x')|.$$

Prove that the metric entropy of the class $\mathcal{F}_L$ with respect to the sup-norm scales as

$$\log_2 \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) \asymp \frac{L}{\epsilon} \quad \text{for suitable small } \epsilon > 0,$$

where $\asymp$ denotes

$$\log_2 \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) = O\left(\frac{L}{\epsilon}\right) \quad \text{and} \quad \frac{L}{\epsilon} = O\left(\log_2 \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon)\right).$$

Proof. A typical method to this problem is to find an ϵ-covering and a 2ϵ-packing of the class $\mathcal{F}_L$ consisting of the same number of elements K_ϵ. Then we have

$$\mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) \le K_\epsilon \quad \text{and} \quad \mathcal{P}(\mathcal{F}_L, \|\cdot\|_\infty, 2\epsilon) \ge K_\epsilon,$$

and hence by applying the inequality

$$\mathcal{P}(\mathcal{F}_L, \|\cdot\|_\infty, 2\epsilon) \le \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon),$$

we obtain

$$\mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) = \mathcal{P}(\mathcal{F}_L, \|\cdot\|_\infty, 2\epsilon) = K_\epsilon.$$

ϵ-covering step

We need to construct an ϵ-covering of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$. To do so, defining $n \coloneqq \lfloor \frac{1}{\epsilon} \rfloor$, we divide the interval $[0,1]$ into $n+1$ segments $I_k = [x_{k-1}, x_k]$ for $k = 1, 2, \ldots, n+1$ with

$$x_k = k\epsilon, \quad \text{for } k = 0, 1, \ldots, n, \quad \text{and} \quad x_{n+1} = 1.$$

Moreover, we define the function $\Phi : \mathbb{R} \to \mathbb{R}$ via

$$\Phi(u) \coloneqq \begin{cases} 0 & \text{for } u < 0, \\ u & \text{for } 0 \le u \le 1, \\ 1 & \text{otherwise.} \end{cases}$$

For each binary sequence $\boldsymbol{\beta} \in \{-1, +1\}^n$, we may define a function $f_{\boldsymbol{\beta}} : [0,1] \to [-L, L]$ via

$$f_{\boldsymbol{\beta}}(y) = \sum_{k=1}^{n} \beta_k L\epsilon \Phi\left(\frac{y - x_k}{\epsilon}\right).$$

By construction, each function is piecewise linear and continuous, with slope either $+L$ or $-L$ over each of the intervals I_k for $k = 2, \ldots, n+1$, and constant on the remaining interval I_1, see Figure 1 for an illustration.

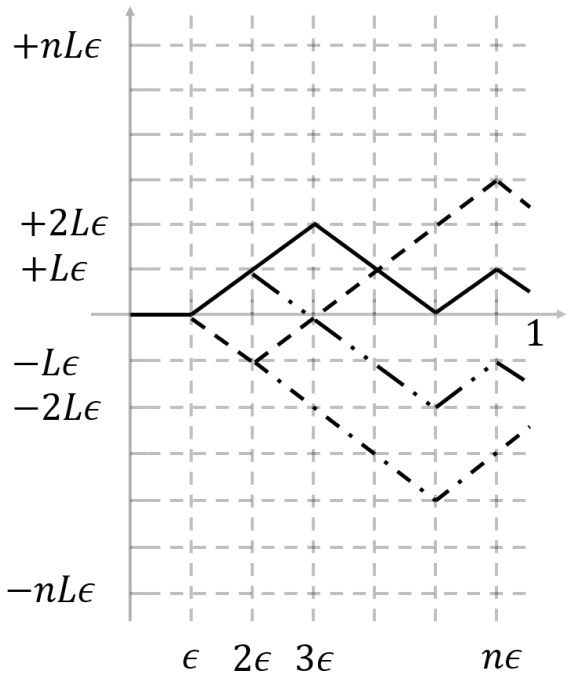

Figure 1: The function class $\{f_{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \in \{-1,+1\}^n\}$ used to construct a covering of the class $\mathcal{F}_L$.

We first show that for any choice of $\boldsymbol{\beta}$, $f_{\boldsymbol{\beta}} \in \mathcal{F}_L$. It's obvious that $f_{\boldsymbol{\beta}}(0) = 0$. Without loss of generality, we let $y \in (x_p, x_{p+1}]$ and $y' \in (x_q, x_{q+1}]$ and $n \geq p > q \geq 0$. By definition of $f_{\boldsymbol{\beta}}$, we have

$$\begin{aligned}
|f_{\boldsymbol{\beta}}(y) - f_{\boldsymbol{\beta}}(y')| &= \left| \sum_{k=1}^{p} \beta_k L\epsilon\Phi\left(\frac{y - x_k}{\epsilon}\right) - \sum_{k=1}^{q} \beta_k L\epsilon\Phi\left(\frac{y - x_k}{\epsilon}\right) \right| \\
&= L\epsilon \left| \sum_{k=q+1}^{p-1} \beta_k + \beta_p \left(\frac{y - x_p}{\epsilon}\right) + \beta_q \left(1 - \frac{y' - x_q}{\epsilon}\right) \right| \\
&\leq L\epsilon \left(\sum_{k=q+1}^{p-1} |\beta_k| + |\beta_p| \left(\frac{y - x_p}{\epsilon}\right) + |\beta_q| \left(1 - \frac{y' - x_q}{\epsilon}\right) \right) \\
&= L\epsilon \left(p - q + \left(\frac{y - x_p}{\epsilon}\right) - \left(\frac{y' - x_q}{\epsilon}\right) \right) \\
&= L\epsilon \left(p - q - \left(\frac{x_p - x_q}{\epsilon}\right) + \left(\frac{y - y'}{\epsilon}\right) \right) \\
&= L|y - y'|.
\end{aligned}$$

Hence for all $\boldsymbol{\beta}$, $f_{\boldsymbol{\beta}} \in \mathcal{F}_L$. Next we will prove by induction on $K = 1, \ldots, n+1$ that for every $f \in \mathcal{F}_L$ there is a $\boldsymbol{\beta}$ such that

$$\sup_{x \in \bigcup_{k=1}^{K} I_k} |f(x) - f_{\boldsymbol{\beta}}(x)| \leq L\epsilon.$$

For $K = 1$, for any choice of $\boldsymbol{\beta}$, we have $f_{\boldsymbol{\beta}}(x) = 0$, $\forall x \in [0, \epsilon]$. By L-Lipschitz assumption, for

every $f \in \mathcal{F}_L$, we have

$$|f(x) - f_{\boldsymbol{\beta}}(x)| = |f(x)| = |f(x) - f(0)| \leq L|x - 0| \leq L\epsilon,$$

which proves the property for $K = 1$.

Assuming the property is satisfied up to some rank K, we consider the rank $K+1$. Given a function $f \in \mathcal{F}_L$, by induction hypothesis, let $\boldsymbol{\beta}$ be such that

$$\sup_{x \in \bigcup_{k=1}^{K} I_k} |f(x) - f_{\boldsymbol{\beta}}(x)| \leq L\epsilon, \ \forall x \in [0, x_K].$$

By L-Lipschitz assumption, $|f(x_{K+1}) - f(x_K)| \leq L|x_{K+1} - x_K| = L\epsilon$. Then we have

$$f_{\boldsymbol{\beta}}(x_K) - 2L\epsilon \leq f(x_K) - L\epsilon \leq f(x_{K+1}) \leq f(x_K) + L\epsilon \leq f_{\boldsymbol{\beta}}(x_K) + 2L\epsilon.$$

Thus, either $f(x_{K+1}) \in [f_{\boldsymbol{\beta}}(x_K) - 2L\epsilon, f_{\boldsymbol{\beta}}(x_K)]$ or $f(x_{K+1}) \in [f_{\boldsymbol{\beta}}(x_K), f_{\boldsymbol{\beta}}(x_K) + 2L\epsilon]$. We treat only the first case, the method being the same for the second one. We take $\boldsymbol{\beta}$ such that the K-th term is -1 (also note that we take $\boldsymbol{\beta}$ such that the K-th term is 1 for the second case). Assume that we have some $x \in [x_K, x_{K+1}]$ such that $f(x) < f_{\boldsymbol{\beta}}(x) - L\epsilon$, then

$$\begin{aligned} f(x_K) - f(x) &> (f_{\boldsymbol{\beta}}(x_K) - L\epsilon) - (f_{\boldsymbol{\beta}}(x) - L\epsilon) \\ &= f_{\boldsymbol{\beta}}(x_K) - f_{\boldsymbol{\beta}}(x) \\ &= L(x - x_K), \end{aligned}$$

which is a contradiction with the L-Lipschitz assumption on f. On the contrary, if there is some $x \in [x_K, x_{K+1}]$ such that $f(x) > f_{\boldsymbol{\beta}}(x) + L\epsilon$, using $f(x_{K+1}) \leq f_{\boldsymbol{\beta}}(x_K)$, then

$$\begin{aligned} f(x) - f(x_{K+1}) &> (f_{\boldsymbol{\beta}}(x) + L\epsilon) - f_{\boldsymbol{\beta}}(x_K) \\ &= (f_{\boldsymbol{\beta}}(x) + L\epsilon) - (f_{\boldsymbol{\beta}}(x_{K+1}) + L\epsilon) \\ &= f_{\boldsymbol{\beta}}(x) - f_{\boldsymbol{\beta}}(x_{K+1}) \\ &= L(x_{K+1} - x), \end{aligned}$$

which is a contradiction with the L-Lipschitz assumption on f. Thus we have proven by induction that for all $K = 1, \ldots, n+1$, there is a $\boldsymbol{\beta}$ such that such that $f_{\boldsymbol{\beta}}$ is $L\epsilon$-close to a given f on $[0, x_K]$. In particular, set $K = n + 1$ and we conclude that the set of functions $\{f_{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \in \{-1, +1\}^n\}$ is an $L\epsilon$-covering of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$. Substituting $\frac{\epsilon}{L}$ for ϵ, then we get an ϵ-covering of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$.

2ϵ-packing step

We need to construct a 2ϵ-packing of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$. To do so, for each binary sequence $\boldsymbol{\beta} \in \{-1, +1\}^n$, we may define a function $h_{\boldsymbol{\beta}} : [0, 1] \to [-L, L]$ via

$$h_{\boldsymbol{\beta}}(y) = \sum_{k=1}^{n} \beta_k L\epsilon \Phi\left(\frac{y - x_{k-1}}{\epsilon}\right).$$

By construction, each function is piecewise linear and continuous, with slope either $+L$ or $-L$ over each of the intervals I_k for $k = 1, \ldots, n$, and constant on the remaining interval I_{n+1}, see Figure 2 for an illustration.

By using the same method as in the ϵ-covering step, we can verify that $h_{\boldsymbol{\beta}}(0) = 0$ and that $h_{\boldsymbol{\beta}} \in \mathcal{F}_L$. Given a pair of distinct strings $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ and the two associated functions $h_{\boldsymbol{\beta}_1}$ and $h_{\boldsymbol{\beta}_2}$, there is at least one interval I_k, with $1 \leq k \leq n$, where the functions start at the same point, and have opposite

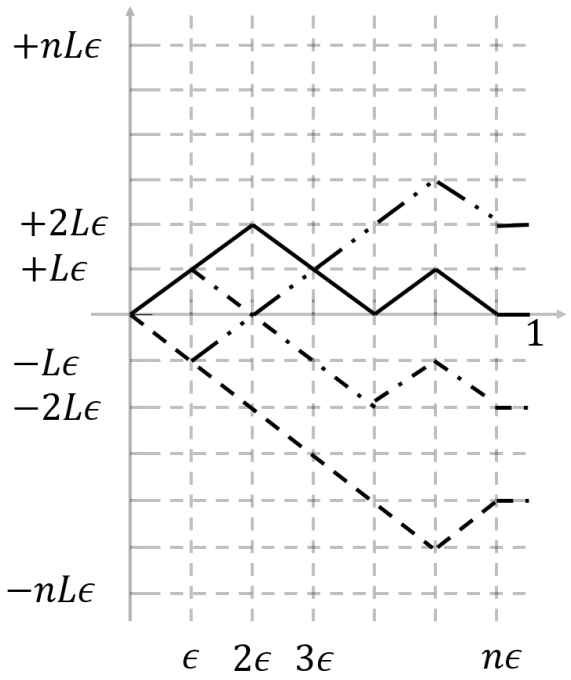

Figure 2: The function class $\{h_{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \in \{-1,+1\}^n\}$ used to construct a packing of the class $\mathcal{F}_L$.

slopes over I_k. Since the functions have slope $+L$ and $-L$ over I_k, respectively, we are guaranteed that $\|h_{\boldsymbol{\beta}_1} - h_{\boldsymbol{\beta}_2}\|_\infty \geq 2L\epsilon$, showing that the set $\{h_{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \in \{-1,+1\}^n\}$ forms a $2L\epsilon$-packing of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$. Substituting $\frac{\epsilon}{L}$ for ϵ, then we get a 2ϵ-packing of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$.

We have constructed an ϵ-covering and a 2ϵ-packing of the metric space $(\mathcal{F}_L, \|\cdot\|_\infty)$, both of cardinality $K_\epsilon = 2^n = 2^{\lfloor \frac{L}{\epsilon} \rfloor}$. Therefore, using the following inequality

$$K_\epsilon \leq \mathcal{P}(\mathcal{F}_L, \|\cdot\|_\infty, 2\epsilon) \leq \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) \leq K_\epsilon,$$

then we obtain

$$\log_2 \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \epsilon) \asymp \frac{L}{\epsilon}.$$

□

6 Solutions to Homework 6 (April 25, 2022)

6.1 Problem 1 (Conditional distributions of the multivariate normal distribution)

Let $X = \begin{pmatrix} Y \\ Z \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{pmatrix} \succcurlyeq 0$, then $Z|Y \sim \mathcal{N}(\mu_{z|y}, \Sigma_{z|y})$ with

$$\begin{aligned} \mu_{z|y} &= \mu_z + \Sigma_{zy}\Sigma_{yy}^{-1}(y - \mu_y), \\ \Sigma_{z|y} &= \Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}. \end{aligned} \tag{6.1}$$

Proof. We first assume that Y is an $n_1 \times 1$ vector, Z is an $n_2 \times 1$ vector and X is an $n \times 1$ vector, where $n = n_1 + n_2$. By construction, the joint distribution of Y and Z is $\mathcal{N}(\mu, \Sigma)$. Moreover, the marginal distribution of Y is $\mathcal{N}(\mu_y, \Sigma_{yy})$. According to the law of conditional probability, it holds that

$$p(z|y) = \frac{p(z, y)}{p(y)}.$$

Since we have

$$p(z|y) = \frac{\mathcal{N}(x; \mu, \Sigma)}{\mathcal{N}(y; \mu_y, \Sigma_{yy})},$$

and then we use the probability density function of the multivariate normal distribution, this yields

$$\begin{aligned} p(z|y) &= \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]}{\frac{1}{\sqrt{(2\pi)^{n_1} |\Sigma_{yy}|}} \cdot \exp\left[-\frac{1}{2}(y-\mu_y)^\top \Sigma_{yy}^{-1}(y-\mu_y)\right]} \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_1}}} \cdot \sqrt{\frac{|\Sigma_{yy}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \frac{1}{2}(y-\mu_y)^\top \Sigma_{yy}^{-1}(y-\mu_y)\right]. \end{aligned} \tag{6.2}$$

Write the inverse of Σ as

$$\Sigma^{-1} = \begin{bmatrix} (\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy})^{-1} & -(\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy})^{-1}\Sigma_{yz}\Sigma_{zz}^{-1} \\ -\Sigma_{zz}^{-1}\Sigma_{zy}(\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy})^{-1} & \Sigma_{zz}^{-1} + \Sigma_{zz}^{-1}\Sigma_{zy}(\Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy})^{-1}\Sigma_{yz}\Sigma_{zz}^{-1} \end{bmatrix} \triangleq \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix}$$

and plug this into (6.2), and we have

$$\begin{aligned} p(z|y) &= \frac{1}{\sqrt{(2\pi)^{n-n_1}}} \cdot \sqrt{\frac{|\Sigma_{yy}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \frac{1}{2}(y-\mu_y)^\top \Sigma_{yy}^{-1}(y-\mu_y)\right] \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_1}}} \cdot \sqrt{\frac{|\Sigma_{yy}|}{|\Sigma|}} \cdot \exp\Big\{ -\frac{1}{2}\big[(y-\mu_y)^\top \Sigma^{11}(y-\mu_y) + 2(y-\mu_y)^\top \Sigma^{12}(z-\mu_z) \\ &\quad + (z-\mu_z)^\top \Sigma^{22}(z-\mu_z)\big] + \frac{1}{2}(y-\mu_y)^\top \Sigma_{yy}^{-1}(y-\mu_y)\Big\} \\ &= \frac{1}{\sqrt{(2\pi)^{n-n_1}}} \cdot \sqrt{\frac{|\Sigma_{yy}|}{|\Sigma|}} \cdot \exp\Big\{ -\frac{1}{2}\left[z - (\mu_z + \Sigma_{zy}^\top \Sigma_{yy}^{-1}(y-\mu_y))\right]^\top (\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz})^{-1} \\ &\quad \left[z - (\mu_z + \Sigma_{zy}^\top \Sigma_{yy}^{-1}(y-\mu_y))\right]\Big\}, \end{aligned}$$

where we use the fact that $\Sigma_{zy}^\top = \Sigma_{yz}$ and that ${\Sigma^{21}}^\top = \Sigma^{12}$ since both Σ and Σ^{-1} are symmetric matrices. Note that the determinant of Σ is

$$|\Sigma| = \begin{vmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{vmatrix} = |\Sigma_{yy}| \cdot |\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}|,$$

with this and $n = n_1 + n_2$, we finally arrive at

$$p(z|y) = \frac{1}{\sqrt{(2\pi)^{n_2}|\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}|}} \cdot \exp\Big\{ -\frac{1}{2}\left[z - (\mu_z + \Sigma_{zy}^{\top}\Sigma_{yy}^{-1}(y - \mu_y))\right]^{\top} (\Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz})^{-1} \left[z - (\mu_z + \Sigma_{zy}^{\top}\Sigma_{yy}^{-1}(y - \mu_y))\right] \Big\},$$

which is exactly the density function of multivariate normal distribution $\mathcal{N}(\mu_{z|y}, \Sigma_{z|y})$. □

7 Solutions to Homework 7 (May 9, 2022)

7.1 Problem 1 (Property of gradient descent for strongly convex objectives)

Assume that f is L-smooth and m-strongly convex, prove that the gradient descent (GD) algorithm

$$x_{k+1} = x_k - \alpha_k \nabla f(w_k)$$

with fixed step-size $\alpha_k = \alpha = \frac{2}{L+m}$ satisfies

$$f(x_k) - f(z) + \frac{m}{2}\|x_k - z\|^2 \leq \left(\frac{L-m}{L+m}\right)^k \left(f(x_0) - f(z) + \frac{m}{2}\|x_0 - z\|^2\right), \quad \forall z \in \mathbb{R}^d.$$

Proof. First, since f is L-smooth, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{7.1}$$

Let $y = x - \frac{2}{L+m}\nabla f(x)$ and plug y in (7.1), we get

$$f(y) \leq f(x) - \frac{2m}{(L+m)^2}\|\nabla f(x)\|^2. \tag{7.2}$$

By strong convexity of f, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{7.3}$$

Therefore, for any $z \in \mathbb{R}^d$, it follows that

$$\begin{aligned}
\frac{m}{2}\|y - z\|^2 &= \frac{m}{2}\|x - \frac{2}{L+m}\nabla f(x) - z\|^2 \\
&= \frac{m}{2}\|x - z\|^2 - \frac{2m}{L+m}\langle \nabla f(x), x - z\rangle + \frac{2m}{(L+m)^2}\|\nabla f(x)\|^2 \\
&\overset{(i)}{\leq} \frac{m}{2}\|x - z\|^2 + \frac{2m}{m+L}\left(f(z) - f(x) - \frac{m}{2}\|z - x\|^2\right) + \frac{2m}{(L+m)^2}\cdot\frac{(L+m)^2}{2m}(f(x) - f(y)) \\
&= \frac{m(L-m)}{2(L+m)}\|x - z\|^2 - \frac{2}{L+m}(f(x) - f(z)) + f(x) - f(y),
\end{aligned}$$

where in (i) we use (7.2) and (7.3) to derive the inequality. Then, adding $f(y) - f(z)$ to both sides of the above inequality yields

$$f(y) - f(z) + \frac{m}{2}\|y - z\|^2 \leq \left(\frac{L-m}{L+m}\right)\left(f(x) - f(z) + \frac{m}{2}\|x - z\|^2\right), \quad \forall z \in \mathbb{R}^d.$$

Set $x = x_{i-1}$ and $y = x_i$ for $i = 1, \ldots, k$, then by recursion we conclude that

$$f(x_k) - f(z) + \frac{m}{2}\|x_k - z\|^2 \leq \left(\frac{L-m}{L+m}\right)^k \left(f(x_0) - f(z) + \frac{m}{2}\|x_0 - z\|^2\right), \quad \forall z \in \mathbb{R}^d.$$

□

7.2 Problem 2 (Lyapunov analysis of gradient descent dynamic)

Assume that f is L-smooth and m-strongly convex, prove the following convergence rates:

(a) For gradient flow $\frac{d}{dt}X_t = -\nabla f(X_t)$, let $\mathcal{E}_t = e^{\frac{2mL}{m+L}t}\frac{1}{2}\|x^* - X_t\|^2$ be the Lyapunov function, then we have $f(X_t) - f(x^*) \leq O(\frac{L}{2}e^{-\frac{2mL}{m+L}t})$.

(b) For gradient descent $\frac{x_{k+1}-x_k}{\delta} = -\nabla f(x_k)$ with $0 < \delta \leq \frac{2}{m+L}$, let $E_k = \left(1 - \frac{2mL}{m+L}\delta\right)^{-k}\frac{1}{2}\|x^* - x_k\|^2$ be the Lyapunov function, then we have $f(x_k) - f(x^*) \leq O(\frac{L}{2}e^{-\frac{2mL}{m+L}\delta k})$.

Proof. (a). First, since f is L-smooth and f is m-strongly convex ($m \leq L$), we have

$$\langle \nabla f(X_t), x^* - X_t\rangle \leq -\frac{mL}{m+L}\|x^* - X_t\|^2 - \frac{1}{m+L}\|\nabla f(X_t)\|^2. \tag{7.4}$$

We provide a proof of this bound in section 7.4. Using (7.4), it follows that

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}_t &= e^{\frac{2mL}{m+L}t}\left(\frac{mL}{m+L}\|x^* - X_t\|^2 - \langle \frac{d}{dt}X_t, x^* - X_t\rangle\right)\\
&= e^{\frac{2mL}{m+L}t}\left(\frac{mL}{m+L}\|x^* - X_t\|^2 + \langle \nabla f(X_t), x^* - X_t\rangle\right)\\
&= e^{\frac{2mL}{m+L}t}\left(\frac{mL}{m+L}\|x^* - X_t\|^2 - \frac{mL}{m+L}\|x^* - X_t\|^2 - \frac{1}{m+L}\|\nabla f(X_t)\|^2\right)\\
&\leq 0.
\end{aligned}$$

By integral, we obtain the statement

$$\mathcal{E}_t - \mathcal{E}_0 = \int_0^t \frac{d}{dt}\mathcal{E}_s ds \leq 0.$$

Thus we have

$$\frac{1}{2}\|x^* - X_t\|^2 \leq e^{-\frac{2mL}{m+L}t}\mathcal{E}_0.$$

In addition, by the property of L-smoothness (7.1), we can subsequently conclude the upper bound for the optimality gap as follows:

$$\begin{aligned}
f(X_t) - f(x^*) &\leq \langle \nabla f(x^*), X_t - x^*\rangle + \frac{1}{2}L\|x^* - X_t\|^2\\
&= \frac{1}{2}L\|x^* - X_t\|^2\\
&\leq e^{-\frac{2mL}{m+L}t}\frac{L}{2}\|x^* - X_0\|^2.
\end{aligned}$$

Hence we obtain $f(X_t) - f(x^*) \leq O(\frac{L}{2}e^{-\frac{2mL}{m+L}t})$.

(b). For gradient descent, as long as the function f is L-smooth and m-strongly convex, along with $0 < \delta \leq \frac{2}{m+L}$, the following function,

$$E_k = \left(1 - \frac{2mL}{(m+L)}\delta\right)^{-k}\frac{1}{2}\|x^* - x_k\|^2,$$

is a Lyapunov function. We check,

$$\begin{aligned}
\frac{E_{k+1}-E_k}{\delta} &= \left(1-\frac{2mL}{(m+L)}\delta\right)^{-(k+1)} \frac{1}{\delta}\left[\frac{1}{2}\|x^*-x_{k+1}\|^2 - \frac{1}{2}\left(1-\frac{2mL}{m+L}\delta\right)\|x^*-x_k\|^2\right] \\
&= \left(1-\frac{2mL}{(m+L)}\delta\right)^{-(k+1)} \left[\frac{\left(\frac{1}{2}\|x^*-x_{k+1}\|^2 - \frac{1}{2}\|x^*-x_k\|^2\right)}{\delta} + \frac{mL}{m+L}\|x^*-x_k\|^2\right] \\
&= \left(1-\frac{2mL}{(m+L)}\delta\right)^{-(k+1)} \left(-\langle\frac{x_{k+1}-x_k}{\delta}, x^*-x_k\rangle + \varepsilon_k^1 + \frac{mL}{m+L}\|x^*-x_k\|^2\right) \\
&= \left(1-\frac{2mL}{(m+L)}\delta\right)^{-(k+1)} \left(\langle\nabla f(x_k), x^*-x_k\rangle + \varepsilon_k^1 + \frac{mL}{m+L}\|x^*-x_k\|^2\right) \\
&\overset{(i)}{\leq} \left(1-\frac{2mL}{(m+L)}\delta\right)^{-(k+1)} \left(-\frac{mL}{m+L}\|x^*-x_k\|^2 + \varepsilon_k^2 + \frac{mL}{m+L}\|x^*-x_k\|^2\right) \\
&= \left(1-\frac{2mL}{(m+L)}\delta\right)^{-(k+1)} \varepsilon_k^2 \\
&\overset{(ii)}{\leq} 0,
\end{aligned}$$

where $\varepsilon_k^1 = \frac{\delta}{2}\|\nabla f(x_k)\|^2$ and $\varepsilon_k^2 = -\left(\frac{1}{m+L} - \frac{\delta}{2}\right)\|\nabla f(x_k)\|^2$. In (i) we use (7.4) to derive the inequality, in (ii) we use $\varepsilon_k^2 \leq 0$ since we take $\delta \in \left(0, \frac{2}{m+L}\right]$. By summing, we obtain the statement $E_k - E_0 \leq \sum_{i=0}^k \frac{E_{i+1}-E_i}{\delta}\delta \leq 0$, thus we have

$$\frac{1}{2}\|x^*-x_k\|^2 \leq \left(1-\frac{2mL}{(m+L)}\delta\right)^k E_0.$$

In addition, by the property of L-smoothness (7.1), we can subsequently conclude the upper bound for the optimality gap as follows:

$$\begin{aligned}
f(x_k) - f(x^*) &\leq \langle\nabla f(x^*), x_k - x^*\rangle + \frac{L}{2}\|x^*-x_k\|^2 \\
&= \frac{L}{2}\|x^*-x_k\|^2 \\
&\leq e^{-\frac{2mL}{m+L}\delta k}\frac{L}{2}\|x^*-x_0\|^2.
\end{aligned}$$

Hence we obtain $f(x_k) - f(x^*) \leq O(\frac{L}{2}e^{-\frac{2mL}{m+L}\delta k})$. □

7.3 Problem 3 (Proximal operator of the nuclear norm)

Consider the proximal operators of matrices

$$\mathrm{prox}_h(Y) \coloneqq \underset{X\in\mathbb{R}^{m\times n}}{\arg\min}\left\{\frac{1}{2}\|X-Y\|_F^2 + h(X)\right\},$$

and let $h(X) = \lambda\|X\|_*$. Show that the proximal operator of the nuclear norm is

$$\mathrm{prox}_h(Y) = U\mathrm{diag}(\{(\sigma_i-\lambda)_+\}_{1\leq i\leq r})V^\top,$$

where $Y = U\Sigma V^\top$, $\Sigma = \mathrm{diag}(\{\sigma_i\}_{1\leq i\leq r}) \in \mathbb{R}^{r\times r}$ is the SVD of $Y \in \mathbb{R}^{m\times n}$ of rank r, $\lambda \geq 0$, $t_+ = \max\{0, t\}$, $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_*$ denotes the nuclear norm.

Proof. Since $h_0(X) \coloneqq \frac{1}{2}\|X-Y\|_F^2 + \lambda\|X\|_*$ is strictly convex, it easy to see that there exists a unique minimizer, and we thus need to prove that it is equal to $U\mathrm{diag}(\{(\sigma_i-\lambda)_+\}_{1\le i\le r})V^\top$. We denote $\hat{X} = \mathrm{prox}_h(Y)$ for convenience. Note that $\hat{X}$ minimizes $h_0(X)$ if and only if $0 \in \partial h_0(X)$, i.e.,

$$0 \in \hat{X} - Y + \lambda\partial\|\hat{X}\|_*, \tag{7.5}$$

where $\partial\|\hat{X}\|_*$ is the subdifferential of the nuclear norm. Let $X \in \mathbb{R}^{m\times n}$ be an arbitary matrix and $U\Sigma V^\top$ be its SVD. Recall that the subdifferential of the nuclear norm is

$$\partial\|X\|_* = \left\{UV^\top + W \mid U^\top W = 0, WV = 0, \|W\|_2 \le 1, W \in \mathbb{R}^{m\times n}\right\},$$

where $\|\cdot\|_2$ denotes the spectral norm. Now we set $\hat{X} \coloneqq U\mathrm{diag}(\{(\sigma_i-\lambda)_+\}_{1\le i\le r})V^\top$ for short. In order to show that $\hat{X}$ satisfies (7.5), we decompose the SVD of Y as

$$Y = U_1\Sigma_1V_1^\top + U_2\Sigma_2V_2^\top,$$

where U_1, V_1 are the singular vectors associated with singular values greater than λ, and U_1, V_1 are the singular vectors associated with singular values smaller or equal to λ. With these notations, we have

$$\hat{X} = U_1(\Sigma_1 - \lambda I)V_1^\top,$$

and, therefore,

$$Y - \hat{X} = \lambda(U_1V_1^\top + W), \quad W = U_2\left(\frac{1}{\lambda}\Sigma_2\right)V_2^\top.$$

By definition, the columns of U and V are orthonormal. Thus we have $U_1^\top W = \frac{1}{\lambda}U_1^\top U_2\Sigma_2V_2^\top = 0$, and $WV_1 = \frac{1}{\lambda}U_2\Sigma_2V_2^\top V_1 = 0$. Since the diagonal elements of $\frac{1}{\lambda}\Sigma_2$ have magnitudes bounded by 1, we also have $\|W\|_2 \le 1$. Hence $Y - \hat{X} \in \lambda\partial\|\hat{X}\|_*$, which concludes the proof. □

7.4 Additional Proof

We provide a proof of inequality (7.4) here.

Proof. Define $\phi(x) \coloneqq f(x) - \frac{m}{2}\|x\|^2$, then $\nabla\phi(x) = \nabla f(x) - mx$. Since f is L-smooth, we have

$$(\nabla f(x) - \nabla f(y)^\top(x-y) \le L\|x-y\|^2,$$

and this implies that

$$(\nabla\phi(x) - \nabla\phi(y))^\top(x-y) \le (L-\mu)\|x-y\|^2.$$

Thus ϕ is $(L-\mu)$-smooth. This, in turn, implies

$$\langle\nabla\phi(x) - \nabla\phi(y), x-y\rangle \ge \frac{1}{L-m}\|\nabla\phi(x) - \nabla\phi(y)\|^2.$$

Expanding the above inequality and rearranging, then we get

$$\langle\nabla f(x) - \nabla f(y), x-y\rangle \ge \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Let $x = X_t$, $y = x^*$, and note that $\nabla f(x^*) = 0$, this yields

$$\langle\nabla f(X_t), X_t - x^*\rangle \ge \frac{mL}{m+L}\|X_t - x^*\|^2 + \frac{1}{m+L}\|\nabla f(X_t)\|^2.$$

Multiply both sides by -1 and we obtain the inequality (7.4). □

8 Solutions to Homework 8 (May 16, 2022)

8.1 Problem 1 (Convergence property of proximal gradient descent)

Let $F(x) \coloneqq f(x) + h(x)$, assume that f is L-smooth and that h is convex. Consider the proximal gradient method

$$\begin{aligned} x_{k+1} &= x_k - \operatorname{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \\ &\coloneqq x_k - \alpha_k G_{\alpha_k}(x_k) \end{aligned}$$

with fixed step-size $\alpha_k = \alpha = \frac{1}{L}$, and suppose that there exists some scalar $R > 0$ such that for all $k \in \mathbb{N}$,

$$\|x_k - x^*\| \leq R,$$

and that

$$G_\alpha(x)^\top (x - x^*) \geq \gamma (F(x) - F^*). \tag{8.1}$$

holds for some scalar $\gamma > 0$ and for all $x \in \mathbb{R}^d$. Then, for all $k \in \mathbb{N}$, the optimality gap satisfies

$$F(x_k) - F^* \leq \frac{2LR^2\epsilon_0}{k\gamma^2\epsilon_0 + 2LR^2} = O\left(\frac{1}{k}\right),$$

where $\epsilon_0 \coloneqq F(w_0) - F^*$ and $F^* \coloneqq F(x^*)$ is the minimum of function F.

Proof. First, since f is L-smooth, we have

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{8.2}$$

In particular, let $y = x - \alpha G_\alpha(x)$, we have

$$f(x - \alpha G_\alpha(x)) \leq f(x) - \alpha \nabla f(x)^\top G_\alpha(x) + \frac{\alpha^2 L}{2}\|G_\alpha(x)\|^2. \tag{8.3}$$

Since h is convex, then by subgradient characterization we have

$$G_\alpha(x) - \nabla f(x) \in \partial h(x - \alpha G_\alpha(x)).$$

By definition of subgradient, we obtain

$$h(z) \geq h(x - \alpha G_\alpha(x)) + (G_\alpha(x) - \nabla f(x))^\top (z - x + \alpha G_\alpha(x)), \quad \forall z \in \mathbb{R}^d. \tag{8.4}$$

Let $z = x$ in (8.4), we get

$$h(x - \alpha G_\alpha(x)) \leq h(x) + \alpha \nabla f(x)^\top G_\alpha(x) - \alpha \|G_\alpha(x)\|^2. \tag{8.5}$$

Adding (8.3) and (8.5) together gives

$$F(x - \alpha G_\alpha(x)) \leq F(x) - \alpha\left(1 - \frac{\alpha L}{2}\right)\|G_\alpha(x)\|^2. \tag{8.6}$$

Set $x = x_k$ and $\alpha = 1/L$ in (8.6), then

$$F(x_{k+1}) \leq F(x_k) - \frac{1}{2L}\|G_{\frac{1}{L}}(x_k)\|^2. \tag{8.7}$$

By (8.1) and Cauchy-Schwarz inequality,

$$\gamma(F(x_k) - F^*) \leq G_{\frac{1}{L}}(x_k)^\top (x_k - x^*) \leq \|G_{\frac{1}{L}}(x_k)\|\|x_k - x^*\|.$$

Denote $\epsilon_k \coloneqq F(x_k) - F^*$, then we have

$$\epsilon_k \leq \frac{1}{\gamma}\|G_{\frac{1}{L}}(x_k)\|\|x_k - x^*\| \leq \frac{R}{\gamma}\|G_{\frac{1}{L}}(x_k)\|. \tag{8.8}$$

Combining (8.7) and (8.8) yields

$$\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L}\left(\frac{\epsilon_k \gamma}{R}\right)^2. \tag{8.9}$$

Therefore,

$$\frac{1}{\epsilon_{k+1}} - \frac{1}{\epsilon_k} = \frac{\epsilon_k - \epsilon_{k+1}}{\epsilon_k \epsilon_{k+1}} \geq \frac{\epsilon_k - \epsilon_{k+1}}{\epsilon_k^2} \geq \frac{\gamma^2}{2LR^2},$$

Sum over the above inequality for $i \leq k-1$ and we get

$$\sum_{i=0}^{k-1}\left(\frac{1}{\epsilon_{i+1}} - \frac{1}{\epsilon_i}\right) \geq \frac{k\gamma^2}{2LR^2}.$$

Rearranging it yields

$$\frac{1}{\epsilon_k} \geq \frac{1}{\epsilon_0} + \frac{k\gamma^2}{2LR^2}.$$

Hence we conclude that

$$F(x_k) - F^* = \epsilon_k \leq \frac{2LR^2\epsilon_0}{k\gamma^2\epsilon_0 + 2LR^2} = O\left(\frac{1}{k}\right).$$

□

9 Solutions to Homework 9 (May 30, 2022)

9.1 Problem 1 (Every RKHS has a unique reproducing kernel)

Every RKHS (reproducing kernel Hilbert space) $\mathcal{H}$ has a unique reproducing kernel k.

Proof. To establish existence of a reproducing kernel in an RKHS, we will make use of the Riesz representation theorem, which tells us that in an RKHS, evaluation itself can be represented as an inner product. We will first show the existence of the reproducing kernel k, and then show the uniqueness of it.

Existence of the reproducing kernel

By definition, since $\mathcal{H}$ is an RKHS, then its evaluation functionals δ_x are continuous linear operators. Assume that $\delta_x \in \mathcal{H}'$ (topological dual space of $\mathcal{H}$), i.e., $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional. The Riesz representation theorem states that there exists an element $f_{\delta_x} \in \mathcal{H}$ such that

$$\delta_x f = \langle f, f_{\delta_x} \rangle_{\mathcal{H}}, \ \forall f \in \mathcal{H}.$$

Define $k(x', x) = f_{\delta_x}(x'), \ \forall x, x' \in \mathcal{X}$. Then, clearly we have $k(\cdot, x) = f_{\delta_x} \in \mathcal{H}$, and $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \delta_x f = f(x)$, which is exactly the reproducing property. Thus, k is the reproducing kernel.

Uniqueness of the reproducing kernel

Assume that $\mathcal{H}$ has two reproducing kernels k_1 and k_2. Then,

$$\langle f, k_1(\cdot, x) - k_2(\cdot, x) \rangle_{\mathcal{H}} = f(x) - f(x) = 0, \ \forall f \in \mathcal{H}, \forall x \in \mathcal{X}.$$

In particular, if we take $f = k_1(\cdot, x) - k_2(\cdot, x)$, we obtain $\|k_1(\cdot, x) - k_2(\cdot, x)\|_{\mathcal{H}}^2 = 0, \ \forall x \in \mathcal{X}$, i.e., $k_1 = k_2$. Hence the reproducing kernel is unique. □

9.2 Problem 2 (Every kernel has a unique RKHS)

For every kernel k, there corresponds a unique RKHS (reproducing kernel Hilbert space) $\mathcal{H}$, for which k is a reproducing kernel.

Remark. *The proof for this problem is largely based on RKHS notes. However, authors of the notes do not give a clear proof of the uniqueness of the RKHS $\mathcal{H}$ associated to kernel k, and we supplement the detailed proof of this part in the last section.*

Proof. Starting with the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we will construct a pre-RKHS $\mathcal{H}_0$, from which we will form the RKHS $\mathcal{H}$. The pre-RKHS $\mathcal{H}_0$ must satisfy two properties:

(1) The evaluation functionals δ_x are continuous on $\mathcal{H}_0$.

(2) Any Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0$ which converges pointwise to 0 also converges in $\mathcal{H}_0$-norm to 0.

The last result has an important implication: Any Cauchy sequence f_n in $\mathcal{H}_0$ which converges pointwise to $f \in \mathcal{H}_0$, also converges to f in $\|\cdot\|_{\mathcal{H}_0}$, since in that case $\{f_n - f\}$ converges pointwise to 0, and thus $\|f_n - f\|_{\mathcal{H}_0} \to 0$.

Proof technique overview

First, we can already say what the pre-RKHS $\mathcal{H}_0$ will look like: it is the set of functions

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x). \tag{9.1}$$

After the proof, we will show in the last section that these functions satisfy properties (1) and (2) of the pre-RKHS $\mathcal{H}_0$.

Next, define $\mathcal{H}$ to be the set of functions $f \in \mathbb{R}^{\mathcal{X}}$ for which there exists an $\mathcal{H}_0$-Cauchy sequence $\{f_n\} \in \mathcal{H}_0$ converging pointwise to f. Note that $\mathcal{H}_0 \subset \mathcal{H}$, since the limits of these Cauchy sequences might not be in $\mathcal{H}_0$. Our goal is to prove that $\mathcal{H}$ is an RKHS. The two properties above hold if and only if

- $\mathcal{H}_0 \subset \mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ and the topology induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ on $\mathcal{H}_0$ coincides with the topology induced on $\mathcal{H}_0$ by $\mathcal{H}$.
- $\mathcal{H}$ has reproducing kernel $k(x, y)$.

In the first four sections, we concern ourselves with proving that two properties of the pre-RKHS $\mathcal{H}_0$, namely (1) and (2), imply the above bullet points, since the reverse direction is easy to prove. This takes four steps:

- We define the inner product between $f, g \in \mathcal{H}$ as the limit of an inner product of the Cauchy sequences $\{f_n\}, \{g_n\}$ converging pointwise to f and g respectively. Is the inner product well defined and independent of the sequences used? This is proved in the first section.
- Recall that an inner product space must satisfy $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$. Is this true when we define the inner product on $\mathcal{H}$ as above? This is proved in the second section. (Note that we also check that the remaining requirements for an inner product on $\mathcal{H}$ hold, and these are straightforward)
- Are the evaluation functionals still continuous on $\mathcal{H}$? This is proved in the third section.
- Is $\mathcal{H}$ complete? I.e., is it a Hilbert space? This is proved in the fourth section.

Finally, in the last two sections we'll see that the functions (9.1) actually define a valid pre-RKHS $\mathcal{H}_0$, and that the kernel k has the **reproducing property** on the RKHS $\mathcal{H}$. We'll also show that the RKHS $\mathcal{H}$ associated to kernel k is unique.

Is the inner product well defined in $\mathcal{H}$?

In this section we prove that if we define the inner product in $\mathcal{H}$ of all limits of Cauchy sequences as (9.2) below, then this limit is well defined: **(1)** it converges, and **(2)** it depends only on the limits of the Cauchy sequences, and not the particular sequences themselves.

Lemma 9.1. *For $f, g \in \mathcal{H}$ and Cauchy sequences (wrt the $\mathcal{H}_0$ norm) $\{f_n\}, \{g_n\}$ converging pointwise to f and g, define $\alpha_n = \langle f_n, g_n \rangle_{\mathcal{H}_0}$. Then, $\{\alpha_n\}$ is convergent and its limits only on f and g. We thus define*

$$\langle f, g \rangle_{\mathcal{H}} \coloneqq \lim_{n \to \infty} \langle f_n, g_n \rangle_{\mathcal{H}_0}. \tag{9.2}$$

Proof of Lemma 9.1. We first show that $\alpha_n = \langle f_n, g_n \rangle_{\mathcal{H}_0}$ is convergent. For $n, m \in \mathbb{N}$, we have

$$\begin{aligned}
|\alpha_n - \alpha_m| &= |\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_m \rangle_{\mathcal{H}_0}| \\
&= |\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_n \rangle_{\mathcal{H}_0} + \langle f_m, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_m \rangle_{\mathcal{H}_0}| \\
&\leq |\langle f_n - f_m, g_n \rangle_{\mathcal{H}_0}| + |\langle f_m, g_n - g_m \rangle_{\mathcal{H}_0}| \\
&\leq \|g_n\|_{\mathcal{H}_0} \|f_n - f_m\|_{\mathcal{H}_0} + \|f_m\|_{\mathcal{H}_0} \|g_n - g_m\|_{\mathcal{H}_0}.
\end{aligned}$$

Since every Cauchy sequence is bounded, so $\exists A, B \in \mathbb{R}$, $|f_m\|_{\mathcal{H}_0} \leq A$, $\|g_n\|_{\mathcal{H}_0} \leq B$, $\forall n, m \in \mathbb{N}$. For any $\epsilon > 0$, by taking $N_1 \in \mathbb{N}$, s.t. $\|f_n - f_m\|_{\mathcal{H}_0} \leq \frac{\epsilon}{2B}$, for $n, m \geq N_1$, and $N_2 \in \mathbb{N}$, s.t. $\|g_n - g_m\|_{\mathcal{H}_0} \leq \frac{\epsilon}{2A}$. Thus we have that $|\alpha_n - \alpha_m| < \epsilon$, for $n, m \geq \max\{N_1, N_2\}$, which means that $\{\alpha_n\}$ is a Cauchy sequence in $\mathbb{R}$, which is complete, hence the sequence is convergent in $\mathbb{R}$.

Next we show that the limit defined as (9.2) is independent of Cauchy sequence chosen. If some $\mathcal{H}_0$-Cauchy sequences $\{f'_n\}$, $\{g'_n\}$ also converge pointwise to f and g, and $\alpha'_n = \langle f'_n, g'_n \rangle_{\mathcal{H}_0}$, one similarly shows that

$$|\alpha_n - \alpha'_n| \leq \|g_n\|_{\mathcal{H}_0}\|f_n - f'_n\|_{\mathcal{H}_0} + \|f'_n\|_{\mathcal{H}_0}\|g_n - g'_n\|_{\mathcal{H}_0}.$$

Now, since both $\{f_n\}$ and $\{f'_n\}$ both converge pointwise to f, $\{f_n - f'_n\}$ converges pointwise to 0, and so does $\{g_n - g'_n\}$. But then they also converge to 0 in $\|\cdot\|_{\mathcal{H}_0}$ by the pre-RKHS axiom (2), and therefore $\{\alpha_n\}$ and $\{\alpha'_n\}$ must have the same limit. □

Does it hold that $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$?

In this section, we verify all the expected properties of an inner product hold for $\mathcal{H}$. Assume $f, f^1, f^2, g \in \mathcal{H}$ and Cauchy sequences (wrt the $\mathcal{H}_0$ norm) $\{f_n\}, \{f^1_n\}, \{f^2_n\}, \{g_n\}$ converging pointwise to f, f^1, f^2 and g.

(1) Linearity:

$$\begin{aligned}
\langle \beta_1 f^1 + \beta_2 f^2, g \rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle \beta_1 f^1_n + \beta_2 f^2_n, g_n \rangle_{\mathcal{H}_0} \\
&= \lim_{n\to\infty} \beta_1 \langle f^1_n, g_n \rangle_{\mathcal{H}_0} + \beta_2 \langle f^2_n, g_n \rangle_{\mathcal{H}_0} \\
&= \beta_1 \lim_{n\to\infty} \langle f^1_n, g_n \rangle_{\mathcal{H}_0} + \beta_2 \lim_{n\to\infty} \langle f^2_n, g_n \rangle_{\mathcal{H}_0} \\
&= \beta_1 \langle f^1, g \rangle_{\mathcal{H}} + \beta_2 \langle f^2, g \rangle_{\mathcal{H}}.
\end{aligned}$$

(2) Symmetric Property:

$$\begin{aligned}
\langle f, g \rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle f_n, g_n \rangle_{\mathcal{H}_0} \\
&= \lim_{n\to\infty} \langle g_n, f_n \rangle_{\mathcal{H}_0} \\
&= \langle g, f \rangle_{\mathcal{H}}.
\end{aligned}$$

(3) Positive Definite Property:

$$\langle f, f \rangle_{\mathcal{H}} = \lim_{n\to\infty} \langle f_n, f_n \rangle_{\mathcal{H}_0} \geq 0,$$

$$\langle f, f \rangle_{\mathcal{H}} = 0 \overset{?}{\Longleftrightarrow} f = 0.$$

It turns out that the only challenging property to show is the third one, the others follow from the inner product definition on the pre-RKHS as above. Therefore, we need to establish a lemma before we obtain the result.

Lemma 9.2. *Let $\{f_n\}$ be Cauchy sequence in $\mathcal{H}_0$ converging pointwise to $f \in \mathcal{H}$. If $\langle f, f \rangle_{\mathcal{H}} = \lim_{n\to\infty} \langle f_n, f_n \rangle_{\mathcal{H}_0} = 0$, then $f(x) = 0$ pointwise for $x \in \mathcal{X}$.*

Proof of Lemma 9.2. For any $x \in \mathcal{X}$, we have

$$\begin{aligned}
f(x) &= \lim_{n\to\infty} f_n(x) \\
&= \lim_{n\to\infty} \delta_x(f_n) \\
&\overset{(a)}{\leq} \lim_{n\to\infty} \|\delta_x\| \|f_n\|_{\mathcal{H}_0} \\
&\overset{(b)}{=} 0,
\end{aligned}$$

where in (a) we use that the evaluation functional δ_x is continuous on $\mathcal{H}_0$, by pre-RKHS axiom (1) (hence bounded, with a well defined operator norm $\|\delta_x\|$); and in (b) we use the assumption in the lemma that f_n converges to 0 in $\|\cdot\|_{\mathcal{H}_0}$. □

By Lemma 9.2 we prove that $\langle f, f\rangle_{\mathcal{H}} \implies f = 0$. To derive the other direction, pre-RKHS axiom (2) indicates that any Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0$ which converges pointwise to 0 also converges in $\mathcal{H}_0$-norm to 0, i.e., $f = 0 \implies \langle f, f\rangle_{\mathcal{H}} = \lim_{n\to\infty}\langle f_n, f_n\rangle_{\mathcal{H}_0} = \lim_{n\to\infty}\|f_n\|^2_{\mathcal{H}_0} = 0$. Therefore, $\langle f, f\rangle_{\mathcal{H}} = 0 \iff f = 0$ holds, and all the expected properties of an inner product hold for $\mathcal{H}$.

Are the evaluation functionals continuous on $\mathcal{H}$?

Here we need to establish a preliminary lemma, before we can continue to discuss the continuity of the evaluation functionals.

Lemma 9.3. *$\mathcal{H}_0$ is dense in $\mathcal{H}$.*

Proof of Lemma 9.3. It suffices to show that given any $f \in \mathcal{H}$ and its associated Cauchy sequence $\{f_n\}$ wrt $\mathcal{H}_0$ converging pointwise to f (which exists by definition), $\{f_n\}$ also converges to f in $\|\cdot\|_{\mathcal{H}}$ (note that this is the new norm which we defined above in terms of limits of Cauchy sequences in $\mathcal{H}_0$).

Since $\{f_n\}$ is Cauchy in $\mathcal{H}_0$-norm, for any $\epsilon > 0$, there exists $N \in \mathbb{N}$, s.t. $\|f_m - f_n\|_{\mathcal{H}_0} < \epsilon$, $\forall m, n \geq N$. Fix $n^* \geq N$, the sequence $\{f_m - f_{n^*}\}_{m=1}^{\infty}$ converges pointwise to $f - f_{n^*}$. We now simply use the definition of the inner product in $\mathcal{H}_0$ from (9.2),

$$\|f - f_{n^*}\|^2_{\mathcal{H}} = \lim_{m\to\infty}\|f_m - f_{n^*}\|^2_{\mathcal{H}_0} \leq \epsilon^2,$$

whereby $\{f_n\}_{n=1}^{\infty}$ converges to f in $\|\cdot\|_{\mathcal{H}}$. □

Lemma 9.4. *The evaluation functionals are continuous on $\mathcal{H}$.*

Proof of Lemma 9.4. We show that δ_x is continuous at $f = 0$, since this implies by linearity that it is continuous everywhere. Let $x \in \mathcal{X}$, and $\epsilon > 0$. By pre-RKHS axiom (1), δ_x is continuous on $\mathcal{H}_0$. Thus, $\exists\eta > 0$, s.t.

$$\|g - 0\|_{\mathcal{H}_0} = \|g\|_{\mathcal{H}_0} < \eta \implies |\delta_x(g)| = |g(x)| < \frac{\epsilon}{2}. \tag{9.3}$$

To complete the proof, we just need to show that there is a $g \in \mathcal{H}_0$ close (in $\mathcal{H}$-norm) to some $f \in \mathcal{H}$ with small norm, and that this function is also close at each point.

We take $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} < \eta/2$. By Lemma 9.3 there is a Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0$ converging both pointwise to f and in $\|\cdot\|_{\mathcal{H}}$ to f, so one can find $N \in \mathbb{N}$, s.t.

$$|f(x) - f_N(x)| < \frac{\epsilon}{2},$$

$$\|f - f_N\|_{\mathcal{H}} < \frac{\eta}{2}.$$

We have from these definitions that

$$\|f_N\|_{\mathcal{H}_0} = \|f_N\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} + \|f - f_N\|_{\mathcal{H}} < \eta.$$

Thus $\|f\|_{\mathcal{H}} < \eta/2$ implies $\|f_N\|_{\mathcal{H}_0} < \eta$. Using (9.3) and setting $g \coloneqq f_N$, we have that $\|f_N\|_{\mathcal{H}_0} < \eta$ implies $|f_N(x)| < \epsilon/2$, and thus $|f(x)| \leq |f(x) - f_N(x)| + |f_N(x)| < \epsilon$. In other words, $\|f\|_{\mathcal{H}} < \eta/2$ is shown to imply $|f(x)| < \epsilon$. This means that δ_x is continuous at 0 in the $\|\cdot\|_{\mathcal{H}}$ sense, and thus by linearity on all $\mathcal{H}$. □

Is $\mathcal{H}$ complete (a Hilbert space)?

The idea here is to show that every Cauchy sequence wrt the $\mathcal{H}$-norm converges to a function in $\mathcal{H}$.

Lemma 9.5. *$\mathcal{H}$ is complete.*

Proof of Lemma 9.5. Let $\{f_n\}$ be any Cauchy sequence in $\mathcal{H}$. Since evaluation functionals are linear continuous on $\mathcal{H}$ by Lemma 9.4, then for any $t \in \mathcal{X}$, $\{f_n(t)\}$ is convergent in $\mathbb{R}$ to some $f(t) \in \mathbb{R}$ (since $\mathbb{R}$ is complete, it contains this limit). The question is thus whether the function $f(t)$ defined pointwise in this way is still in $\mathcal{H}$ (recall that $\mathcal{H}$ is defined as containing the limit of $\mathcal{H}_0$-Cauchy sequences that converge pointwise).

The proof strategy is to define a sequence of functions $\{g_n\}$, where $\{g_n\} \in \mathcal{H}_0$, which is "close" to the $\mathcal{H}$-Cauchy sequence $\{f_n\}$. These functions will then be shown **(1)** to converge pointwise to f, and **(2)** to be Cauchy in $\mathcal{H}_0$. Hence by our original construction of $\mathcal{H}$, we have $f \in \mathcal{H}$. Finally, we show $f_n \to f$ in $\mathcal{H}$-norm.

Define $f(x) \coloneqq \lim_{n\to\infty} f_n(x)$. For $n \in \mathbb{N}$, choose $g_n \in \mathcal{H}_0$ such that $\|g_n - f_n\|_{\mathcal{H}} < \frac{1}{n}$. This can be done since $\mathcal{H}_0$ is dense in $\mathcal{H}$, and we have

$$\begin{aligned} |g_n(x) - f(x)| &\le |g_n(x) - f_n(x)| + |f_n(x) - f(x)| \\ &\le |\delta_x(g_n - f_n)| + |f_n(x) - f(x)|. \end{aligned}$$

The first term in this sum goes to zero due to the continuity of δ_x on $\mathcal{H}$ (Lemma 9.4), and thus $\{g_n(x)\}$ converges to $f(x)$, satisfying criterion (1). For criterion (2), we have

$$\begin{aligned} \|g_m - g_n\|_{\mathcal{H}_0} &= \|g_m - g_n\|_{\mathcal{H}} \\ &\le \|g_m - f_m\|_{\mathcal{H}} + \|f_m - f_n\|_{\mathcal{H}} + \|f_n - g_n\|_{\mathcal{H}} \\ &\le \frac{1}{m} + \frac{1}{n} + \|f_m - f_n\|_{\mathcal{H}}, \end{aligned}$$

hence $\{g_n\}$ is Cauchy in $\mathcal{H}_0$.

Finally, since $\{g_n\}$ wrt $\mathcal{H}_0$ converges pointwise to f and by Lemma 9.3, we know that $\{g_n\}$ also converges to f in $\|\cdot\|_{\mathcal{H}}$. Moreover,

$$\begin{aligned} \|f_n - f\|_{\mathcal{H}} &\le \|f_n - g_n\|_{\mathcal{H}} + \|g_n - f\|_{\mathcal{H}} \\ &\le \frac{1}{n} + \|g_n - f\|_{\mathcal{H}}. \end{aligned}$$

Take $n \to \infty$, we get $\|f_n - f\|_{\mathcal{H}} \to 0$, thus $\{f_n\}$ converges to f in $\mathcal{H}$-norm. Hence $\mathcal{H}$ is complete. □

How to build a valid pre-RKHS $\mathcal{H}_0$?

Here we show how to build a valid pre-RKHS. Imporantly, in doing this, we prove that for every positive definite kernel k, there corresponds an RKHS $\mathcal{H}$.

Theorem 9.1 (Moore-Aronszajn)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive definite. There is an RKHS with reproducing kernel k. Moreover, if space $\mathcal{H}_0 = [\{k(\cdot, x)\}_{x\in\mathcal{X}}]$ is endowed with the inner product*

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(y_i, x_j), \tag{9.4}$$

where $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$, then $\mathcal{H}_0$ is a valid pre-RKHS.

Proof of Theorem 9.1. We first need to show that (9.4) is a **valid inner product**. First, it is independent of the particular α_i and β_i used to define f, g, since

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(y_j).$$

As a useful consequence of this result, we get the **reproducing property** on $\mathcal{H}_0$, by setting $g = k(x, \cdot)$,

$$\langle f, g\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x).$$

Next we check that the form (9.4) is indeed a valid inner product on $\mathcal{H}_0$. As what we have discussed in the second section, the only nontrivial axiom to be verified here is

$$\langle f, f\rangle_{\mathcal{H}_0} = 0 \Longrightarrow f = 0.$$

This is true since

$$\forall x \in \mathcal{X},\ f(x) = \langle f, k(x, \cdot)\rangle \overset{(a)}{\leq} \sqrt{k(x, x)}\|f\|_{\mathcal{H}_0} = 0,$$

where in (a) we use Cauchy-Schwarz inequality. We now proceed to the main proof and we will show that $\mathcal{H}_0$ satisfies the pre-RKHS axioms.

Let $t \in \mathcal{X}$, note that for $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ we have

$$\langle f, k(\cdot, t)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i k(t, x_i) = f(t),$$

and thus for $f, g \in \mathcal{H}_0$,

$$\begin{aligned} |\delta_x(f) - \delta_x(g)| &= |\langle f - g, k(\cdot, x)\rangle_{\mathcal{H}_0}| \\ &\leq \sqrt{k(x, x)}\|f - g\|_{\mathcal{H}_0}, \end{aligned}$$

meaning δ_x is continuous on $\mathcal{H}_0$, thus the **first** pre-RKHS requirement is satisfied.

Now, take any $\epsilon > 0$ and define a Cauchy $\{f_n\}$ in $\mathcal{H}_0$ that converges pointwise to 0. Since Cauchy sequences are bounded, there exists some $A > 0$, s.t. $\|f_n\|_{\mathcal{H}_0} < A$, $\forall n \in \mathbb{N}$. One can find $N_1 \in \mathbb{N}$, s.t. $\|f_n - f_m\|_{\mathcal{H}_0} < \frac{\epsilon}{2A}$, for $n, m \geq N_1$. Write $f_{N_1} = \sum_{i=1}^{k} \alpha_i k(\cdot, x_i)$, one can also find $N_2 \in \mathbb{N}$, s.t. $|f_n(x_i)| < \frac{\epsilon}{2k|\alpha_i|}$, for $i \in [k]$. Now, for $n \geq \max\{N_1, N_2\}$ we have

$$\begin{aligned} \|f_n\|_{\mathcal{H}_0} &\leq |\langle f_n - f_{N_1}, f_n\rangle_{\mathcal{H}_0}| + |\langle f_{N_1}, f_n\rangle_{\mathcal{H}_0}| \\ &\leq \|f_n - f_{N_1}\|_{\mathcal{H}_0} + \sum_{i=1}^{k} |\alpha_i f_n(x_i)| \\ &< \epsilon, \end{aligned}$$

so f_n converges to 0 in $\|\cdot\|_{\mathcal{H}}$, which is exactly the **second** pre-RKHS requirement. Therefore, all the pre-RKHS axioms are satisfied, and thus $\mathcal{H}$ is an RKHS.

To see that the **reproducing kernel** on $\mathcal{H}$ is k, simply note that if $f \in \mathcal{H}$, and $\{f_n\}$ in $\mathcal{H}_0$ converges to f pointwise,

$$\begin{aligned} \langle f, k(\cdot, x)\rangle_{\mathcal{H}} &\overset{(a)}{=} \langle f_n, k(\cdot, x)\rangle_{\mathcal{H}_0} \\ &= \lim_{n\to\infty} f_n(x) \\ &= f(x), \end{aligned}$$

where in (a) we use the definition of an inner product on $\mathcal{H}$ in (9.2). □

Is RKHS $\mathcal{H}$ associated to kernel k unique?

Let's now prove that for every kernel k, it has only one RKHS. To this end, let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two RKHSs of kernel k. We have seen in the previous step (Lemma 9.3) that $\mathcal{H}_0$ is dense in both $\mathcal{H}_1$ and $\mathcal{H}_2$, and that the norms of $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide on $\mathcal{H}_0$. Let's choose any $f \in \mathcal{H}_1$, then there exists a sequence $\{f_n\}$ in $\mathcal{H}_0$ such that $\|f_n - f\|_{\mathcal{H}_1} \to 0$. Since $\mathcal{H}_0 \subset \mathcal{H}_2$, the sequence $\{f_n\}$ is also contained in $\mathcal{H}_2$, and since the norms of $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide on $\mathcal{H}_0$, the sequence $\{f_n\}$ is a Cauchy sequence in $\mathcal{H}_2$. Therefore, there exists a $g \in \mathcal{H}_2$ such that $\|f_n - g\|_{\mathcal{H}_2} \to 0$. Since convergence with respect to an RKHS norm implies pointwise convergence, we then find $f(x) = \lim_{n\to\infty} f_n(x) = g(x)$ for all $x \in \mathcal{X}$. In other words, $f = g$ and therefore $f \in \mathcal{H}_2$. This shows that $\mathcal{H}_1 \subset \mathcal{H}_2$. Furthermore, $\|f_n - f\|_{\mathcal{H}_1} \to 0$ and $\|f_n - f\|_{\mathcal{H}_2} \to 0$ imply

$$\|f\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_0} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_2} = \|f_n\|_{\mathcal{H}_2},$$

i.e., the norms in $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide and therefore $\mathcal{H}_1$ is isometrically included in $\mathcal{H}_2$. Then by symmetry of the argument, in fact we have that $\mathcal{H}_2$ is also isometrically included in $\mathcal{H}_1$. Thus we obtain $\mathcal{H}_1 = \mathcal{H}_2$ with equal norms, that is, the RKHS $\mathcal{H}$ associated to kernel k is unique.

Summary

So far we have proved that for every kernel k, there corresponds a unique RKHS $\mathcal{H}$, for which k is a reproducing kernel. Moreover, Moore-Aronszajn theorem tells us that every positive definite function is a reproducing kernel, we also know that every reproducing kernel is a kernel and that every kernel is a positive definite function. Therefore, all three notions are exactly the same. □

9.3 Problem 3 (SVM with kernels)

SVM optimization problem with kernels can be derived from

$$\min_{f\in\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=i}^{n} \varphi_{\text{hinge}}(y_i f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{9.5}$$

where $\mathcal{H}$ is an RKHS, $\lambda = \frac{1}{2nC}$, the hinge loss is the function $\varphi : \mathbb{R} \to \mathbb{R}_+$:

$$\varphi_{\text{hinge}}(u) = \max\{0, 1-u\} = \begin{cases} 0 & \text{if } u \geq 1, \\ 1-u & \text{otherwise.} \end{cases}$$

Proof. We first denote K be the kernel associated to the RKHS $\mathcal{H}$ and $\mathbf{K}$ be the kernel matrix associated to K. By representer theorem, the solution satisfies

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \hat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}),$$

where $\hat{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_n)^\top$ solves

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi_{\text{hinge}}(y_i [\mathbf{K}\boldsymbol{\alpha}]_i) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \right\}.$$

Primal optimization problem and support vectors

This is a convex optimization problem, but the objective function is not smooth because of the hinge loss. Let us introduce additional slack variables $\xi_1, \ldots, \xi_n \in \mathbb{R}$. The problem is equivalent to:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\xi}\in\mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \right\} \tag{9.6}$$

$$\text{subject to}^1 \quad y_i[\mathbf{K}\boldsymbol{\alpha}]_i + \xi_i - 1 \geq 0 \ \wedge \ \xi_i \geq 0, \ i \in [n],$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$.

The problem (9.6) is a convex optimization problem since the objective function as well as the constraints are convex and differentiable. Thus the KKT conditions hold and they apply at the optimum. We use these conditions to both analyze the algorithm and demonstrate several its crucial properties, and subsequently derive the dual optimization problem associated to SVM in the next section.

Let us introduce the Lagrange multipliers $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n_+$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top \in \mathbb{R}^n_+$. The Lagragian of the problem can be defined by

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{n}\sum_{i=1}^n \xi_i + \lambda \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} - \sum_{i=1}^n \mu_i \left(y_i[\mathbf{K}\boldsymbol{\alpha}]_i + \xi_i - 1\right) - \sum_{i=1}^n \nu_i \xi_i,$$

or in matrix notations:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{n}\boldsymbol{\xi}^\top \mathbf{1}_n + \lambda \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} - \left(\text{diag}(\mathbf{y})\boldsymbol{\mu}\right)^\top \mathbf{K}\boldsymbol{\alpha} - (\boldsymbol{\mu} + \boldsymbol{\nu})^\top \boldsymbol{\xi} + \boldsymbol{\mu}^\top \mathbf{1}_n.$$

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables $\boldsymbol{\alpha}, \boldsymbol{\xi}$ to zero and by writing the complementary slackness conditions:

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L} = \mathbf{K}(2\lambda\boldsymbol{\alpha} - \text{diag}(\mathbf{y})\boldsymbol{\mu}) = 0 \qquad \Longrightarrow \qquad \boldsymbol{\alpha} = \frac{\text{diag}(\mathbf{y})\boldsymbol{\mu}}{2\lambda} \tag{9.7}$$

$$\nabla_{\boldsymbol{\xi}} \mathcal{L} = \frac{\mathbf{1}_n}{n} - \boldsymbol{\mu} - \boldsymbol{\nu} = 0 \qquad \Longrightarrow \qquad \boldsymbol{\mu} + \boldsymbol{\nu} = \frac{\mathbf{1}_n}{n} \tag{9.8}$$

$$\forall i, \mu_i \left(y_i[\mathbf{K}\boldsymbol{\alpha}]_i + \xi_i - 1\right) = 0 \qquad \Longrightarrow \qquad \mu_i = 0 \vee y_i[\mathbf{K}\boldsymbol{\alpha}]_i = 1 - \xi_i \tag{9.9}$$

$$\forall i, \nu_i \xi_i = 0 \qquad \Longrightarrow \qquad \nu_i = 0 \vee \xi_i = 0. \tag{9.10}$$

If we plug $\boldsymbol{\alpha}$ solved by KKT condition (9.7) into complementary slackness conditions (9.9) and (9.10), in terms of $\boldsymbol{\alpha}$ they can be rewritten as:

$$\begin{cases} \alpha_i[y_i f(\mathbf{x}_i) + \xi_i - 1] = 0, & (9.11) \\ (\alpha_i - \dfrac{y_i}{2n\lambda})\xi_i = 0. & (9.12) \end{cases}$$

Note that here we use the fact that $f(\mathbf{x}_i) = [\mathbf{K}\boldsymbol{\alpha}]_i$. From (9.7) and (9.8), we know that $\alpha_i = \mu_i y_i / 2\lambda$, $\mu_i + \nu_i = 1/n$, $\forall i \in [n]$, also recall that $\mu_i, \nu_i \geq 0$, $\forall i \in [n]$. Then it is obvious to obtain the following:

$$o \leq \mu_i \leq \frac{1}{n}, \ \forall i \in [n];$$

$$0 \leq y_i \alpha_i \leq \frac{1}{2n\lambda}, \ \forall i \in [n].$$

Now we can analyze support vectors as follows:

- If $y_i\alpha_i = 0$, then $\alpha_i = 0$ and $\xi_i = 0$, this implies $y_i f(\mathbf{x}_i) \geq 1$ and $\mathbf{x}_i$ is not a support vector.
- If $0 < y_i\alpha_i < 1/2n\lambda$, then $\alpha_i \neq 0$ and $\xi_i = 0$, this implies $y_i f(\mathbf{x}_i) = 1$, thus $\mathbf{x}_i$ lies on a marginal hyperplane as in the separable case.
- If $y_i\alpha_i = 1/2n\lambda$, then $\alpha_i \neq 0$, this implies $y_i f(\mathbf{x}_i) = 1 - \xi_i$ where $\xi_i \geq 0$, thus $\mathbf{x}_i$ either lies on a marginal hyperplane or is an outlier.

[1] Either $y_i[\mathbf{K}\boldsymbol{\alpha}]_i \geq 1$ and $\xi_i = 0$, or $y_i[\mathbf{K}\boldsymbol{\alpha}]_i < 1$ and $\xi_i > 0$ takes the value satisfying $y_i[\mathbf{K}\boldsymbol{\alpha}]_i = 1 - \xi_i$.

Dual optimization problem

Convex optimization problem (9.6) is a classical QP (minimization of a convex quadratic function with linear constraints) for which any out-of-the-box optimization package can be used. However, the dimension of the problem (9.6) and the number of constraints are $2n$ where n is the number of points. General-purpose QP solvers will have difficulties when n exceeds a few thousands. Solving the dual of this problem (also a QP) will be more convenient and lead to faster algorithms due to the sparsity of the final solution. We therefore obtain the Lagrange dual function:

$$\begin{aligned} q(\boldsymbol{\mu}, \boldsymbol{\nu}) &= \inf_{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\xi} \in \mathbb{R}^n} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\nu}) \\ &= \begin{cases} \boldsymbol{\mu}^\top \mathbf{1}_n - \frac{1}{4\lambda} \boldsymbol{\mu}^\top \mathrm{diag}(\mathbf{y}) \mathbf{K} \mathrm{diag}(\mathbf{y}) \boldsymbol{\mu} & \text{if } \boldsymbol{\mu} + \boldsymbol{\nu} = \frac{1}{n}, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Thus the dual problem is:

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\nu} \in \mathbb{R}^n} q(\boldsymbol{\mu}, \boldsymbol{\nu})$$
$$\text{subject to} \quad \boldsymbol{\mu} \geq \mathbf{0}, \boldsymbol{\nu} \geq \mathbf{0}.$$

Now we consider the following two cases:

- If $\mu_i > 1/n$ for some $i \in [n]$, then there is no $\nu_i \geq 0$ such that $\mu_i + \nu_i = 1/n$, hence $q(\boldsymbol{\mu}, \boldsymbol{\nu}) = -\infty$.
- If $0 \leq \mu_i \leq 1/n$, $\forall i \in [n]$, then the dual function takes finite values that depend only on $\boldsymbol{\mu}$ by taking $\nu_i = 1/n - \mu_i$.

The dual problem is therefore equivalent to:

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \boldsymbol{\mu}^\top \mathbf{1}_n - \frac{1}{4\lambda} \boldsymbol{\mu}^\top \mathrm{diag}(\mathbf{y}) \mathbf{K} \mathrm{diag}(\mathbf{y}) \boldsymbol{\mu}$$
$$\text{subject to} \quad \mathbf{0} \leq \boldsymbol{\mu} \leq \frac{\mathbf{1}_n}{n}.$$

Recall that we derive $\boldsymbol{\alpha}$ in terms of $\boldsymbol{\mu}$ in KKT condition (9.7), $\boldsymbol{\alpha} = \frac{\mathrm{diag}(\mathbf{y})\boldsymbol{\mu}}{2\lambda}$. Then we can directly plug this into the dual problem to obtain the QP that $\boldsymbol{\alpha}$ must solve:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} 2 \sum_{i=1}^n \alpha_i y_i - \sum_{i,j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = 2\boldsymbol{\alpha}^\top \mathbf{y} - \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \tag{9.13}$$
$$\text{subject to} \quad 0 \leq y_i \alpha_i \leq \frac{1}{2n\lambda}, \ \forall i \in [n].$$

In particular, if we set $\lambda = \frac{1}{2nC}$, then we get the formulation:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} 2 \sum_{i=1}^n \alpha_i y_i - \sum_{i,j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{9.14}$$
$$\text{subject to} \quad 0 \leq y_i \alpha_i \leq C, \ \forall i \in [n].$$

The formulation (9.14) is exactly the SVM optimization problem with kernels, and it is often called C-SVM.

□