EECS 127/227AT Optimization Models in Engineering
Spring 2020 Homework 4

This homework is due Friday, February 21, 2020 at 23:00 (11pm).
Self grades are due Friday, February 28, 2020 at 23:00 (11pm).

This version was compiled on 2020-02-25 19:40.

Submission Format: Your homework submission should consist of a single PDF file that contains all of your answers (any handwritten answers should be scanned) as well as your IPython notebook with solutions saved as a PDF.

1. (Practice) Matrix norm

The Frobenius norm of a matrix A is defined as

$$\|A\|_F = \sqrt{\langle A, A\rangle}$$

where for two matrices $A, B \in \mathbb{R}^{m\times n}$, the canonical inner product defined over this space is $\langle A, B\rangle := \mathrm{Tr}(A^\top B) = \sum_{ij} A_{ij}B_{ij}$. The previous definition of the inner product is equivalent to interpreting the matrices A and B as vectors of length n^2 and taking the vector inner product of the respective n^2-dimensional vectors. The Cauchy-Schwarz inequality for the inner product follows in a straightforward way from the Cauchy-Schwarz inequality for vectors:

$$\langle A, B\rangle = \sum_{\substack{1\le i\le m\\ 1\le j\le n}} A_{ij}B_{ij} \le \left(\sum_{\substack{1\le i\le m\\ 1\le j\le n}} A_{ij}^2\right)^{1/2} \left(\sum_{\substack{1\le i\le m\\ 1\le j\le n}} B_{ij}^2\right)^{1/2} = \|A\|_F \|B\|_F .$$

(a) Show that the Frobenius norm satisfies all three properties of a norm.

Solution:

i. $\|A\|_F = 0 \iff A = 0$:
$A = 0 \iff \langle A, A\rangle = 0 \iff \|A\|_F = 0$

ii. For every $\alpha \in \mathbb{R}$, $\|\alpha A\|_F = |\alpha|\|A\|_F$:
$\|\alpha A\|_F = \sqrt{\langle \alpha A, \alpha A\rangle} = \sqrt{\alpha^2\langle A, A\rangle} = |\alpha|\sqrt{\langle A, A\rangle} = |\alpha|\|A\|_F$

iii. $\|A + B\|_F \le \|A\|_F + \|B\|_F$ for all $A, B \in \mathbb{R}^{\gg\times\ltimes}$:
Equivalently, we can show

$$\begin{aligned}
\|A + B\|_F^2 &= \langle A + B, A + B\rangle \\
&= \langle A, A\rangle + 2\langle A, B\rangle + \langle B, B\rangle \\
&= \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle \\
&\le \|A\|_F^2 + \|B\|_F^2 + 2\sqrt{\langle A, A\rangle\langle B, B\rangle} && \text{Cauchy-Schwarz} \\
&= \|A\|_F^2 + \|B\|_F^2 + 2\|A\|_F\|B\|_F \\
&= (\|A\|_F + \|B\|_F)^2
\end{aligned}$$

(b) Write the Frobenius norm squared in terms of singular values.

Solution:

$$\begin{aligned}
\|A\|_F^2 &= \langle A, A\rangle \\
&= \mathrm{Tr}(A^\top A) \\
&= \mathrm{Tr}((U\Sigma V^\top)^\top U\Sigma V^\top) \\
&= \mathrm{Tr}(V\Sigma^\top U^\top U\Sigma V^\top) && U^\top U = I \\
&= \mathrm{Tr}(V\Sigma^\top \Sigma V^\top) \\
&= \mathrm{Tr}(V^\top V\Sigma^\top \Sigma) && \text{Rotation property of trace} \\
&= \mathrm{Tr}(\Sigma^\top \Sigma) \\
&= \sum_i \sigma^2
\end{aligned}$$

(c) Express the Frobenius norm squared in terms of the ℓ_2-norm of the columns of A with $\vec{a}_i$ denoting column i. Concretely, prove $\|A\|_F^2 = \sum_{i=1}^n \|\vec{a}_i\|^2$ where $\vec{a}_i$ are the columns of A.

Solution:

$$\|A\|_F^2 = \langle A, A\rangle = \mathrm{Tr}(A^\top A)$$

$$A^\top A = \begin{bmatrix} \vec{a}_1^\top \\ \vec{a}_2^\top \\ \vdots \\ \vec{a}_n^\top \end{bmatrix} \begin{bmatrix} \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_n \end{bmatrix} = \begin{bmatrix} \vec{a}_1^\top\vec{a}_1 & \vec{a}_1^\top\vec{a}_2 & \cdots & \vec{a}_1^\top\vec{a}_n \\ \vec{a}_2^\top\vec{a}_1 & \vec{a}_2^\top\vec{a}_2 & \cdots & \vec{a}_2^\top\vec{a}_n \\ \vdots & \vdots & \ddots & \vdots \\ \vec{a}_n^\top\vec{a}_1 & \vec{a}_n^\top\vec{a}_2 & \cdots & \vec{a}_n^\top\vec{a}_n \end{bmatrix}$$

$$\mathrm{Tr}(A^\top A) = \sum_i \vec{a}_i^\top \vec{a}_i = \sum_i \|\vec{a}_i\|_2^2$$

(d) A generalization of the least squares problem is to find a *matrix* X that most closely solves the problem $AX = B$. This is sometimes called the *matrix least squares problem*, and when X and B are vectors, reduces to the ordinary least squares problem you are familiar with. Formally, we can define the problem given $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{m\times k}$:

$$\min_{X\in\mathbb{R}^{n\times k}} \|AX - B\|_F$$

However, at this point we only know how to solve the vector least squares problem. Reformulate the above objective in terms of vector least squares problems that way we can solve it.

Hint: The result derived in part (iii) may be particularly useful in addition to the following fact:

$$\min_{\vec{a}_1,\vec{a}_2,\cdots\in\mathbb{R}^n} \|\vec{a}_1\|_2^2 + \|\vec{a}_2\|_2^2 + \cdots = \min_{\vec{a}_1\in\mathbb{R}^n} \|\vec{a}_1\|_2^2 + \min_{\vec{a}_2\in\mathbb{R}^n} \|\vec{a}_2\|_2^2 + \ldots$$

Solution:

Note that the problem is equivalent to $\min_{X\in\mathbb{R}^{n\times k}} \|AX - B\|_F^2$ as the Frobenius norm is a positive function. Using the result from part iii, we can re-express the objective as:

$$\min_{X \in \mathbb{R}^{n \times k}} \|AX - B\|_F^2 = \min_{\vec{x}_1, \vec{x}_2, \ldots \vec{x}_k \in \mathbb{R}^n} \sum_i^k \|A\vec{x}_i - \vec{b}_i\|_2^2 = \sum_i^k \min_{\vec{x}_i \in \mathbb{R}^n} \|A\vec{x}_i - \vec{b}_i\|_2^2$$

This is just the sum of multiple ordinary least squares objectives! So by solving them individually, we are able to obtain an overall optimal matrix X.

2. PCA and low-rank compression

We have a data matrix $X = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{bmatrix}$ of size $n \times m$ containing n data points[1], $x_1, x_2, \ldots, x_n$, with $x_i \in \mathbb{R}^m$. Note that $x_i^\top$ is the ith row of X. Assume that the data matrix is centered, i.e. each column of X is zero mean. In this problem, we will show equivalence between the following three problems:

(P_1) Finding a line going through the origin that maximizes the variance of the scalar projections of the points on the line. Formally P_1 solves the problem:

$$\underset{\vec{u} \in \mathbb{R}^m : \vec{u}^\top \vec{u} = 1}{\text{argmax}} \quad \vec{u}^\top C \vec{u} \tag{1}$$

with $C = \frac{1}{n} \sum_{i=1}^n \vec{x}_i \vec{x}_i^\top$ denoting the covariance matrix associated with the centered data.

(P_2) Finding a line going through the origin that minimizes the sum of squares of the distances from the points to their vector projections. Formally P_2 solves the minimization problem:

$$\underset{\vec{u} \in \mathbb{R}^m : \vec{u}^\top \vec{u} = 1}{\text{argmin}} \sum_{i=1}^n \min_{v_i \in \mathbb{R}} \|\vec{x}_i - v_i \vec{u}\|_2^2 \tag{2}$$

(P_3) Finding a rank-one approximation to the data matrix. Formally P_3 solves the minimization problem:

$$\underset{Y : \text{rank}(Y) \leq 1}{\text{argmin}} \quad \|X - Y\|_F \tag{3}$$

Note that loosely speaking, two problems are said to be "equivalent" if the solution of one can be "easily" translated to the solution of the other. Some form of "easy" translations include adding/subtracting a constant or some quantity depending on the data points.

Note the significance of these results. P_1 is finding the first principal component of X, the direction that maximizes variance of scalar projections. P_2 says that this direction also minimizes the distances between the points to their vector projections along this direction. If we view the distances as errors in approximating the points by their projections along a line, then the error is minimized by choosing the line in the same direction as the first principal component. Finally P_3 tells us that finding a rank one matrix to best approximate the data matrix (in terms of error computed using Frobenius norm) is equivalent to finding the first principal component as well!

[1] Data matrices are sometimes represented as above, and sometimes as the transpose of the matrix here. Make sure you always check this, and recall that based on the definition of the data matrix, the definition of the covariance matrix also changes.

(a) Consider the line $\mathcal{L} = \{\vec{x}_0 \ + v\vec{u} : v \in \mathbb{R}\}$, with $\vec{x}_0 \in \mathbb{R}^m, \vec{u}^\top \vec{u} = 1$. Recall that the vector projection of a point $\vec{x} \in \mathbb{R}^m$ on to the line $\mathcal{L}$ is given by $\vec{z} = \vec{x}_0 + v^*\vec{u}$, where v^* is given by:

$$v^* = \operatorname*{argmin}_v \|\vec{x}_0 + v\vec{u} - \vec{x}\|_2$$

Show that $v^* = (\vec{x} - \vec{x}_0)^\top \vec{u}$. Use this to show that the square of the distance between x and its vector projection on $\mathcal{L}$ is given by:

$$d^2 = \|\vec{x} - \vec{x}_0\|_2^2 - ((\vec{x} - \vec{x}_0)^\top \vec{u})^2$$

Solution: The projection of point $\vec{x}$ on $\mathcal{L}$ corresponds to the following problem:

$$v^* = \min_v \ \|\vec{x}_0 + v\vec{u} - \vec{x}\|_2.$$

The squared objective writes

$$\|\vec{x}_0 + v\vec{u} - \vec{x}\|_2^2 = v^2 - 2v(\vec{x} - \vec{x}_0)^T\vec{u} + \|\vec{x} - \vec{x}_0\|_2^2.$$

By taking the derivative of the above expression with respect to v and setting it to 0, we obtain the optimal value of v as

$$v^* = (\vec{x} - \vec{x}_0)^\top \vec{u}.$$

At optimum, the squared objective function, which equals the minimum squared distance $\|\vec{z} - \vec{x}\|_2^2$, takes the desired value:

$$\|\vec{x}_0 + v^*\vec{u} - \vec{x}\|_2^2 = \|\vec{x} - \vec{x}_0\|_2^2 - ((\vec{x} - \vec{x}_0)^\top \vec{u})^2.$$

(b) Show that P_2 is equivalent to P_1.
Hint: Start with equation (2) and using the result from part (a) show that it is equivalent to equation (1).
Solution: From part (a), we have the following decomposition of P_2:

$$\begin{aligned}
\operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1} \sum_{i=1}^n \min_{v_i\in\mathbb{R}} \|\vec{x}_i - v_i\vec{x}_i\|^2 &= \operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1} \sum_{i=1}^n \|\vec{x}_i\|^2 - (\vec{x}_i^\top\vec{u})^2 \\
&= \sum_{i=1}^n \|\vec{x}_i\|^2 - \operatorname*{argmax}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1} \sum_{i=1}^n \vec{u}^\top\vec{x}_i\vec{x}_i^\top\vec{u} \\
&= \sum_{i=1}^n \|\vec{x}_i\|^2 - \operatorname*{argmax}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1} \vec{u}^\top(\sum_{i=1}^n \vec{x}_i\vec{x}_i^\top)\vec{u} \\
&= \sum_{i=1}^n \|\vec{x}_i\|^2 - n \operatorname*{argmax}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1} \vec{u}^\top C\vec{u}.
\end{aligned}$$

From the above equation, we see that a solution for P_1 constitutes a solution for P_2 and vice-versa.

(c) Show that every matrix $Y \in \mathbb{R}^{n\times m}$ with rank at most 1, can be expressed as $Y = \vec{v}\vec{u}^\top$ for some $\vec{v} \in \mathbb{R}^n$, $\vec{u} \in \mathbb{R}^m$ and $\|\vec{u}\| = 1$.

Hint: Use the SVD.

Solution: First, consider the case where Y is rank-0. If Y is rank 0, all of its all of its singular values must be 0 and hence, Y must be the 0 matrix. Therefore, we can express $Y = \vec{v}\vec{u}^\top$ by setting $\vec{v} = 0$ and $\vec{u}$ being any arbitrary unit-length vector. Now let Y be a rank 1 matrix. Then its has the following SVD: $Y = \sigma\vec{w}\vec{u}^\top$ where $\sigma \neq 0$. It follows that $Y = \vec{v}\vec{u}^\top$ for $\vec{v} = \sigma\vec{w}$.

(d) Show that P_3 is equivalent to P_2.

Hint: Use the result from part (c) to show that P_3 is equivalent to:

$$\operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n} \|X - \vec{v}\vec{u}^\top\|_F^2$$

Prove that this is equivalent to equation (2).

Solution: From the previous part, we have that the set of matrices, Y, with rank at most 1 is equivalent to the set $\{\vec{v}\vec{u}^\top : \|\vec{u}\| = 1, \vec{u} \in \mathbb{R}^m, \vec{v} \in \mathbb{R}^m\}$. Therefore, we may equivalently reformulate P_3 as:

$$\operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n} \left\|X - \vec{v}\vec{u}^\top\right\|_F^2. \tag{4}$$

Now, we expand the objective in the above equation as $\left\|X - \vec{v}\vec{u}^\top\right\|_F^2 = \sum_{i=1}^n \|\vec{x}_i - v_i\vec{u}\|^2$. With this reformulation, we see that any solution $(\vec{u}^*, \vec{v}^*)$ must satisfy $\vec{v}^* = \operatorname{argmin}_{\vec{v}} \sum_{i=1}^n \|\vec{x}_i - v_i\vec{u}\|^2$. Therefore,

$$\begin{aligned}\vec{u}^* &= \operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n} \min_{\vec{v}\in\mathbb{R}^n} \sum_{i=1}^n \|\vec{x}_i - v_i\vec{u}\|^2 \\ &= \operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n} \sum_{i=1}^n \min_{v_i\in\mathbb{R}} \|\vec{x}_i - v_i\vec{u}\|^2.\end{aligned}$$

Therefore, $\vec{u}^*$ is also a solution to P_2.

3. Quadratics and Least Squares

In this question, we will see that every least squares problem can be considered as minimization of a quadratic cost function; whereas not every quadratic minimization problem corresponds to a least-squares problem. To begin with, consider the quadratic function, $f : \mathbb{R}^2 \to \mathbb{R}$ given by:

$$f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top\vec{w} + c$$

where $A \in \mathbb{S}_+^2$ (set of symmetric positive semidefinite matrices in $\mathbb{R}^{2\times2}$), $\vec{b} \in \mathbb{R}^2$ and $c \in \mathbb{R}$.

(a) Assume $c = 0$, and assume that setting $\nabla f(\vec{w}) = 0$ allows us to find the unique minimizer. Give a concrete example of a matrix $A \succ 0$ and a vector $\vec{b}$ such that the point $\vec{w}^* = \begin{bmatrix}-1 & 1\end{bmatrix}^\top$ is the unique minimizer of the quadratic function $f(\vec{w})$.

Solution: First, let $A \succ 0$. Now, by taking the gradient of $f(\vec{w})$ and setting it to zero, we get:

$$\nabla f(\vec{w}^*) = 2A\vec{w}^* - 2b = 0.$$

Since A is positive definite, it is invertible and therefore, the above minimizer is unique. Concretely, let $A = I$. By setting the gradient to zero, we obtain

$$\nabla f(\vec{w}^*) = (A + A^\top)\vec{w}^* - 2\vec{b} = 0 \implies \vec{w}^* = \vec{b}.$$

Then $\vec{w}^* = [-1\ 1]^\top$ is the unique minimizer if $\vec{b} = [-1\ 1]^\top$ and $A = I$.

(b) Assume $c = 0$. Give a concrete example of a matrix $A \succeq 0$, and a vector $\vec{b}$ such that the quadratic function $f(\vec{w})$ has infinitely many minimizers and all of them lie on the line $w_1 + w_2 = 0$. *Hint: Take the gradient of the expression and set it to zero. What needs to be true for there to be infinitely many solutions to the equation?*

Solution: Since $A \in \mathbb{R}^{2\times 2}$ is positive semidefinite, setting gradient to zero shows us that each minimizer $\vec{w}^*$ satisfies

$$\nabla f(\vec{w}^*) = (A + A^\top)\vec{w}^* - 2\vec{b} = 2A\vec{w}^* - 2\vec{b} = 0.$$

In order to have infinitely many solutions, the positive semidefinite matrix A cannot have full rank. Since $A \in \mathbb{S}^2_+$, this amounts to A having rank at most 1. In other words, there must exist a vector $\vec{v} \in \mathbb{R}^2$ such that $A = \vec{v}\vec{v}^\top$. By setting $A = \vec{v}\vec{v}^\top$, each minimizer $\vec{w}^*$ should satisfy

$$A\vec{w}^* - \vec{b} = \vec{v}\vec{v}^\top\vec{w}^* - \vec{b} = 0. \tag{5}$$

Note that each point on the line

$$\mathcal{L} = \{\vec{w} \in \mathbb{R}^2 : w_1 + w_2 = 0\} = \left\{\alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix} : \alpha \in \mathbb{R}\right\}$$

is a minimizer of f. Along with (5), this implies that

$$\vec{v}\vec{v}^\top \left(\alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right) - \vec{b} = 0 \quad \forall \alpha \in \mathbb{R}.$$

This is satisfied only when $\vec{b} = 0$ and $\vec{v} \perp [-1\ 1]^\top$. Choosing $\vec{v} = [1\ 1]^\top$, we have

$$A = \beta \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \vec{b} = 0 \quad \text{for some } \beta > 0.$$

(c) Assume $c = 0$. Let $\vec{w} = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$. Give a concrete example of a **non-zero** matrix $A \succeq 0$ and a vector $\vec{b}$ such that the quadratic function $f(\alpha\vec{w})$ tends to $-\infty$ as $\alpha \to \infty$. *Hint: Use the eigenvalue decomposition to write $A = \sigma_1\vec{u}_1\vec{u}_1^\top + \sigma_2\vec{u}_2\vec{u}_2^\top$ and express $\vec{w}$ in the basis formed by $\vec{u}_1, \vec{u}_2$.* **Solution:**

Let $\vec{w} = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$. We first expand $f(\alpha\vec{w})$ as follows:

$$f(\alpha\vec{w}) = (\vec{w}^\top A\vec{w})\alpha^2 - 2b_1\alpha + c.$$

Note that since A is PSD, $\vec{w}^\top A\vec{w} \geq 0$. Therefore, for $f(\alpha\vec{w})$ to tend to $-\infty$ as $\alpha \to \infty$, we must have $\vec{w}^\top A\vec{w} = 0$ and $b_1 > 0$.

Using the spectral theorem since A is symmetric positive semidefinite it can be written as $A = \sigma_1\vec{u}_1\vec{u}_1^\top + \sigma_2\vec{u}_2\vec{u}_2^\top$ with $\sigma_1 \geq \sigma_2 \geq 0$ and $\vec{u}_1, \vec{u}_2$ orthonormal vectors that form a basis for $\mathbb{R}^2$. Further $\sigma_1 > 0$ since A is not the zero matrix.

Thus we can write $\vec{w} = \beta_1\vec{u}_1 + \beta_2\vec{u}_2$. Substituting this we have, $\vec{w}^\top A\vec{w} = \sigma_1\beta_1^2 + \sigma_2\beta_2^2$. For this to be zero, we must have $\beta_1 = 0$ since $\sigma_1 > 0$.

This implies $\vec{w} = \beta_2\vec{u}_2$ and we must have $\vec{u}_2 = \pm\vec{w} = \pm[1, 0]^\top$ and $\beta_2 = \pm 1$ since both $\vec{w}$ and $\vec{u}$ are unit norm. Using the fact that $\beta_2^2 = 1$ we require $\sigma_2 = 0$ for $\vec{w}^\top A\vec{w}$ to be zero. Further since $\vec{u}_1$ and $\vec{u}_2$ are orthonormal we have $\vec{u}_1 = \pm[0, 1]^\top$.

Putting everything together we can construct one example of A and b as,

$A = (1)\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}$,

and $\vec{b} = \begin{bmatrix}1\\0\end{bmatrix}$.

(d) Say that we have the data set $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1,\ldots,n}$ of features $\vec{x}^{(i)} \in \mathbb{R}^d$ and values $y^{(i)} \in \mathbb{R}$. Define $X = \begin{bmatrix}\vec{x}^{(1)} & \ldots & \vec{x}^{(n)}\end{bmatrix}^\top$ and $\vec{y} = \begin{bmatrix}y^{(1)} & \ldots & y^{(n)}\end{bmatrix}^\top$. In terms of X and $\vec{y}$, find a matrix A, a vector $\vec{b} \in \mathbb{R}^d$ and a scalar c, so that we can express the sum of the square losses $\sum_{i=1}^{n}(\vec{w}^\top\vec{x}^{(i)} - y^{(i)})^2$ as the quadratic function $f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top\vec{w} + c$.

Solution:

$$\sum_{i=1}^{n}(\vec{w}^\top\vec{x}^{(i)} - y^{(i)})^2 = \sum_{i=1}^{n}\left(\vec{w}^\top\vec{x}^{(i)}(\vec{x}^{(i)})^\top\vec{w} - 2\vec{w}^\top(y^{(i)}\vec{x}^{(i)}) + (y^{(i)})^2\right)$$

Rearranging terms, we have

$$A = \sum_{i=1}^{n}\vec{x}^{(i)}\left(\vec{x}^{(i)}\right)^\top = X^\top X, \quad \vec{b} = \sum_{i=1}^{n}y^{(i)}\vec{x}^{(i)}, \quad c = \sum_{i=1}^{n}\left(y^{(i)}\right)^2.$$

(e) Here are three statements with regards to the minimization of a quadratic loss function:

i. It can have a unique minimizer.

ii. It can have infinitely many minimizers.

iii. It can be unbounded from below, i.e. there is some direction, $\vec{w}$ so that $f(\alpha\vec{w})$ goes to $-\infty$ as $\alpha \to \infty$.

All three statements apply to general minimization of a quadratic cost function. Parts (a), (b) and (c) give concrete examples of quadratic cost functions where (i), (ii) and (iii) apply respectively. However, notice that statement (iii) cannot apply to the least squares problem as the objective is always positive.

The least-squares problem can have infinitely many minimizers though. How? Consider the gradient of the least squares problem in part (d) at an optimal solution $\vec{w}^*$:

$$\nabla f(\vec{w}^*) = 2X^\top X\vec{w}^* - 2\vec{b} = 0.$$

Therefore, the least squares problem only has multiple solutions if $X^\top X$ is not full rank. This means that $\text{rank}(X^\top X) = \text{rank}(X) < d$. Finally, the rank of X is less than d when the data points $\{\vec{x}^{(i)}\}_{i=1}^n$ do not span $\mathbb{R}^d$. This can happen when the number of data points n is less than d or when $\{\vec{z}_i\}_{i=1}^d$ are linearly dependent where $\vec{z}_i$ are the columns of X.

To complete this subpart of the question, make sure you understand the discussion above and indicate that in the solution.

4. Proof of the Eckart-Young Theorem

Given a matrix $A \in \mathbb{R}^{m\times n}$ with singular value decomposition $A = U\Sigma V^\top$, define the matrix $A_k = \sum_{i=1}^k \sigma_i \vec{u}_i \vec{v}_i^\top$ where $\vec{u}$ and $\vec{v}$ denote the ith left and right singular vectors of A and σ_i denotes the ith singular value. Recall that the Eckart-Young Theorem states that:

$$A_k = \underset{\substack{B\in\mathbb{R}^{m\times n}\\ \text{rank}(B)\le k}}{\text{argmin}} \|A-B\|_2 \qquad \text{Spectral Norm Approximation}$$

$$A_k = \underset{\substack{B\in\mathbb{R}^{m\times n}\\ \text{rank}(B)\le k}}{\text{argmin}} \|A-B\|_F \qquad \text{Frobenius Norm Approximation.}$$

That is, the matrix A_k is the best rank-k approximation of A in both the Spectral and Frobenius norms. In the question, we will prove the Eckart-Young Theorem.

(a) Prove the Spectral Norm Approximation result from the Eckart-Young Theorem

Solution: We first observe that $\|A - A_k\|_2 = \sigma_{k+1}$ as $(A - A_k) = \sum_{i=k+1}^n \sigma_i \vec{u}\vec{v}^\top$ is a valid singular value decomposition of $A - A_k$. Therefore, it suffices to prove that for any matrix B with rank at most k, we have $\|A - B\|_2 \ge \sigma_{k+1}$.

To do this, we need to exhibit a vector $\vec{w}$ such that $\|(A - B)\vec{w}\| \ge \sigma_{k+1}$. Since, B is of rank at most k, its nullspace, $\mathcal{N}(B)$, is of dimension at least $n - k$. Now, let $\mathcal{V}$, denote the rank $k + 1$ subspace spanned by $\{\vec{v}_i\}_{i=1}^{k+1}$, the first $k + 1$ right singular vectors of A. Since, $\mathcal{V}$ is of dimension $k + 1$ and $\mathcal{N}(B)$ is of dimension at least $n - k$, there exists $\vec{w} \in \mathcal{V} \cap \mathcal{N}(B)$ with $\|\vec{w}\| = 1$. Therefore, we have $\vec{w} = \sum_{i=1}^{k+1} \alpha_i \vec{v}_i$ and for this $\vec{w}$, we get:

$$\|(A - B)\vec{w}\|_2^2 = \|A\vec{w}\|_2^2 = \left\|U\Sigma V^\top \vec{w}\right\|_2^2 = \left\|\Sigma V^\top \vec{w}\right\|_2^2 = \sum_{i=1}^{k+1} \alpha_i^2 \sigma_i^2 \ge \sigma_{k+1}^2.$$

This proves the first part of the Eckart-Young Theorem.

(b) Prove the Frobenius Norm Approximation result from the Eckart-Young Theorem

Solution: As in the first part, we note that $\|A - A_k\|_F^2 = \sum_{i=k+1}^n \sigma_i^2$. Therefore, it suffices to show that for any matrix B of rank at most k, we have $\|A - B\|_F^2 \ge \sum_{i=k+1}^n \sigma_i^2$.

Recall that for any matrix, C, we have $\|C\|_F^2 = \sum_{i=1}^n \sigma_i^2(C)$. We will now prove that $\|A - B\|_F^2 \ge \sum_{i=k+1}^n \sigma_i^2$ by showing that $\sigma_r(A - B) \ge \sigma_{k+r}$ for any r. We will also make use of the following recursive characterization of the singular values of a matrix. That is for any matrix, C, we have:

$$\sigma_r(C) = \|C - C_{r-1}\|_2,$$

where $C_i = \sum_{j=1}^i \sigma_j(C)\vec{u}_j(C)\vec{v}_j(C)^\top$.

From the above result, we have $\sigma_r(A - B) = \|A - B - (A - B)_{r-1}\|_2$. Since, the matrix $B - (A - B)_{r-1}$ is of rank at most $k + r - 1$ because the rank of B is at most k and the rank of $(A - B)_{r-1}$ is at most $r - 1$, we have from the previous part:

$$\sigma_r(A - B) = \|A - B - (A - B)_{r-1}\|_2 \ge \sigma_{k+r}.$$

Since, $\sigma_r(A - B) \ge \sigma_{k+r}$ for any r, we have $\sum_{i=1}^n \sigma_i^2(A - B) \ge \sum_{i=k+1}^n \sigma_i^2$. This concludes the proof of the result.

5. (Practice) Ridge Regression

Prove that the optimal solution to the ridge regression problem:

$$\min_{\vec{w}\in\mathbb{R}^p} \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2,$$

where $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and $\vec{y} \in \mathbb{R}^n$, is given by:

$$\vec{w}^* = (X^\top X + \lambda I)^{-1} X^\top \vec{y}.$$

Solution: We by taking the gradient of of the objective function

$$f(\vec{w}) = \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2 = \vec{w}^\top X^\top X\vec{w} - 2\vec{y}^\top X\vec{w} + \|\vec{y}\|^2 + \lambda \|\vec{w}\|^2$$

with respect to $\vec{w}$ and setting it to zero, we get:

$$\nabla f(\vec{w}^*) = 2X^\top X\vec{w}^* + 2\lambda\vec{w}^* - 2X^\top\vec{y} = 0 \implies \vec{w}^* = (X^\top X + \lambda I)^{-1} X^\top \vec{y}.$$

6. Variants on least sqaures: a playground

In this problem we will explore four different types of regression: Ordinary least squares, ridge regression, weighted least squares and Tikhonov regularization. This problem has an associated ipython notebook 'regression_playground.ipynb'. You will write lines of code and answer questions asked (marked with TODO) in the notebook itself. **Solution:** Please the ipython-notebook for solutions.

7. Homework process

Whom did you work with on this homework? List the names and SIDs of your group members.