<u>עבודה 3 – פתרון</u>

שאלה 3.

א. נוכיח כי הפונקציה $K(x,x') \coloneqq \big(2x(7)+x(3)\big)\cdot x'(2)$ לא יכולה להיות פונקציית קרנל לאף מרחב תכונות.
נניח בשלילה כי קיים מרחב תכונות ψ עבורו פונקציה זו הינה פונקציית הקרנל שלו.
כלומר לכל $x,x' \in \mathcal{X}$ $K(x,x') = \langle \psi(x), \psi(x') \rangle$.
$\mathcal{X} = \mathbb{R}^d$ ולכן עבור $x = \left(0,0,1,0,0,0,2,\underbrace{0,\dots,0}_{d-7}\right), x' = \left(0,1,\underbrace{0,\dots,0}_{d-2}\right) \in \mathcal{X}$ נקבל כי

$$K(x,x') = (2\cdot 1+2)\cdot 1 = 4$$
$$K(x',x) = (0+0)\cdot 0 = 0$$

אבל $K(x,x') = \langle \psi(x), \psi(x') \rangle \overset{\text{הרמטיות}}{=} \langle \psi(x'), \psi(x) \rangle = K(x',x)$ והגענו לסתירה בכך הראינו ששוויון זה לא מתקיים.

ב. נוכיח כי הפונקציה $K(x,x') \coloneqq 5-(x(1)-x(2)(x'(1)-x'(2))$ לא יכולה להיות פונקציית קרנל לאף מרחב תכונות.
נניח בשלילה כי קיים מרחב תכונות ψ עבורו פונקציה זו הינה פונקציית הקרנל שלו.
כלומר לכל $x,x' \in \mathcal{X}$ $K(x,x') = \langle \psi(x), \psi(x') \rangle$.
יהי $x \in \mathcal{X} = \mathbb{R}^d$ כך ש $x = \left(10,\underbrace{0,\dots,0}_{d-1}\right)$,
אזי $K(x,x) = 5-\big(x(1)-x(2)\big)^2 = 5-100 = -95 < 0$
אבל $K(x,x) = \langle \psi(x), \psi(x) \rangle = \big|\big|\psi(x)\big|\big|^2$
והגענו לסתירה שכן נורמה הינה תמיד גדולה שווה מ-0.

ג. נראה כי
$f(x,x') = \big(x(1)x'(1)\big)^6 + e^{x(3)+x(5)+x'(3)+x'(5)} + \frac{1}{x(1)x'(1)} + (x(4)+x(6))(x'(4)+x'(6))$
הינה פונקציית קרנל.

$$\psi(x) = \left(x(1)^6, e^{x(3)+x(5)}, \frac{1}{x(1)}, x(4)+x(6)\right)$$

נראה כי עבור $x,x' \in X$ $f(x,x') = \langle \psi(x), \psi(x') \rangle$

$$\langle \psi(x), \psi(x') \rangle = \left\langle \left(x(1)^6, e^{x(3)+x(5)}, \frac{1}{x(1)}, x(4)+x(6)\right), \left(x'(1)^6, e^{x'(3)+x'(5)}, \frac{1}{x'(1)}, x'(4)+x'(6)\right) \right\rangle$$
$$= x(1)^6 x'(1)^6 + e^{x(3)+x(5)+x'(3)+x'(5)} + \frac{1}{x(1)x'(1)} + \big(x(4)+x(6)\big)\big(x'(4)+x'(6)\big) =$$
$$\big(x(1)x'(1)\big)^6 + e^{x(3)+x(5)+x'(3)+x'(5)} + \frac{1}{x(1)x'(1)}\big(x(4)+x(6)\big)\left(x'(4)+x'(6)\right)$$
$$= f(x,x')$$

שאלה 4.

a.

Consider $k = 1$, $a_1 = -1$, and a convex function f_1. Then, $g(u) = -f_1(u)$

Assume that g is a convex function, then

$$g(\alpha u + (1-\alpha)v) \leq \alpha g(u) + (1-\alpha)g(v)$$

which implies

$$-f_1(\alpha u + (1-\alpha)v) \leq -\alpha f_1(u) - (1-\alpha)f_1(v)$$

and therefore

$$f_1(\alpha u + (1-\alpha)v) \geq \alpha f_1(u) + (1-\alpha)f_1(v)$$

which contradicts the convexity of f_1 unless f_1 is linear and therefore

$$f_1(\alpha u + (1-\alpha)v) = \alpha f_1(u) + (1-\alpha)f_1(v)$$

for all $\alpha \in [0,1],\ u, v \in \mathbb{R}^d$.

This proves that g is not necessarily convex if $a_i \in \mathbb{R}$.

So, every non-linear convex function will be counterexample e.g.

$$f_1(u) = u^2, a_1 = -1, k = 1$$

b.

Now $g(u) = \sum_{i=1}^{k} a_i f_i(u)$ is with convex f_i and $a_i \geq 0$ for all $i \in \{1, \dots, k\}$.

Then, for all $\alpha \in [0,1],\ u, v \in \mathbb{R}^d$, we have

$$g(\alpha u + (1-\alpha)v) = \sum_{i=1}^{k} a_i f_i(\alpha u + (1-\alpha)v)$$

$$\leq \sum_{i=1}^{k} a_i\big(\alpha f_i(u) + (1-\alpha)f_i(v)\big) = (*)$$

where the inequality is due to the convexity of f_i and the non-negativity of a_i for all i.

Then,

$$(*) = \alpha \sum_{i=1}^{k} a_i f_i(u) + (1-\alpha) \sum_{i=1}^{k} a_i f_i(v) = \alpha g(u) + (1-\alpha)g(v)$$

To conclude, we got

$$g(\alpha u + (1-\alpha)v) \leq \alpha g(u) + (1-\alpha)g(v)$$

for all $i \in [0,1],\ u, v \in \mathbb{R}^d$, and therefore g is a convex function.

שאלה 5.

1. The optimization can be written as

$$\underset{w\in\mathbb{R}^d}{\text{minimize}}\ \lambda\|w-v\|_2^2+\|\mathbf{X}^Tw-y\|_2^2.$$

We are asked to formulate a closed-form expression for the w that solves the optimization problem, namely, to formulate $\widehat{w}$ where

$$\widehat{w}=\underset{w\in\mathbb{R}^d}{\text{argmin}}\ \lambda\|w-v\|_2^2+\|\mathbf{X}^Tw-y\|_2^2.$$

Due to the convexity of the minimization problem, the optimal solution $\widehat{w}$ satisfies the optimality condition

$$\lambda\nabla\|w-v\|_2^2+\nabla\|\mathbf{X}^Tw-y\|_2^2=0.$$

The gradients can be formulated as

$$\begin{aligned}&\nabla\|w-v\|_2^2=2\left(w-v\right)\\&\nabla\|\mathbf{X}^Tw-y\|_2^2=2\mathbf{X}\left(\mathbf{X}^Tw-y\right)\end{aligned}$$

and therefore the optimality condition can be developed as follows

$$\begin{aligned}&\lambda\cdot 2\left(w-v\right)+2\mathbf{X}\left(\mathbf{X}^Tw-y\right)=0\\\rightarrow\ &\lambda w+\mathbf{X}\mathbf{X}^Tw=\lambda v+\mathbf{X}y\\\rightarrow\ &\left(\lambda I+\mathbf{X}\mathbf{X}^T\right)w=\lambda v+\mathbf{X}y\\\rightarrow\ &w=\left(\lambda I+\mathbf{X}\mathbf{X}^T\right)^{-1}\left(\lambda v+\mathbf{X}y\right)\end{aligned}$$

2. The gradient of the optimization objective is

$$\lambda\nabla\|w-v\|_2^2+\nabla\|\mathbf{X}^Tw-y\|_2^2.$$

Hence, the Gradient Descent update step is

$$w^{(t+1)}\leftarrow w^{(t)}-\eta\left(\lambda\nabla\|w-v\|_2^2+\nabla\|\mathbf{X}^Tw-y\|_2^2\right).$$

that can be explicitly formulated as

$$w^{(t+1)}\leftarrow w^{(t)}-2\eta\left(\lambda\left(w-v\right)+\mathbf{X}\left(\mathbf{X}^Tw-y\right)\right).$$

3. Note that we can write the optimization as

$$\begin{aligned}\widehat{w} &= \operatorname*{argmin}_{w\in\mathbb{R}^d} \lambda\|w-v\|_2^2 + \sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2 \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^d} \lambda\|w-v\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} m(\langle w, x_i\rangle - y_i)^2\end{aligned}$$

The gradient of the entire optimization objective is

$$\lambda\nabla\|w-v\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} m\nabla(\langle w, x_i\rangle - y_i)^2$$

but for SGD the averaging over the squared loss of the training examples (here scaled by m) is approximated by the gradient of the loss of a single training example that is uniformly randomly drawn from the sample in each SGD iteration. Namely, the SGD iteration is to draw a random i uniformly from $\{1,\ldots,m\}$ and to update according to

$$w^{(t+1)} \leftarrow w^{(t)} - 2\eta\left(\lambda\nabla\|w-v\|_2^2 + m\nabla(\langle w, x_i\rangle - y_i)^2\right)$$

that can be explicitly formulated as

$$w^{(t+1)} \leftarrow w^{(t)} - 2\eta\left(\lambda\left(w-v\right) + m x_i(\langle w, x_i\rangle - y_i)\right).$$

שאלה 6.

1. There is no error, because $x_t(3)$ and $x_t(4)$ can be expressed as linear combinations of $x_t(1)$ and $x_t(2)$.

$$x_t(3) = 3x_t(1) + x_t(2),\ x_t(4) = 2x_t(2) - 4x_t(3) = -2x_t(2) - 12x_t(1). \tag{1}$$

   We would like to show that $X^{\mathrm{T}}X \in \mathbb{R}^{4\times 4}$ has two 0 eigenvalues, where $X \in \mathbb{R}^{m\times 4}$ is the data matrix. Recall that the Rank of a matrix is the dimension of the vector space spanned by its columns, or equivalently, the dimension of the vector space spanned by its rows.

   From *Eq.* (1), we know that the dimension of the column space of X is at most 2, as the maximal set of linearly independent column vectors is at most 2. Similarly, the dimension of the row space of X^{T} is at most 2. We get that $\mathrm{Rank}(X^{\mathrm{T}}), \mathrm{Rank}(X) \leq 2$.

   We know also that $\mathrm{Rank}(AB) \leq \min(\mathrm{Rank}(A), \mathrm{Rank}(B))$, this is elementary claim from linear algebra. So we have $\mathrm{Rank}(X^{\mathrm{T}}X) \leq 2$. The "Rank-nullity theorem" asserts that for a matrix $A : \mathbb{R}^n \to \mathbb{R}^n$, $n = \mathrm{Rank}(A) + \dim(\mathrm{Ker}(A))$. We conclude that $\dim\big(\mathrm{Ker}\big(X^{\mathrm{T}}X\big)\big) \geq 2$. Moreover, the eigenspace corresponding to the 0 eigenvalue is exactly the Kernel space (or null space) of a linear transformation. So the dimension of the 0 eigenspace is at least 2, and $X^{\mathrm{T}}X$ has at least two 0 eigenvalues.

2. Take $x_1 = (1,0,1,0)\,, x_2 = (0,1,1,0)\,, x_3 = (-1,-1,0,1)$, then

$$X^{\mathrm{T}}X = \begin{pmatrix} 2 & 1 & 1 & -1 \\ 1 & 2 & 1 & -1 \\ 1 & 1 & 2 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix}$$

   and the eigenvalues are $0, 1, -\sqrt{2}+3, \sqrt{2}+3$. The error is equal to the smallest $d-k$ eigenvalues, here $d = 4$ and $k = 2$, so the distortion is 1.