HW 1 for CSCI 5525

YangyangLi
li002252@umn.edu

1 Problem 1

1.1 a

Show that $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B})$ and so $\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{M}^\intercal \boldsymbol{M})}$
PROOF:

$\because \langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{ij} a_{ij} b_{ij}$ and $\boldsymbol{A}$ and $\boldsymbol{B}$ have same size.

$\therefore \langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_j \boldsymbol{a}_j^\intercal \boldsymbol{b}_j$, where $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ are column vector of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.

$\because \boldsymbol{A}^\intercal \boldsymbol{B} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_j]^\intercal [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_j]$ and $\operatorname{tr}(\boldsymbol{M}) = \sum_i m_{ii}$.

$$\therefore \boldsymbol{A}^\intercal \boldsymbol{B} = \begin{bmatrix} \boldsymbol{a}_1^\intercal \boldsymbol{b}_1 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \boldsymbol{a}_2^\intercal \boldsymbol{b}_2 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \boldsymbol{a}_j^\intercal \boldsymbol{b}_j \end{bmatrix}$$

$\therefore \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \sum_j \boldsymbol{a}_j^\intercal \boldsymbol{b}_j = \langle \boldsymbol{A}, \boldsymbol{B} \rangle$ and $\operatorname{tr}(\boldsymbol{M}^\intercal \boldsymbol{M}) = \langle \boldsymbol{M}, \boldsymbol{M} \rangle$

Summarizing:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) \tag{1}$$

$$\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{M}^\intercal \boldsymbol{M})} \tag{2}$$

1.2 b

Show that $\operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$
PROOF:

$\because \langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) \, and \, \langle \boldsymbol{A}, \boldsymbol{B} \rangle = \langle \boldsymbol{B}, \boldsymbol{A} \rangle$
$\therefore \operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \langle \boldsymbol{A}, \boldsymbol{B} \rangle = \langle \boldsymbol{B}, \boldsymbol{A} \rangle = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$

Summarizing:

$$\operatorname{tr}(\boldsymbol{A}^\intercal \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A}) \tag{3}$$

1.3 c

Assume $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same size. In general, $\boldsymbol{A}\boldsymbol{B}^\intercal$ and $\boldsymbol{B}^\intercal \boldsymbol{A}$ have different sizes, but $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \operatorname{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$. Show it.

PROOF:

$\because$ $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \langle \boldsymbol{A}^\intercal, \boldsymbol{B}^\intercal \rangle = \sum_{ji} a_{ji} b_{ji}$, where a_{ji} and b_{ji} are (j, i)-th element of $\boldsymbol{A}^\intercal$ and $\boldsymbol{B}^\intercal$ respectively.

$\because$ $\mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{A}) = \langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{ij} a_{ij} b_{ij}$, where a_{ij} and b_{ij} are (i, j)-th element of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.

$\therefore$ $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{A})$

Summarizing:

$$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{A}) \tag{4}$$

1.4 d

Show that $\mathrm{tr}(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{M}_3) = \mathrm{tr}(\boldsymbol{M}_3\boldsymbol{M}_1\boldsymbol{M}_2) = \mathrm{tr}(\boldsymbol{M}_2\boldsymbol{M}_3\boldsymbol{M}_1)$, assuming that the sizes of $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ are compatible with all the matrix multiplications. This is known as the *cyclic property* of matrix traces. (Hint: think of (c))
PROOF:

Provided that $\boldsymbol{M}_1\boldsymbol{M}_2 = \boldsymbol{A}$, $\boldsymbol{M}_3 = \boldsymbol{B}^\intercal$, according to the Eq. (4):

$$\mathrm{tr}(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{M}_3) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal) = \mathrm{tr}(\boldsymbol{B}^\intercal \boldsymbol{A}) = \mathrm{tr}(\boldsymbol{M}_3\boldsymbol{M}_1\boldsymbol{M}_2)$$

Similarly,

$$\mathrm{tr}(\boldsymbol{M}_3\boldsymbol{M}_1\boldsymbol{M}_2) = \mathrm{tr}(\boldsymbol{M}_2\boldsymbol{M}_3\boldsymbol{M}_1)$$

Summarizing:

$$\mathrm{tr}(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{M}_3) = \mathrm{tr}(\boldsymbol{M}_3\boldsymbol{M}_1\boldsymbol{M}_2) = \mathrm{tr}(\boldsymbol{M}_2\boldsymbol{M}_3\boldsymbol{M}_1) \tag{5}$$

1.5 e

For any matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ of compatible sizes, we always have $\langle \boldsymbol{A}\boldsymbol{C}\boldsymbol{B}, \boldsymbol{D} \rangle = \langle \boldsymbol{C}\boldsymbol{B}, \boldsymbol{A}^\intercal \boldsymbol{D} \rangle = \langle \boldsymbol{A}\boldsymbol{C}, \boldsymbol{D}\boldsymbol{B}^\intercal \rangle$, i.e., we can always move the **leading** matrix of one side of the inner product to the other side as **leading** matrix **once transposed** (if these matrices are complex-valued, should be conjugate transposed), and similarly the **trailing** matrix to the other side as **trailing** matrix once **transposed**. Why? (Hint: think of the above results and also try to remember this important property that will be useful for calculation later)
PROOF:

In terms of the Eq. (1):

$$\langle \boldsymbol{A}\boldsymbol{C}\boldsymbol{B}, \boldsymbol{D} \rangle = \langle \boldsymbol{D}, \boldsymbol{A}\boldsymbol{C}\boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{D}^\intercal \boldsymbol{A}\boldsymbol{C}\boldsymbol{B})$$

Provided that $\boldsymbol{D}^\intercal \boldsymbol{A} = \boldsymbol{M}_1$ and $\boldsymbol{C}\boldsymbol{B} = \boldsymbol{M}_2$. In terms of the Eq. (1):

$$\mathrm{tr}(\boldsymbol{D}^\intercal \boldsymbol{A}\boldsymbol{C}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{M}_1\boldsymbol{M}_2) = \langle \boldsymbol{M}_1^\intercal, \boldsymbol{M}_2 \rangle = \langle \boldsymbol{A}^\intercal \boldsymbol{D}, \boldsymbol{C}\boldsymbol{B} \rangle = \langle \boldsymbol{C}\boldsymbol{B}, \boldsymbol{A}^\intercal \boldsymbol{D} \rangle$$

Similarly, in terms of the Eq. (5):

$$\langle \boldsymbol{ACB}, \boldsymbol{D} \rangle = \langle \boldsymbol{D}, \boldsymbol{ACB} \rangle = \operatorname{tr}\left(\boldsymbol{D}^\intercal \boldsymbol{ACB}\right) = \operatorname{tr}\left(\boldsymbol{BD}^\intercal \boldsymbol{AC}\right) = \langle \boldsymbol{DB}^\intercal, \boldsymbol{AC} \rangle$$

Summarizing:

$$\langle \boldsymbol{ACB}, \boldsymbol{D} \rangle = \langle \boldsymbol{CB}, \boldsymbol{A}^\intercal \boldsymbol{D} \rangle = \langle \boldsymbol{AC}, \boldsymbol{DB}^\intercal \rangle \tag{6}$$

1.6 f

For $\boldsymbol{M}$, let's perform a *compact SVD* (if not sure, check up Wikipedia! https://en.wikipedia.org/wiki/Singular_value_decomposition#Compact_SVD) to obtain $\boldsymbol{M} = \boldsymbol{U\Sigma V}^\intercal$, so that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal (not necessarily square) matrices, i.e., $\boldsymbol{U}^\intercal \boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^\intercal \boldsymbol{V} = \boldsymbol{I}$. Use the cyclic property of trace and that $\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}\left(\boldsymbol{M}^\intercal \boldsymbol{M}\right)}$ to show that

$$\|\boldsymbol{M}\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2},$$

assuming the rank of $\boldsymbol{M}$ is r. Here σ_i's are the singular values of $\boldsymbol{M}$.
PROOF:

since $\boldsymbol{M} = \boldsymbol{U\Sigma V}^\intercal$, $\|\boldsymbol{M}\|_F = \sqrt{\operatorname{tr}\left(\boldsymbol{M}^\intercal \boldsymbol{M}\right)}$, $\boldsymbol{U}^\intercal \boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^\intercal \boldsymbol{V} = \boldsymbol{I}$ so,

$$\|M\|_F = \sqrt{\operatorname{tr}\left(\boldsymbol{V\Sigma U}^\intercal \boldsymbol{U\Sigma V}^\intercal\right)} = \sqrt{\operatorname{tr}\left(\boldsymbol{V\Sigma\Sigma V}^\intercal\right)} = \sqrt{\operatorname{tr}\left(\boldsymbol{V}^\intercal \boldsymbol{V\Sigma\Sigma}\right)} = \sqrt{\operatorname{tr}\left(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_r\right)} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$$

2 Problem 2

2.1 a

Let $\boldsymbol{A}$ be a square matrix. Deriving the gradient and Hessian of the quadratic function $f(\boldsymbol{x}) = \boldsymbol{x}^\intercal \boldsymbol{Ax} + \boldsymbol{b}^\intercal \boldsymbol{x}$. Please include your calculation details. (Hint: note that Hessian must be a symmetric matrix.
SOLUTION:

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{\sigma}) &= (\boldsymbol{x}+\boldsymbol{\sigma})^\intercal \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{\sigma}) + \boldsymbol{b}^\intercal(\boldsymbol{x}+\boldsymbol{\sigma}) \\ &= \left(\boldsymbol{x}^\intercal + \boldsymbol{\sigma}^\intercal\right)(\boldsymbol{Ax} + \boldsymbol{A\sigma}) + \boldsymbol{b}^\intercal \boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{\sigma} \\ &= \boldsymbol{x}^\intercal \boldsymbol{Ax} + \boldsymbol{x}^\intercal \boldsymbol{A\sigma} + \boldsymbol{\sigma}^\intercal \boldsymbol{Ax} + \boldsymbol{\sigma}^\intercal \boldsymbol{A\sigma} + \boldsymbol{b}^\intercal \boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{\sigma} \\ &= f(\boldsymbol{x}) + \boldsymbol{\sigma}^\intercal \boldsymbol{Ax} + \boldsymbol{x}^\intercal \boldsymbol{A\sigma} + \boldsymbol{b}^\intercal \boldsymbol{\sigma} + \boldsymbol{\sigma}^\intercal \boldsymbol{A\sigma} \end{aligned} \tag{7}$$

$\because \boldsymbol{\sigma}^\intercal \boldsymbol{Ax}$ is a scalar

$\therefore \boldsymbol{\sigma}^\intercal \boldsymbol{Ax} = \left(\boldsymbol{\sigma}^\intercal \boldsymbol{Ax}\right)^\intercal = \boldsymbol{x}^\intercal \boldsymbol{A}^\intercal \boldsymbol{\sigma}$

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{\sigma}) &= f(\boldsymbol{x}) + \left(\boldsymbol{x}^\intercal \boldsymbol{A}^\intercal + \boldsymbol{x}^\intercal \boldsymbol{A} + \boldsymbol{b}^\intercal\right)\boldsymbol{\sigma} + \boldsymbol{\sigma}^\intercal \boldsymbol{A\sigma} \\ &= f(\boldsymbol{x}) + \langle \boldsymbol{Ax} + \boldsymbol{A}^\intercal \boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\sigma} \rangle + \langle \boldsymbol{\sigma}, \boldsymbol{A\sigma} \rangle \end{aligned} \tag{8}$$

Consequently,

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^\intercal \boldsymbol{x} + \boldsymbol{b}$$

$$\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}$$

2.2 b

Let $p(\boldsymbol{x};\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}}}$. The log-likelihood for logistic regression with two classes is (assuming N samples of the form $(\boldsymbol{x}_i, y_i)$)

$$\begin{aligned} f(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left[y_i \log p(\boldsymbol{x}_i;\boldsymbol{\beta}) + (1-y_i)\log(1-p(\boldsymbol{x}_i;\boldsymbol{\beta})) \right] \\ &= \sum_{i=1}^{N} \left[y_i \boldsymbol{\beta}^\intercal \boldsymbol{x}_i - \log\left(1+e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}_i}\right)\right]. \end{aligned}$$

Derive the gradient and Hessian of $f(\boldsymbol{\beta})$. Please include your calculation details. (1/12) For logistic regression, we are going to maximize $f(\boldsymbol{\beta})$, which is equivalent to minimize $-f(\boldsymbol{\beta})$. Does the minimization problem has a unique minimizer or not?
SOLUTION:

denote $f(\boldsymbol{\beta})$ as matrix form,

$$\begin{aligned} f(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left[y_i \boldsymbol{\beta}^\intercal \boldsymbol{x}_i - \log\left(1+e^{\boldsymbol{\beta}^\intercal \boldsymbol{x}_i}\right)\right] \\ &= \boldsymbol{y}^\intercal \boldsymbol{X}\boldsymbol{\beta} - \mathbf{1}^\intercal \log\left(\mathbf{1} + e^{\boldsymbol{X}\boldsymbol{\beta}}\right) \end{aligned} \tag{9}$$

where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \boldsymbol{x}_2^\intercal \\ \vdots \\ \boldsymbol{x}_N^\intercal \end{bmatrix}$, $\boldsymbol{y} = [y_1, \dots, y_N]^\intercal$, $\mathbf{1} = [1, \dots, 1]$.

Since $f(\boldsymbol{\beta})$ map vector to scalar, $df = \sum_i \frac{\partial f}{\partial \beta_i} d\beta_i = \left(\frac{\partial f}{\partial \boldsymbol{\beta}}\right)^\intercal d\boldsymbol{\beta} = \nabla_{\boldsymbol{\beta}} f^\intercal d\boldsymbol{\beta}$

$$\begin{aligned} df &= \boldsymbol{y}^\intercal \boldsymbol{X} d\boldsymbol{\beta} - \frac{\mathbf{1}^\intercal \left(e^{\boldsymbol{X}\boldsymbol{\beta}} \circ d\boldsymbol{X}\boldsymbol{\beta}\right)}{\mathbf{1} + e^{\boldsymbol{X}\boldsymbol{\beta}}} \\ &= \boldsymbol{y}^\intercal \boldsymbol{X} d\boldsymbol{\beta} - \frac{\left(e^{\boldsymbol{X}\boldsymbol{\beta}}\right)^\intercal d\boldsymbol{X}\boldsymbol{\beta}}{\mathbf{1} + e^{\boldsymbol{X}\boldsymbol{\beta}}} \end{aligned} \tag{10}$$

Let $\sigma\left(a\right)=\frac{e^{a}}{1+e^{a}}$, $\sigma'\left(a\right)=\frac{e^{a}}{\left(1+e^{a}\right)^{2}}$,

$$\begin{aligned} df &= \boldsymbol{y}^{\intercal}\boldsymbol{X}d\boldsymbol{\beta}-\frac{\mathbf{1}^{\intercal}\left(e^{\boldsymbol{X}\boldsymbol{\beta}}\circ d\boldsymbol{X}\boldsymbol{\beta}\right)}{\mathbf{1}+e^{\boldsymbol{X}\boldsymbol{\beta}}} \\ &= \boldsymbol{y}^{\intercal}\boldsymbol{X}d\boldsymbol{\beta}-\frac{\left(e^{\boldsymbol{X}\boldsymbol{\beta}}\right)^{\intercal}\boldsymbol{X}d\boldsymbol{\beta}}{\mathbf{1}+e^{\boldsymbol{X}\boldsymbol{\beta}}} \\ &= \boldsymbol{y}^{\intercal}\boldsymbol{X}d\boldsymbol{\beta}-\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right)^{\intercal}\boldsymbol{X}d\boldsymbol{\beta} \\ &= \left(\boldsymbol{y}^{\intercal}-\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right)^{\intercal}\right)\boldsymbol{X}d\boldsymbol{\beta} \end{aligned} \tag{11}$$

So $\nabla f=\boldsymbol{X}^{T}\left(\boldsymbol{y}-\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)$. $\nabla^{2}f=\frac{\partial^{2}f}{\partial\beta\partial\beta}$, so,

$$\begin{aligned} d\nabla f &= d\left(\boldsymbol{X}^{\intercal}\left(\boldsymbol{y}-\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\right) \\ &= -\boldsymbol{X}^{\intercal}d\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right) \\ &= -\boldsymbol{X}^{\intercal}\left(\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)\circ d\boldsymbol{X}\boldsymbol{\beta}\right) \\ &= -\boldsymbol{X}^{\intercal}\operatorname{diag}\left(\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\boldsymbol{X}d\boldsymbol{\beta} \end{aligned} \tag{12}$$

Similarly,

$\nabla^{2}f=-\boldsymbol{X}^{\intercal}\operatorname{diag}\left(\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\boldsymbol{X}$, where $\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)=\frac{e^{\boldsymbol{X}\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}\boldsymbol{\beta}}\right)^{2}}$

Consequently,

$$\begin{gathered} \nabla f=\boldsymbol{X}^{T}\left(\boldsymbol{y}-\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right), \\ \nabla^{2}f=-\boldsymbol{X}^{\intercal}\operatorname{diag}\left(\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\boldsymbol{X} \end{gathered}$$

where $\sigma\left(\boldsymbol{X}\boldsymbol{\beta}\right)=\frac{e^{\boldsymbol{X}\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}\boldsymbol{\beta}}}$, $\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)=\frac{e^{\boldsymbol{X}\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}\boldsymbol{\beta}}\right)^{2}}$.

Also,

$$\begin{aligned} \nabla^{2}\text{-}f &= \boldsymbol{X}^{T}\operatorname{diag}\left(\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\boldsymbol{X} \\ &= \begin{bmatrix} \sum_{i}^{N}\lambda_{i}x_{i,1}^{2} & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \sum_{i}^{N}\lambda_{i}x_{i,M}^{2} \end{bmatrix} \end{aligned} \tag{13}$$

Where λ_i is diagonal entries of $\operatorname{diag}\left(\sigma'\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)$, and M is the number of column vector of $\boldsymbol{X}$. As $\sigma'\left(a\right)>0$, diagonal entries of ∇^{2}-f more than 0. So, ∇^{2}-f is positive definite.

Therefore, only one unique minimizer exists.

2.3 c

Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \leq n$, then given a $\boldsymbol{y} \in \mathbb{R}^m$, the least-squares problem $\min_{\boldsymbol{x}} \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ has infinitely many solutions. Now let's say we want a solution with a small ℓ_2 norm, then it is reasonable to put a penalty on the ℓ_2 norm:

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_2^2$$

with a chosen $\lambda > 0$. This is *ridge regression*. Now we know that for an unconstrained first-order differentiable function $g(\boldsymbol{x})$, any of its local minimizer $\boldsymbol{x}_*$ must satisfy the *first-order optimality condition*: $\nabla g(\boldsymbol{x}_*) = \boldsymbol{0}$. Use this to derive $\boldsymbol{x}_*$ (1/12). Is the $\boldsymbol{x}_*$ unique? Why?

SOLUTION:

$$\begin{aligned} g(\boldsymbol{x} + \boldsymbol{\sigma}) &= \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x} + \boldsymbol{\sigma})\|_2^2 + \lambda \|\boldsymbol{x} + \boldsymbol{\sigma}\|_2^2 \\ &= \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|x\|_2^2 + \lambda \|\sigma\|_2^2 + 2\lambda \langle \boldsymbol{x}, \boldsymbol{\sigma} \rangle - 2 \langle \boldsymbol{A}^\intercal (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}), \boldsymbol{\sigma} \rangle + \|\boldsymbol{A}\boldsymbol{\sigma}\|_2^2 \\ &= g(\boldsymbol{x}) + \boldsymbol{\sigma}^\intercal (\lambda \boldsymbol{I} + \boldsymbol{A}^\intercal \boldsymbol{A}) \boldsymbol{\sigma} + 2 (\lambda \boldsymbol{x}^\intercal - (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^\intercal \boldsymbol{A}) \boldsymbol{\sigma} \\ &= g(\boldsymbol{x}) + 2 \langle \lambda \boldsymbol{x} - \boldsymbol{A}^\intercal (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}), \boldsymbol{\sigma} \rangle + \langle \boldsymbol{\sigma}, (\lambda \boldsymbol{I} + \boldsymbol{A}^\intercal \boldsymbol{A}) \boldsymbol{\sigma} \rangle \end{aligned} \tag{14}$$

So,

$$\nabla g = 2 (\lambda \boldsymbol{x} - \boldsymbol{A}^\intercal \boldsymbol{y} + \boldsymbol{A}^\intercal \boldsymbol{A}\boldsymbol{x})$$
$$\nabla^2 g = 2 (\lambda \boldsymbol{I} + \boldsymbol{A}^\intercal \boldsymbol{A})$$

Let $\nabla g(\boldsymbol{x}_*)) = 0$,

$$\boldsymbol{x}_* = (\lambda \boldsymbol{I} + \boldsymbol{A}^\intercal \boldsymbol{A})^{-1} \boldsymbol{A}^\intercal \boldsymbol{y}$$

$\boldsymbol{A}^\intercal \boldsymbol{A}$ is positive definite in that its diagonal entries are all positive. Hence, $\nabla^2 g = 2 (\lambda \boldsymbol{I} + \boldsymbol{A}^\intercal \boldsymbol{A})$ is positive definite as well, which indicates $\boldsymbol{x}_*$ is unique.

3 Problem 3

The solution for this problem has been finished in the Colab, and I will upload this notebook *HW1_CS5525_YangyangLi.ipynb* as well. Feel free to check that.

Have a good day!