EE578 Homework #1 Jake Gonzales

1. Linear Algebra and Matrix Analysis Review

a) Let $A, B \in S^n$ and for $1 \leq i \leq n$ let $\lambda_i(A)$ and $\lambda_i(B)$ be the i-th largest eigenvalues of A and B respectively. If $\lambda_i(A) \leq \lambda_i(B)$ for all i, then $A \preceq B$.

True

We know that the largest eigenvalue of a real symmetric matrix A is $\max\{x^TAx : \|x\| = 1\}$ and similarly for B its $\max\{x^TBx : \|x\| = 1\}$.

$A \preceq B$ means that $x^TAx \leq x^TBx$ for all x.

This is also related to the Rayleigh quotient $R(A) = \frac{x^TAx}{x^Tx}$, where the largest eigenvalue is the max of that quotient.

Putting this together it is clear to see that if

$\lambda_i(A) \leq \lambda_i(B)$ for $1 \leq i \leq n$ then $A \preceq B$. ∎

b) Let $A, B \in S^n$. If $A \succeq B \succeq 0$, then $A^{-1} \preceq B^{-1}$.
True.

proof Lets first look at the case where $B = I$, where I is the identity matrix. If $A \succeq I$, then all of the eigenvalues of A are > 1. It follows that since the eigenvalues of A^{-1} are the inverses of the eigenvalues of A so we have that $A^{-1} \preceq I$.

Now consider $B \in S^n$ such that $A \succeq B \succeq 0$. We can write $B^{-1/2} A B^{-1/2} \succeq I$. If we take the inverse we get $(B^{-1/2} A B^{-1/2})^{-1} \preceq I$ then we have:

$$B^{1/2} A^{-1} B^{1/2} \preceq I \Rightarrow A^{-1} \preceq B^{-1}.$$

Thus we have extended the original proof to the general symmetric positive definite matrix B. ☒

multivariate Calculus review

a) Let $f(x) = \|Ax-b\|_2^2 + \lambda \|x-x_0\|_2^2$, where $x \in \mathbb{R}^n$ and $\lambda > 0$. Compute the Hessian of $f(x)$ with respect to x, denoted by $\nabla^2 f(x)$.

Let $g(x) = \|Ax-b\|_2^2$ and $h(x) = \lambda \|x-x_0\|_2^2$.

First

$$g(x) = (Ax+b)^T(Ax+b) = (Ax)^T + b^T)(Ax+b)$$
$$= x^TA^TAx + x^TA^Tb + b^TAx + b^Tb$$

$$\nabla g(x) = (2x^TA^TA)^T + A^Tb + (b^TA)^T$$
$$= 2A^TAx + 2A^Tb = 2A^T(Ax+b)$$

Then, $\nabla^2 g(x) = 2A^TA$.

Next we compute the first derivative of $h(x)$

$$\nabla h(x) = 2\lambda(x-x_0)^T I$$

and the Hessian

$$\nabla^2 h(x) = 2\lambda I.$$

Therefore the Hessian of $f(x)$ with respect to x is

$$\nabla^2 f(x) = 2A^TA + 2\lambda I,$$

b) Let $F(X) = Tr(A^TX)$, where $A, X \in \mathbb{R}^{n \times n}$. Compute the gradient of $f(X)$ with respect to X, denoted by $\nabla_X f(X)$.

We have $Tr(A^TX) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_{ij}$

then $\nabla_X Tr(A^TX) = \nabla_X \left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_{ij}\right)$ →

We can see that $\nabla_X Tr(A^T X) = \frac{\partial Tr(A^T X)}{\partial x_{ji}} = \begin{bmatrix} a_{1,1} & \cdots & a_{n,1} \\ \vdots & & \vdots \\ a_{1,n} & & a_{n,n} \end{bmatrix} = a_{ij}$.

Therefore $\nabla_X f(X) = \nabla_X Tr(A^T X) = A$.

C). Let $f(x) = \ln\left[\sum_{i=1}^{m} \exp(a_i^T x + b)\right]$ where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for all $1 \leq i \leq m$. Compute the gradient of $f(x)$ with respect to $X \in \mathbb{R}^n$.

We can compute the gradient by writing it as a composition of the affine function $Ax + b$, where $A \in \mathbb{R}^{m \times n}$ with rows $a_1^T, \ldots, a_m^T$ and the function $g: \mathbb{R}^m \to \mathbb{R}$ given from the problem statement as $g(y) = \ln\left(\sum_{i=1}^{m} \exp(y_i)\right)$.

By the chain rule, we can write

$$\nabla g(y) = \frac{1}{\sum_{i=1}^{m} \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix},$$

So by the composition

$$\nabla f(x) = \frac{1}{1^T \exp(a_i^T x + b)} A^T \exp(a_i^T x + b)$$

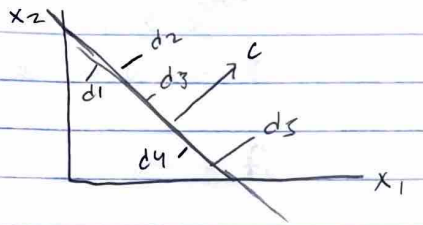
or more compactly

$$\nabla f(x) = \frac{1}{1^T z} A^T z$$

where $z_i = \exp(a_i^T x + b_i)$, $i = 1, \ldots, m$.

Reference: Boyd & V, Convex Optimization, pg. 643, App. A.

3. We are given m data points $d_1, \ldots, d_m \in \mathbb{R}^n$, and we seek a hyperplane $\mathcal{H}(c,b) := \{x \in \mathbb{R}^n \mid c^T x = b\}$, where $c \in \mathbb{R}^n$, $c \neq 0$, and $b \in \mathbb{R}$, that best fits the given points, in the sense of a minimum sum of squared distances criterion, see figure 1.



Formally, we need to solve the optimization problem:

$$\min_{c,b} \sum_{i=1}^{m} \text{dist}^2(d_i, \mathcal{H}(c,b)) \; : \; \|c\|_2 = 1,$$

where $\text{dist}(d_i, \mathcal{H})$ is the Euclidean (l_2-norm) distance from a point d to $\mathcal{H}$. Here, the constraint on c is imposed W.L.O.G., in a way that does not favor a particular direction in space.

a) Show that the distance from a given point $d \in \mathbb{R}^n$ to $\mathcal{H}$ is given by

$$\text{dist}(d, \mathcal{H}(c,b)) = |c^T d - b|.$$

The minimum Euclidean distance from a point to a hyperplane $\mathcal{H}$ is the l_2-norm of the proj. of the point onto the plane. Given our data points d, hyperplane $\mathcal{H}$, and our normal vector and offset we can write

$$\text{dist}(d, \mathcal{H}) = \left| \begin{bmatrix} d^T & -1 \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} \right|$$

which is equivalent to

$$\left| \begin{bmatrix} d^T & -1 \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} \right| = |c^T d - b|.$$

b) Show that the problem can be expressed as

$$\min_{b,c:\|c\|_2=1} f_0(b,c)$$

where f_0 is a certain quadratic function.

Let $D \in \mathbb{R}^{n,m}$ with columns d_i such that, $D = [d_1, \dots, d_m]$, we can write the objective as,

$$f_0 = \sum_{i=1}^{m} \left([d_i^T \; -1] \begin{bmatrix} c \\ b \end{bmatrix} \right)^2$$

$$= \left\| [D^T \; -1] \begin{bmatrix} c \\ b \end{bmatrix} \right\|_2^2$$

$$= \begin{bmatrix} c \\ b \end{bmatrix}^T \begin{bmatrix} DD^T & m\bar{d} \\ -m\bar{d} & m \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix}$$

$$f_0 = c^T (DD^T) c - 2mbc^T \bar{d} + mb^2$$

where m is the number of data points and $\bar{d} = \frac{1}{m} \sum_{i=1}^{m} d_i$, which is the average of the data points $[d_1, \dots d_m]$.

c) show that the problem can be reduced to

$$\min_{c} \; c^T (\tilde{D}\tilde{D}^T) c$$

$$\text{s.t. } \|c\|_2 = 1,$$

where $\tilde{D}$ is the matrix of centered data points.

We know that at optimum, the partial derivative of the objective function with respect to b must be zero we can write

$$\frac{\partial f_0}{\partial b} = -2m^T \bar{d} + 2mb = 0$$

which gives $b = c^T \bar{d}$, which we substitute back into $\bar{f}_0$ so,

$$\bar{f}_0(c) = c^T H c$$

where $H = DD^T - m\bar{d}\bar{d}^T = D(I - \frac{1}{m} 11^T) D^T$

otherwise written $H = \tilde{D}\tilde{D}^T$

Where the i-th column of $\tilde{D}$ is $d_i - \bar{d}$ where $\bar{d} := \frac{1}{m} \sum_{i=1}^{m} d_i$ so $\tilde{D} = D - \bar{d}1^T$. and $\bar{d}$ is the average of the data points.

So we've shown that with our objective $\bar{f}_0(c)$ we can write the optimization problem as

$$\min_{c} c^T(\tilde{D}\tilde{D}^T)c : \|c\|_2 = 1.$$

d) Explain how to find the hyperplane via SVD.

The optimal objective value of our optimization problem above is the minimum eigenvalue of $\tilde{D}\tilde{D}^T$, which would be the smallest singular value of $\tilde{D}$. So the optimal value of our problem is

$$f_0^* = \lambda_{min}(\tilde{D}\tilde{D}^T) = \sigma_n^2$$

With σ_n being the smallest singular value of $\tilde{D}$. Thus we find the corresponding left singular vector denoted u_n and write the best fitting hyperplane as

$$\mathcal{H} = \{x : u_n^T x = u_n^T \bar{d}\},$$

where $\bar{d}$ is the average of the data points.