

5525 HW1

Problem 1

a) $\langle A, B \rangle = \text{tr}(A^T B)$

$$= \sum_i (A^T B)_{ii}$$

$$= \sum_i \sum_k A_{ki} B_{ik}$$

$$= \sum_i \sum_j a_{ij} b_{ij}$$

$$\langle A, B \rangle = \sum_{i,j} a_{ij} b_{ij}$$

$$\text{tr}(M) = \sum_i m_{ii}$$

$$(A^T B)_{ij} = \sum_k A_{ik}^T B_{kj}$$

$$(A^T B)_{ii} = \sum_k A_{ik}^T B_{ki}$$

$$A_{ik}^T = A_{ki}$$

$$(A^T B)_{ii} = \sum_k A_{ki} B_{ki}$$

$$\|M\|_F = \sqrt{\text{tr}(M^T M)}$$

$$= \sqrt{\sum_i \sum_k m_{ki}^2}$$

$$= \sqrt{\sum_{ij} m_{ij}^2}$$

$$\|M\|_F = \sqrt{\sum_{ij} m_{ij}^2}$$

$$(M^T M)_{ij} = \sum_k m_{ik}^T m_{kj}$$

$$(M^T M)_{ii} = \sum_k m_{ik}^T m_{ki}$$

$$(M^T M)_{ii} = \sum_k m_{ki} m_{ki}$$

$$= \sum_k m_{ki}^2$$

b) $\text{tr}(A^T B) = \text{tr}(B^T A)$

$$(A^T B)_{ij} = \sum_k a_{ik}^T b_{kj}$$

$$(A^T B)_{ij} = \sum_k a_{ki} b_{kj}$$

$$\therefore (B^T A)_{ij} = \sum_k b_{ki} a_{kj}$$

$$a_{ik}^T = a_{ki}$$

$$\text{trace}(A^T B) = \sum (A^T B)_{ii}$$

$$\therefore \text{tr}(B^T A) = \sum (B^T A)_{ii}$$

$$\therefore \sum (A^T B) = \sum (B^T A)$$

$$\sum_i \sum_k a_{ki} b_{kj} = \sum_i \sum_k b_{ki} a_{kj}$$

$$\therefore \sum_{i,k} a_{ki} b_{kj} = \sum_{i,k} a_{kj} b_{ki}$$

$$\begin{aligned}
 c) \text{tr}(AB^T) &= \text{tr}(B^TA) \\
 &= \sum_i (B^TA)_{ii} \\
 &= \sum_i \sum_k b_{ik}^T a_{ki} \\
 &= \sum_i \sum_k a_{ik} b_{ki}^T \\
 &= \sum_i (AB^T)_{ii} \\
 &= \text{tr}(AB^T)
 \end{aligned}$$

Using $(B^TA)_{ij} = \sum_k b_{ik}^T a_{kj}$

$$(AB^T)_{ij} = \sum_k a_{ik} b_{kj}^T$$

$$\text{tr}(M) = \sum_i m_{ii}$$

we know this to be true as the sum over both i & k will be equal as matrix element multiplication is commutative

$$\begin{aligned}
 d) \text{tr}(M_1 M_2 M_3) &= \text{tr}(M_3 M_1 M_2) = \text{tr}(M_2 M_3 M_1) \\
 \text{tr}(M_1 M_2 M_3) &= \sum_i (M_1 M_2 M_3)_{ii} \\
 &= \sum_i \sum_j M_{1,ij} M_{2,j} M_{3,j} \\
 &= \sum_i \sum_j \sum_k M_{1,ij} M_{2,jk} M_{3,k} \quad * \text{ based on part c} \\
 &= \sum_i \sum_j \sum_k M_{1,ki} M_{2,ij} M_{3,jk} = \sum_i \sum_j \sum_k M_{2,jk} M_{3,kj} M_{1,ij} \\
 &= \sum_i (M_3 M_1 M_2) \\
 &= \text{tr}(M_2 M_3 M_1)
 \end{aligned}$$

$$e) \langle A C B, D \rangle = \langle A C, D B^T \rangle$$

As shown before: $\langle A C B, D \rangle = \text{tr}((ACB)^T D)$

$$\text{tr}((ACB)^T D) =$$

$$\text{tr}(B^T C^T A^T D) =$$

$$\text{tr}(C^T A^T D B^T) =$$

$$\text{tr}(A C^T D B^T) =$$

$$\langle A C, D B^T \rangle = \langle A C B, D \rangle = \langle C B, A^T D \rangle$$

$$\text{tr}((ACB)^T D) =$$

$$\text{tr}(B^T C^T A^T D) =$$

$$\text{tr}((CB)^T (A^T D)) =$$

$$\langle C B, A^T D \rangle$$

$$(AB)^T = B^T A^T$$

As shown from part D

$$f) \text{ Goal: } \|M\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$$

$$\|M\|_F = \sqrt{\text{tr}(M^T M)}$$

$$= \sqrt{\text{tr}((U\Sigma V^T)^T (U\Sigma V^T))}$$

$$= \sqrt{\text{tr}(V\Sigma^T U^T U\Sigma V^T)}$$

$$= \sqrt{\text{tr}(V\Sigma^T I \Sigma V^T)}$$

$$= \sqrt{\text{tr}(V\Sigma^2 V^T)}$$

$$= \sqrt{\text{tr}(V\Sigma^2 V^T)}$$

$$= \sqrt{\text{tr}(V^T V \Sigma^2)}$$

$$= \sqrt{\text{tr}(\Sigma^2)}$$

Given

$$\|M\|_F = \sqrt{\text{tr}(M^T M)}$$

$$M = U\Sigma V^T$$

σ_i are diagonal entries of Σ

$$\sigma_i = \sqrt{\sum_{j=1}^r \sigma_{ij}^2}$$

* each diagonal element is σ_i^2

Problem 2

$$a) f(x) = x^T A x + b^T x$$

$$\frac{\partial f(x)}{\partial x} = \frac{\partial}{\partial x} (x^T A x + b^T x)$$

$$\text{Gradient} = 2Ax + b$$

Hessian

$$= 2A(x + \delta) + b$$

$$= 2Ax + 2A\delta + b$$

$$= 2Ax + b + 2A\delta$$

$$\nabla f(x)$$

$\therefore \nabla^2 f(x) = 2A$ because the Hessian is symmetric
over something like $A + A^T$

$$b) p(x; B) = \frac{e^{B^T x}}{1 + e^{B^T x}}$$

$$f(B) = \sum_{i=1}^N [y_i B^T x_i - \log(1 + e^{B^T x_i})]$$

Gradient $\nabla_B (y_i B^T x_i) = y_i x_i$ $\nabla_B (\log(1 + e^{B^T x_i})) = \frac{e^{B^T x_i}}{1 + e^{B^T x_i}} x_i$

$$= p(x_i; B) x_i$$

$$\nabla_B f(B) = \sum_{i=1}^N (y_i - p(x_i; B)) x_i$$

Hessian

$$\nabla_B (y_i x_i) = 0$$

$$\nabla_B (-p(x_i; B) x_i) = \frac{\partial p(x_i; B)}{\partial B} x_i^T$$

$$\frac{\partial p(x_i; B)}{\partial B} = p(x_i; B)(1 - p(x_i; B)) x_i$$

$$\nabla_B^2 f(B) = \sum_{i=1}^N p(x_i; B)(1 - p(x_i; B)) x_i x_i^T$$

Minimize $-f(B)$

$$\text{If } \sum_{i=1}^N p(x_i; B)(1 - p(x_i; B)) x_i x_i^T \geq 0 \text{ for every}$$

vector then it is convex & positive semi-definite. However, it may not have a unique minimizer unless it is positive definite & strictly convex. This is the case when

$$\sum_{i=1}^N p(x_i; B)(1 - p(x_i; B)) x_i x_i^T > 0 \text{ for every vector.}$$

This only happens when every x_i is independent & span the feature matrix X . If this is the case then there is a unique minimizer.

$$c) g(x) = \|y - Ax\|_2^2 + \lambda \|x\|_2^2$$

$$\|v\|_2^2 = v^T v$$

$$= (y - Ax)^T (y - Ax) + \lambda x^T x$$

$$= y^T y - 2y^T A x + x^T A^T A x + \lambda x^T x$$

$$\nabla g(x) = 0 - 2A^T y + 2A^T A x + 2\lambda x$$

$$= 2(-A^T y + A^T A x + \lambda x)$$

Solve for $\nabla g(x) = 0$

$$0 = -A^T y + A^T A x + \lambda x$$

$$A^T y = A^T A x + \lambda x$$

$$A^T y = x(A^T A + \lambda I) \quad \text{local minimizer}$$

$$(A^T A + \lambda I)^{-1} A^T y = x_*$$

Local minimizer is anytime the function $\nabla g(x) = 0$, so x_* is a local minimizer.

Global minimizer

- $A^T A + \lambda I$ is positive definite because $\lambda > 0$

showing all eigenvalues are > 0

- Because of this the function is also a global minimizer.

Uniqueness

- $A^T A + \lambda I$ is invertible, proving x_* is also unique.

Problem 3

a) Show A has full-rank $\Leftrightarrow A^T A$ is invertible

A is full-rank $\rightarrow A^T A$ is invertible

$Az = 0, z = 0$ when A is full-rank

$A^T A z = 0, z = 0$ when A is invertible

$$z^T A^T A z = 0$$

$$(Az)^T (Az) = 0$$

$$\|Az\|_2^2 = 0$$

$$\therefore Az = 0$$

$A^T A$ is invertible $\rightarrow A$ is full rank

$Az = 0$ * suppose A is not full rank

$$A^T A z = 0$$

$$A^T 0 = 0$$

$\therefore z$ is in null-space of $A^T A$

However $A^T A z = 0$, when $z \neq 0$ as it is invertible
creating a contradiction that $z \neq 0$, so A
must have full rank.

b) $\min_{z \in \text{span}(A)} g(z) = \|y - z\|_2^2 \quad z_0 = Ax_0$

$$\min_{x \in \mathbb{R}^n} \|y - Ax_0\|_2^2$$

$$f(x) = \|y - Ax\|_2^2 \quad \text{solve for } x$$

$$\nabla f(x) = -2A^T(y - Ax) = 0$$

$$A^T y = A^T A x \rightarrow \text{normal eqn.}$$

$$x = (A^T A)^{-1} A^T y$$

$$z_0 = A((A^T A)^{-1} A^T y) \rightarrow \text{closed form of } z_0$$

It is unique, because A is full rank to show $A^T A$ is invertible. It means the normal equation, $A^T A x_0 = A^T y$, is unique for x_0 which consequently means z_0 .

Show $\langle z_0 - y, w \rangle = 0$, where $w \in \text{span}(A)$

$$\langle Ax_0 - y, w \rangle = 0$$

$$\langle Ax_0 - y, Ax_w \rangle = 0, w = Ax_w \text{ as some } x_w \in \mathbb{R}^n \text{ so that } w = Ax_w$$

$$(Ax_0 - y)^T Ax_w = 0$$

$$x_w^T A^T (Ax_0 - y) = 0$$

We already know $A^T (Ax_0 - y) = 0$
from the normal equation earlier

$\therefore \langle z_0 - y, w \rangle = 0$ showing $z_0 - y$ is orthogonal to $\text{span}(A)$

c) See submitted code

d) From before: $A^T A x_0 = A^T y$ $\ker(A) = \{z \in \mathbb{R}^n \mid Az = 0\}$
(global min.)

Let x be another global min \rightarrow This implies $x - x_0 \in \ker(A^T A)$

$$A^T A x = A^T y$$

$$A^T A(x - x_0) = 0$$

or the difference is in the null space.

Because $A^T A$ has the same null space as A

$$x - x_0 \in \ker(A^T A)$$

$$\therefore \{x_0 + z \mid z \in \ker(A)\}$$

Proving $\{x_0 + z \mid z \in \ker(A)\}$

Let x_0 = global min

Let $x_1 = x_0 + z$ that is global min

$$A^T A x_1 = A^T A(x_0 + z) = A^T A x_0 + A^T A z = A^T A x_0 = A^T y$$

$\therefore x_1$ satisfies normal eqn. and is a global min
proving the set has a global min

$$e) (i) \nabla f(x) = 2A^T(Ax - y) \quad \text{Grad. descent: } x^{(k+1)} = x^{(k)} - \gamma \nabla f(x^{(k)})$$

$$x^0 = 0$$

$$\begin{aligned} x^1 &= x^0 - \gamma A^T(Ax^0 - y) \\ &= 0 + \gamma A^T y \end{aligned}$$

$$\begin{aligned} x_2 &= x_1 - \gamma A^T(Ax^1 - y) \\ &= \gamma A^T y - \gamma A^T(\gamma A^T y - y) \\ &= \gamma A^T y - 4\gamma^2 A^T A A^T y + \gamma A^T y \\ &= 4\gamma A^T y - 4\gamma^2 A^T A A^T y \end{aligned}$$

$$\begin{aligned} x^3 &= x^2 - \gamma A^T(Ax^2 - y) \\ &= 4\gamma A^T y - \gamma A^T(\gamma A^T y - 4\gamma^2 A^T A A^T y - y) \\ &\quad - 4\gamma^2 A^T A A^T y \\ &= 4\gamma A^T y - 4\gamma^2 A^T A A^T y - 8\gamma^3 A^T A A^T y + 8\gamma^3 A^T A A^T A A^T y + 2\gamma^3 A^T y \\ &= 6\gamma A^T y - 12\gamma^2 A^T A A^T y + 8\gamma^3 A^T A A^T A A^T y \end{aligned}$$

Showing $x^k \in \text{row}(A)$

- Each update of x^{k+1} involves $A^T v$ for some vector v . These are in $\text{row}(A^T)$ which is the same as $\text{row}(A)$
- $x^0 \in \text{row}(A)$ & each increment of x^{k+1} is from vectors of $\text{row}(A)$
 : By induction, $x^k \in \text{row}(A)$ for $\forall k \geq 0$

(ii) Show x_* is global min w/ l_2 norm is smallest

$x_* \in \text{row}(A)$ where null-space component = 0 as it is completely in the row space of A.

Now consider a second global min x , where
 $x = x_r + x_n$, $x_r \in \text{row}(A)$, $x_n \in \text{null}(A)$

$$\therefore Ax = Ax_r \text{ as } Ax_n = 0$$

$$l_2 \text{ norm: } \|x\|_2 = \sqrt{\|x_r\|_2^2 + \|x_n\|_2^2}$$

$$\|x_*\|_2 = \|x_r\|_2 \text{ as } \|x_n\|_2 = 0$$

Yet, any other global min has $\|x_n\|_2 > 0$

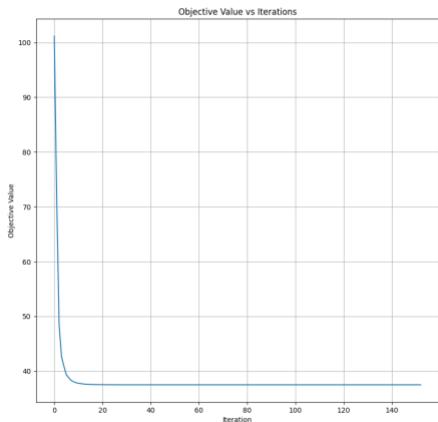
$$\therefore \|x\|_2 > \|x_r\|_2$$

This proves x_* is the global min. with the smallest l_2 norm as it lies entirely on $\text{row}(A)$

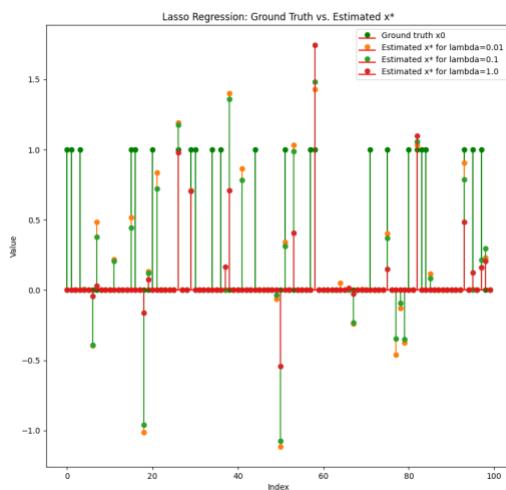
- 5) As λ increases the estimated x_* will become sparser as l_1 helps more coefficients go to 0.
(see code)

Problem 3

Part C



Part F



Problem 4

Generated Example:

```
y_1 = 1.302983480245089
w* = [ 0.49671415 -0.1382643  0.64768854  1.52302986 -0.23415337 -0.23413696
       1.57921282  0.76743473 -0.46947439  0.54256004 -0.46341769 -0.46572975
       0.24196227 -1.91328024 -1.72491783 -0.56228753 -1.01283112  0.31424733
      -0.90802408 -1.4123037 ]
X_1 = [ 1.46564877 -0.2257763   0.0675282  -1.42474819 -0.54438272  0.11092259
      -1.15099358  0.37569802 -0.60063869 -0.29169375 -0.60170661  1.85227818
      -0.01349722 -1.05771093  0.82254491 -1.22084365  0.2088636  -1.95967012
      -1.32818605  0.19686124]
epsilon_1 = 3.169186787416377
```

Least Squares Estimation Error: 0.726222422795486

Least Absolute Deviations Estimation Error: 0.5642178291022407