Connor Holm

5525 HW1

Problem 1

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

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

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

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

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

$tr(M) = \sum m_{ii}$

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

$\Downarrow$

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

$A^T_{ik} = A_{ki}$

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

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

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

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

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

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

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

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

$= \sum_k m_{ki}^2$

---

b) $tr(A^T B) = tr(B^T A)$

$(A^T B)_{ij} = \sum_k a^T_{ik} b_{kj}$ ← * see a bove

$a^T_{ik} = a_{ki}$

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

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

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

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

$\sum_i (A^T B) = \sum_i (B^T A)$

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

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

c) $tr(AB^T) = tr(B^TA)$

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

$= \sum_i \sum_k b^T_{ik} a_{ki}$

$= \sum_i \sum_k a_{ik} b^T_{ki}$

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

$= tr(AB^T)$

Using $(B^TA)_{ij} = \sum_k b^T_{ik} a_{kj}$ & $(AB^T)_{ij} = \sum_k a_{ik} b^T_{kj}$ & $tr(M) = \sum_i m_{ii}$

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

d) $tr(M_1M_2M_3) = tr(M_3M_1M_2) = tr(M_2M_3M_1)$

$tr(M_1M_2M_3) = \sum_i (M_1M_2M_3)_{ii}$

$= \sum_i \sum_j M_{1,ij} (M_2M_2)_{ji}$

$= \sum_i \sum_j \sum_k M_{1,ij} M_{2,jk} M_{3,ki}$ * based on part c

$= \sum_i \sum_j \sum_k M_{3,ki} M_{1,ij} M_{2,jk}$ $= \sum_i \sum_j \sum_k M_{2,jk} M_{3,jk,i} M_{1,ij}$

$= \sum_i (M_3M_1M_2)$ $= \sum_i (M_2M_3M_1)$

$= tr(M_2M_1M_2)$ $= tr(M_2M_3M_1)$

e) $\langle ACB, D \rangle = \langle AC, DB^T \rangle$ As shown before: $\langle ACB, D \rangle = tr((ACB)^T D)$

$tr((ACB)^T D) =$

$tr(B^TC^TA^TD) =$ $(AB)^T = B^TA^T$

$tr(C^TA^TDB^T) =$ As shown from part D)

$tr((AC)^T DB^T) =$

$\langle AC, DB^T \rangle =$ $\langle ACB, D \rangle = \langle CB, A^TD \rangle$

$tr((ACB)^T D) =$

$tr(B^TC^TA^TD) =$

$tr((CB)^T(A^TD)) =$

$\langle CB, A^TD \rangle$

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

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

$M = U \Sigma V^T$

σ_i are diagonal entries of Σ

$$\|M\|_F = \sqrt{tr(M^T M)}$$
$$= \sqrt{tr((U\Sigma V^T)^T (U\Sigma V^T))}$$
$$= \sqrt{tr(V\Sigma^T U^T U \Sigma V^T)}$$
$$= \sqrt{tr(V\Sigma^T I \Sigma V^T)}$$
$$= \sqrt{tr(V\Sigma^T \Sigma V^T)}$$
$$= \sqrt{tr(V\Sigma^2 V^T)}$$
$$= \sqrt{tr(V^T V \Sigma^2)}$$
$$= \sqrt{tr(\Sigma^2)}$$

$= \sqrt{\sum_{i=1}^{r} \sigma_i^2}$

* each diagonal element is σ_i^2

Problem 2

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

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

Gradient $= 2Ax + b$

Hessian

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

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

$= \underbrace{2Ax + b}_{\nabla f(x)} + 2A\delta$

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

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

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

Gradient

$\nabla_\beta (y_i \beta^T x_i) = y_i x_i$

$\nabla_\beta (\log(1+e^{\beta^T x_i})) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} x_i$

$= p(x_i;\beta) x_i$

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

Hessian

$\nabla_\beta (y_i x_i) = 0$

$\nabla_\beta (-p(x_i;\beta) x_i) = \frac{-\partial p(x_i;\beta)}{\partial \beta} x_i^T$

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

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

Minimize $-f(\beta)$

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

$\sum_{i=1}^{N} p(x_i;\beta)(1-p(x_i;\beta)) x_i x_i^T > 0$ 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) $g(x) = \|y - Ax\|_2^2 + \lambda\|x\|_2^2$ $\quad\quad \|v\|_2^2 = v^T v$

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

$= y^Ty - 2y^TAx + x^TA^TAx + \lambda x^Tx$

$\nabla g(x) = 0 - 2A^Ty + 2A^TAx + 2\lambda x$

$= 2(-A^Ty + A^TAx + \lambda x)$

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

$0 = -A^Ty + A^TAx + \lambda x$

$A^Ty = A^TAx + \lambda x$

$A^Ty = x(A^TA + \lambda)$

$(A^TA + \lambda I)^{-1}A^Ty = x_*$ ← local minimizer

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

Global minimizer

- $A^TA + \lambda I$ is positive definite because $\lambda > 0$ showing all eigenvalues are > 0
- Because of this the function is also a global minimizer.

Uniqueness

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

Problem 3

a) Show A has full-rank $\iff$ A^TA is invertible

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

$Ax = 0$, $x = 0$ when A is full-rank

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

$z^TA^TAz = 0$

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

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

$\therefore Az = 0$

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

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

$A^TAz = 0$

$A^T0 = 0$

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

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

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

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

$f(x) = \|y - Ax\|_2^2$, solve for x

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

$A^Ty = A^TAx \rightarrow$ normal equ.

$x_0 = (A^TA)^{-1}A^Ty$

$z_0 = A((A^TA)^{-1}A^Ty)$ $\rightarrow$ closed form of z_0

It is unique, because A is full rank to show A^TA is invertable. It means the normal equation, $A^TAx_0 = A^Ty$, is unique for x_0 which consequently means z_0

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

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

$\langle Ax_0 - y, Ax_w \rangle = 0$, $w = Ax_w$ as some $x_w \in R^n$ 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 span(A)

c) See submitted code

d) From before: $A^T A x_0 = A^T y$ ← global min.

$\ker(A) = \{z \in R^n \mid Az = 0\}$

Let x be another global min → This implies $x - x_0 \in \ker(A^T A)$ or the difference is in the null space.

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

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

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 equ. and is a global min proving the set has a global min

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

$x^0 = 0$

$x^1 = x^0 - 2t A^T(Ax^0 - y)$

$= 0 + 2tA^Ty$

$x_2 = x_1 - 2t A^T(Ax^1 - y)$

$= 2tA^Ty - 2tA^T(2tAA^Ty - y)$

$= 2tA^Ty - 4t^2A^TAA^Ty + 2tA^Ty$

$= 4tA^Ty - 4t^2A^TAA^Ty$

$x^3 = x^2 - 2tA^T(Ax^2 - y)$

$= 4tA^Ty - 2tA^T(A(4tA^Ty - 4t^2A^TAA^Ty) - y)$

$-4t^2A^TAA^Ty$

$= 4tA^Ty - 4t^2A^TAA^Ty - 8t^2A^TAA^Ty + 8t^3A^TAA^TAA^Ty + 2tA^Ty$

$= 6tA^Ty - 12t^2A^TAA^Ty + 8t^3A^TA\,A^TAA^Ty$

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

- Each update of x^{k+1} involves A^Tv for some vector v. These are in $\text{row}(A^T)$ which is the same as $\text{row}(A)$
- $x^0 = 0$, $x^0 \in \text{row}(A)$ & 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
$x = x_r + x_n$, $x_r \in \text{row}(A)$, $x_n \in \text{null}(A)$

$\therefore Ax = Ax_r$ as $Ax_n = 0$

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

$\|x_*\|_2 = \|x_r\|_2$ 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 entirly on row(A)

f) As λ increases the estimated x_* will become sparser as l_1 helps more coefficients go to 0.

(see code)