EECS 127/227AT Optimization Models in Engineering UC Berkeley Fall 2022

Homework 4

This homework is due at 11 pm on 23 September 2022.
Self grades are due at 11 pm on 30 September 2022.

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 printouts of the completed Jupyter notebooks for this homework.

1. 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:

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

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:

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

Note that the vector projection of $\vec{x}$ on $\vec{u}$ is given by $v^* \vec{u}$, where

$$v^* = \underset{v \in \mathbb{R}}{\operatorname{argmin}} \|\vec{x} - v\vec{u}\|_2^2, \tag{3}$$

and we will show that $v^* = < \vec{x}, \vec{u} >$ in part (a).

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

$$\underset{Y : \mathrm{rk}(Y) \leq 1}{\operatorname{argmin}} \|X - Y\|_F \tag{4}$$

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.

[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.

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!

(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:

$$v^* = \operatorname*{argmin}_v \|\vec{x}_0 + v\vec{u} - \vec{x}\|_2 \tag{5}$$

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:

$$d^2 = \|\vec{x} - \vec{x}_0\|_2^2 - ((\vec{x} - \vec{x}_0)^\top \vec{u})^2 \tag{6}$$

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

$$v^* = \min_v \|\vec{x}_0 + v\vec{u} - \vec{x}\|_2 . \tag{7}$$

The squared objective writes

$$\|\vec{x}_0 + v\vec{u} - \vec{x}\|_2^2 = v^2 - 2v(\vec{x} - \vec{x}_0)^T\vec{u} + \|\vec{x} - \vec{x}_0\|_2^2 . \tag{8}$$

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

$$v^* = (\vec{x} - \vec{x}_0)^\top \vec{u}. \tag{9}$$

The square of the distance between $\vec{x}$ and its vector projection on $\mathcal{L}$ ($\vec{z}$) is given by $\|\vec{z} - \vec{x}\|_2^2$. We have shown that $\vec{z} = \vec{x}_0 + v^*\vec{u} = \vec{x}_0 + (\vec{x} - \vec{x}_0)^\top \vec{u} * \vec{u}$. At optimum, the squared objective function, which equals the minimum squared distance $\|\vec{z} - \vec{x}\|_2^2$, takes the desired value:

$$\left\|\vec{x}_0 + (\vec{x} - \vec{x}_0)^\top \vec{u} * \vec{u} - \vec{x}\right\|_2^2 = \|\vec{x} - \vec{x}_0\|_2^2 - ((\vec{x} - \vec{x}_0)^\top \vec{u})^2. \tag{10}$$

(b) Show that P_2 is equivalent to P_1.

(HINT: Start with P_2 and using the result from part (a) show that it is equivalent to P_1.)

Solution: From part (a), we have the following decomposition of P_2:

$$\operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m : \vec{u}^\top\vec{u}=1} \sum_{i=1}^n \min_{v_i \in \mathbb{R}} \|\vec{x}_i - v_i u\|^2 = \operatorname*{argmin}_{\vec{u}\in\mathbb{R}^m : \vec{u}^\top\vec{u}=1} \sum_{i=1}^n \|\vec{x}_i\|^2 - (\vec{x}_i^\top \vec{u})^2 \tag{11}$$

$$= \operatorname*{argmax}_{\vec{u}\in\mathbb{R}^m : \vec{u}^\top\vec{u}=1} \sum_{i=1}^n \vec{u}^\top \vec{x}_i \vec{x}_i^\top \vec{u} \tag{12}$$

$$= \operatorname*{argmax}_{\vec{u}\in\mathbb{R}^m : \vec{u}^\top\vec{u}=1} \vec{u}^\top C \vec{u}. \tag{13}$$

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 $\|\vec{u}\| = 1$.

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:

$$\underset{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n}{\operatorname{argmin}} \left\|X - \vec{v}\vec{u}^\top\right\|_F^2 \tag{14}$$

Prove that this is equivalent to P_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:

$$\underset{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n}{\operatorname{argmin}} \left\|X - \vec{v}\vec{u}^\top\right\|_F^2. \tag{15}$$

X is a matrix with rows $\vec{x}_i^\top$, and $\vec{v}\vec{u}^\top$ is a matrix with rows $v_i\vec{u}^\top$. We expand the Frobenius norm in the objective in the above equation as

$$\left\|X - \vec{v}\vec{u}^\top\right\|_F^2 = \sum_{i=1}^n \|\vec{x}_i - v_i\vec{u}\|^2, \tag{16}$$

i.e., express the matrix norm as a sum of vector norms, which follows from the definition of the Frobenius norm.

With this reformulation, we see that any solution $(\vec{u}^*, \vec{v}^*)$ must satisfy

$$\vec{v}^* = \underset{\vec{v}}{\operatorname{argmin}} \sum_{i=1}^n \|\vec{x}_i - v_i\vec{u}\|^2, \quad \vec{u}^* = \underset{\vec{u}}{\operatorname{argmin}} \sum_{i=1}^n \|\vec{x}_i - v_i^*\vec{u}\|^2$$

i.e., we can minimize it over $\vec{u}, \vec{v}$ sequentially. We separate the minimization over $\vec{u}$ and $\vec{v}$ to get

$$\vec{u}^* = \underset{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1}{\operatorname{argmin}} \min_{\vec{v}\in\mathbb{R}^n} \sum_{i=1}^n \|\vec{x}_i - v_i\vec{u}\|^2 \tag{17}$$

We now have a minimization of a sum of squares of vector norms $\|\vec{x}_i - v_i\vec{u}\|^2$, each of which depends only on a single element of $\vec{v}$, i.e., v_i.

Note: The objective of an optimization problem $\min_{x,y} f(x,y)$ is said to be separable when the objective can be written as a sum of two functions- one which depends on x, and one on y, i.e.,

$$\min_{x,y} f(x,y) = \min_{x,y}[g(x) + h(y)].$$

If the objective is separable, we can solve the problem separately across the two variables, and

$$(x^*, y^*) = \underset{x,y}{\operatorname{argmin}} f(x,y) = (\underset{x}{\operatorname{argmin}}\, g(x), \underset{y}{\operatorname{argmin}}\, h(y)).$$

We can split the minimization problem in 17 over each individual v_i. We have

$$\vec{u}^* = \underset{\vec{u}\in\mathbb{R}^m:\vec{u}^\top\vec{u}=1,\vec{v}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{v_i\in\mathbb{R}} \|\vec{x}_i - v_i\vec{u}\|^2 . \tag{18}$$

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

2. 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:

$$f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top \vec{w} + c \tag{19}$$

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

$$\nabla f(\vec{w}^*) = 2A\vec{w}^* - 2b = 0. \tag{20}$$

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

$$\nabla f(\vec{w}^*) = (A + A^\top)\vec{w}^* - 2\vec{b} = 0 \implies \vec{w}^* = \vec{b}. \tag{21}$$

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

$$\nabla f(\vec{w}^*) = (A + A^\top)\vec{w}^* - 2\vec{b} = 2A\vec{w}^* - 2\vec{b} = 0. \tag{22}$$

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

$$A\vec{w}^* - \vec{b} = \vec{v}\vec{v}^\top\vec{w}^* - \vec{b} = 0. \tag{23}$$

Note that each point on the line

$$\mathcal{L} = \{\vec{w} \in \mathbb{R}^2 : w_1 + w_2 = 0\} = \left\{ \alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix} : \alpha \in \mathbb{R} \right\} \tag{24}$$

is a minimizer of f. Along with (23), this implies that

$$\vec{v}\vec{v}^\top \left(\alpha \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) - \vec{b} = 0 \quad \forall \alpha \in \mathbb{R}. \tag{25}$$

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

$$A = \beta \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \vec{b} = 0 \quad \text{for some } \beta > 0. \tag{26}$$

(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:

$$f(\alpha\vec{w}) = (\vec{w}^\top A\vec{w})\alpha^2 - 2b_1\alpha + c. \tag{27}$$

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 data points $\vec{x}_i \in \mathbb{R}^d$ 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:

$$\sum_{i=1}^n (\vec{w}^\top\vec{x}_i - y_i)^2 = \sum_{i=1}^n \left(\vec{w}^\top\vec{x}_i(\vec{x}_i)^\top\vec{w} - 2\vec{w}^\top(y_i\vec{x}_i) + (y_i)^2\right) \tag{28}$$

Rearranging terms, we have

$$A = \sum_{i=1}^n \vec{x}_i\left(\vec{x}_i\right)^\top = X^\top X, \quad \vec{b} = \sum_{i=1}^n y_i\vec{x}_i, \quad c = \sum_{i=1}^n \left(y_i\right)^2. \tag{29}$$

(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 $\vec{w}^*$:

$$\nabla f(\vec{w}^*) = 2X^\top X\vec{w}^* - 2\vec{b} = 0. \tag{30}$$

Therefore, the least squares problem only has multiple solutions if $X^\top X$ is not full rank. This means that $\text{rk}\left(X^\top X\right) = \text{rk}(X) < d$. 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 $\{\vec{f}_i\}_{i=1}^d$ are linearly dependent where $\vec{f}_i$ are the columns of X, i.e., the features.

Indicate below that you have read and understood the discussion above.

Solution: Any answer is fine.

3. Gradients, Jacobian matrices and Hessians

The *Gradient* of a scalar-valued function $g : \mathbb{R}^n \to \mathbb{R}$, is the column vector of length n, denoted as ∇g, containing the derivatives of components of g with respect to the variables:

$$(\nabla g(x))_i = \frac{\partial g}{\partial x_i}(x),\ i = 1, \dots n.$$

The *Hessian* of a scalar-valued function $g : \mathbb{R}^n \to \mathbb{R}$, is the $n \times n$ matrix, denoted as $\nabla^2 g$, containing the second derivatives of components of g with respect to the variables:

$$(\nabla^2 g(x))_{ij} = \frac{\partial^2 g}{\partial x_i \partial x_j}(x),\ \ i = 1, \dots, n,\ \ j = 1, \dots, n.$$

The *Jacobian* of a vector-valued function $g : \mathbb{R}^n \to \mathbb{R}^m$ is the $m \times n$ matrix, denoted as Dg, containing the derivatives of components of g with respect to the variables:

$$(Dg)_{ij} = \frac{\partial g_i}{\partial x_j},\ \ i = 1, \dots, m,\ \ j = 1, \dots, n.$$

For the remainder of the class, we will repeatedly have to take gradients, hessians and jacobians of functions we are trying to optimize. This exercise serves as a warm up for future problems.

(a) Compute the gradients and Hessians for the following functions:

i. $g(x) = x^\top A x$

ii. $g(x) = \|Ax - b\|_2^2$

Consider the case now where all vectors and matrices above are scalar; do your answers above make sense? (No need to answer this in your submission)

Solution:

i. Let $A = [a_1,\ a_2,\ \dots\ , a_n]$ where a_i is the i-th column of A. Similarly, let $a_i^\top$ be the i-th row of $A^\top$. For notational convenience, let α_i^T denote the i-th row of A. Finally, let a_{ij} denote the (i, j)th entry of A. Then

$$\begin{aligned} g(x) &= x^\top A x \\ &= x^\top [a_1,\ a_2,\ \dots\ , a_n] x \\ &= x^\top (a_1 x_1 + a_2 x_2 + \ \dots\ + a_n x_n) \\ &= \sum_{i=1}^{n} (x^\top a_i) x_i. \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial g}{\partial x_j}(x) &= \frac{\partial}{\partial x_j}\left[(x^\top a_j) x_j + \sum_{i \neq j} (x^\top a_i) x_i\right] \\ &= x^\top a_j + a_{jj} x_j + \sum_{i \neq j} a_{ji} x_i \\ &= a_j^\top x + \alpha_j^\top x. \end{aligned}$$

It follows that $\nabla g(x) = (A + A^\top)x$. Note if A is symmetric this reduces to $2Ax$. Based on the definition of the Hessian, it follows that the ith column of the hessian is the ith column of $A + A^\top$. Thus $\nabla^2 g(x) = A + A^\top$.

ii. Expanding the norm and using the fact that $c^\top x = x^\top c$ we have that $g(x) = x^\top A^\top Ax - 2b^\top Ax + b^\top b$. Using the previous results and the fact that $A^\top A$ is symmetric, it follows that $\nabla g(x) = 2(A^\top Ax - A^\top b)$ and $\nabla^2 g(x) = 2A^\top A$.

(b) Compute the Jacobians for the following maps

i. $g(x) = Ax$

ii. $g(x) = f(x)x$ where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is once-differentiable

iii. $g(x) = f(Ax + b)x$ where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is once differentiable and $A \in \mathbb{R}^{n\times n}$

Solution:

i. Note $g_i(x) = \alpha_i^\top x$ where $\alpha_i^\top$ is the i-th row of A. Then $\dfrac{\partial g_i}{\partial x_j} = \alpha_{ij}$ which is simply the (i,j) entry of A. It follows that $Dg(x) = A$.

ii. Again $g_i(x) = f(x)x_i$. Then

$$\frac{\partial g_i}{\partial x_i} = f(x) + (\nabla f(x))_i x_i$$
$$\frac{\partial g_i}{\partial x_j} = 0 + (\nabla f(x))_j x_i.$$

It follows that $Dg(x) = x(\nabla f(x))^\top + f(x)I$.

iii. First, we derive $\nabla f(z)$. Let $z = Ax + b$. Let $\alpha_i^\top$ and a_i denote the i-th row and i-th column of A respectively. Finally let a_{ij} denote (i,j)th entry of A. Note $z_i = \alpha_i^\top x + b_i$. Then by the chain rule we have

$$\begin{aligned}\frac{\partial f}{\partial x_j} &= \sum_{i=1}^n \frac{\partial f}{\partial z_i}\frac{\partial z_i}{\partial x_j} \\ &= \sum_{i=1}^n \frac{\partial f}{\partial z_i} a_{ij} \\ &= a_j^\top \nabla f(z).\end{aligned}$$

It follows that $\nabla f(Ax + b) = \nabla f(z) = A^\top \nabla f(Ax + b)$.
Returning to the original problem, we have $g_i(x) = f(Ax + b)x_i$. Then using the derivation in the previous part, it follows that $Dg(x) = x(\nabla f(Ax + b))^\top A + f(Ax + b)I$.

(c) Plot/hand-draw the level sets of the following functions:

i. $g(x_1, x_2) = \dfrac{x_1^2}{4} + \dfrac{x_2^2}{9}$

ii. $g(x_1, x_2) = x_1 x_2$

Also point out the gradient directions in the level-set diagram. Additionally, compute the first and second order Taylor series approximation around the point $(1, 1)$ for each function and comment on how accurately they approximate the true function.

Solution: Figures 1 and 2 contain the level sets and gradient directions for the given functions.

i. We first compute the first and second order partial derivatives of g as follows:

$$\frac{\partial g}{\partial x_1}(x_1, x_2) = \frac{x_1}{2}, \quad \frac{\partial g}{\partial x_2}(x_1, x_2) = \frac{2x_2}{9}, \tag{31}$$

$$\frac{\partial^2 g}{\partial x_1^2}(x_1, x_2) = \frac{1}{2}, \quad \frac{\partial^2 g}{\partial x_2 x_1}(x_1, x_2) = 0, \tag{32}$$

$$\frac{\partial^2 g}{\partial x_2^2}(x_1, x_2) = \frac{2}{9}, \quad \frac{\partial^2 g}{\partial x_1 x_2}(x_1, x_2) = 0. \tag{33}$$

The gradient of g is then given by,

$$\nabla g(x_1, x_2) = \begin{bmatrix} \frac{\partial g}{\partial x_1}(1,1) \\ \frac{\partial g}{\partial x_2}(1,1) \end{bmatrix}, \tag{34}$$

and the Hessian matrix is given by,

$$H(x_1, x_2) = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1^2}(x_1, x_2) & \frac{\partial^2 g}{\partial x_1 x_2}(x_1, x_2) \\ \frac{\partial^2 g}{\partial x_2 x_1}(x_1, x_2) & \frac{\partial^2 g}{\partial x_2^2}(x_1, x_2) \end{bmatrix}. \tag{35}$$

The first order Taylor series approximation around $(1, 1)$ can be computed as:

$$g(x_1, x_2) \approx g(1,1) + (\nabla g(1,1))^\top \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \tag{36}$$

$$= \frac{13}{36} + \frac{x_1}{2} - \frac{1}{2} + \frac{2x_2}{9} - \frac{2}{9} = \frac{x_1}{2} + \frac{2x_2}{9} - \frac{13}{36}. \tag{37}$$

The second order Taylor series approximation around $(1, 1)$ can be computed as:

$$g(x_1, x_2) \approx g(1,1) + (\nabla g(1,1))^\top \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x_1 - 1 & x_2 - 1 \end{bmatrix} H(1,1) \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \tag{38}$$

$$= \frac{x_1}{2} + \frac{2x_2}{9} - \frac{13}{36} + \frac{1}{2}\left(\frac{1}{2}(x_1 - 1)^2 + \frac{2}{9}(x_2 - 1)^2\right) \tag{39}$$

$$= \frac{(x_1 - 1)^2}{4} + \frac{(x_2 - 1)^2}{9} + \frac{x_1}{2} + \frac{2x_2}{9} - \frac{13}{36}. \tag{40}$$

$$= \frac{x_1^2}{4} + \frac{x_2^2}{9} \tag{41}$$

The original function at (1.1,1.1) takes on the value 0.437. The first order approximation returns, evaluated at (1.1,1.1): $\frac{1.1}{2} + \frac{2.2}{9} - \frac{13}{36} = 0.433$. Additionally, observe that the second order approximation simplifies to return the original function!

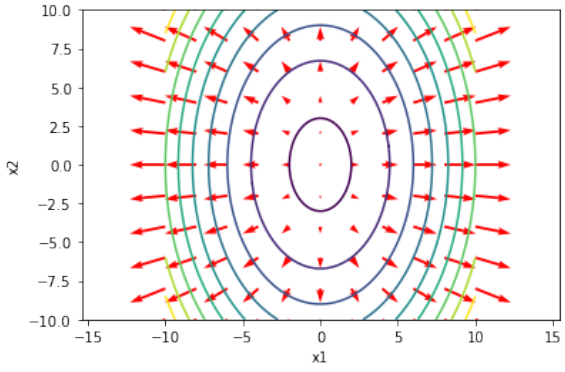

Figure 1: Level sets and gradient directions for the function $g(x_1, x_2) = \frac{x_1^2}{4} + \frac{x_2^2}{9}$.

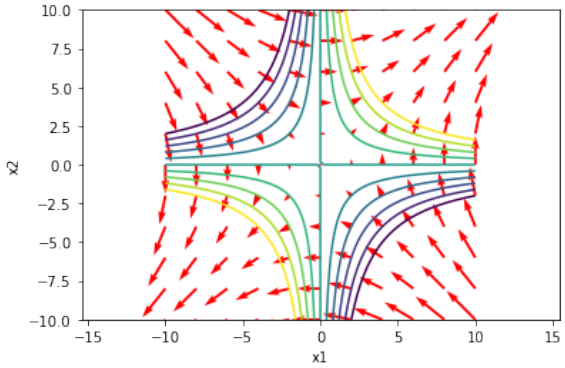

Figure 2: Level sets and gradient directions for the function $g(x_1, x_2) = x_1 x_2$.

ii. We follow the same steps as in the previous part of the problem. The partial derivatives for this g are given by:

$$\frac{\partial g}{\partial x_1}(x_1, x_2) = x_2, \quad \frac{\partial g}{\partial x_2}(x_1, x_2) = x_1, \tag{42}$$

$$\frac{\partial^2 g}{\partial x_1^2}(x_1, x_2) = 0, \quad \frac{\partial^2 g}{\partial x_2 x_1}(x_1, x_2) = 1, \tag{43}$$

$$\frac{\partial^2 g}{\partial x_2^2}(x_1, x_2) = 0, \quad \frac{\partial^2 g}{\partial x_1 x_2}(x_1, x_2) = 1. \tag{44}$$

The first order Taylor series approximation around $(1, 1)$ can be computed as:

$$g(x_1, x_2) \approx g(1, 1) + (\nabla g(1, 1))^\top \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \tag{45}$$

$$= 1 + x_1 - 1 + x_2 - 1 = x_1 + x_2 - 1. \tag{46}$$

The second order Taylor series approximation around $(1, 1)$ can be computed as:

$$g(x_1, x_2) \approx g(1, 1) + (\nabla g(1, 1))^\top \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x_1 - 1 & x_2 - 1 \end{bmatrix} H(1, 1) \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} \tag{47}$$

$$= x_1 + x_2 - 1 + \frac{1}{2} (2(x_1 - 1)(x_2 - 1)) \tag{48}$$

$$= (x_1 - 1)(x_2 - 1) + x_1 + x_2 - 1. \tag{49}$$

$$= x_1 x_2 \tag{50}$$

The original function evaluated at (1.1,1.1) is 1.21. The first order approximation around $(1.1, 1.1)$ is 1.2, but the second order approximation again exactly represents the function!

4. A quadratic-fractional function and its gradient

The goal of this problem is to compute gradients and hessians without explicitly writing out the partial derivatives. As we will see later in the class, most optimization problems can be reformulated as quadratic functionals (such as the famous least squares objective function) and it is essential that you are able to compute symbolic gradients of these without having to write out the partial derivatives. In this problem, we consider the function $f : \mathbb{R}^n \to \mathbb{R}$, with values for $x \in \mathbb{R}^n$ given by

$$f(x) = z^\top (I + xx^\top)^{-1} z.$$

Here, $z \in \mathbb{R}^n$ is given and non-zero. Note that $(I + xx^\top)$ is invertible for all x, as you know from the shift property of eigenvalues shown in lecture.

(a) Show that,

$$f(x) = \|z\|_2^2 - \frac{(z^\top x)^2}{1 + x^\top x}. \tag{51}$$

Hint: prove that for every $x \in \mathbb{R}^n$, we have

$$(I + xx^\top)^{-1} = I - \frac{xx^\top}{1 + x^\top x}.$$

Solution: We first prove that,

$$(I + xx^\top)^{-1} = I - \frac{xx^\top}{1 + x^\top x}. \tag{52}$$

To do this, you can either show $I - \frac{xx^\top}{1+x^\top x}$ is a left-inverse by:

$$\left(I - \frac{xx^\top}{1 + x^\top x}\right)(I + xx^\top) = I - \frac{xx^\top}{1 + x^\top x} + xx^\top - \frac{xx^\top xx^\top}{1 + x^\top x} \tag{53}$$

$$= I - \frac{xx^\top}{1 + x^\top x}(-1 + 1 + x^\top x - x^\top x). \tag{54}$$

$$= I. \tag{55}$$

Or show $I - \frac{xx^\top}{1+x^\top x}$ is a right-inverse by:

$$(I + xx^\top)\left(I - \frac{xx^\top}{1 + x^\top x}\right) = I - \frac{xx^\top}{1 + x^\top x} + xx^\top - \frac{xx^\top xx^\top}{1 + x^\top x} \tag{56}$$

$$= I - \frac{xx^\top}{1 + x^\top x}(-1 + 1 + x^\top x - x^\top x). \tag{57}$$

$$= I. \tag{58}$$

Either way, we have $(I + xx^\top)^{-1} = I - \frac{xx^\top}{1+x^\top x}$. Substituting this in expression for $f(x)$ we get,

$$f(x) = z^\top \left(I - \frac{xx^\top}{1 + x^\top x}\right) z \tag{59}$$

$$= \|z\|_2^2 - \frac{z^\top (xx^\top) z}{1 + x^\top x} \tag{60}$$

$$= \|z\|_2^2 - \frac{(z^\top x)^2}{1 + x^\top x}. \tag{61}$$

Note that in the last equality we used the fact that $x^\top z$ is a scalar and is equal to $z^\top x$.

(b) Compute the gradient of f at a point x. We expect a symbolic answer in terms of $x, Q \doteq zz^\top$, and $\lambda(x) \doteq \frac{x^\top Qx}{1+x^\top x}$. We don't expect answers expressed in terms of components of x and z computed by taking partial derivatives with respect to components of x.
Recall the quotient rule for finding the gradient of $h(x) = \frac{n(x)}{d(x)}$, where $n(x)$ and $d(x)$ are scalar-valued functions. We have,

$$\nabla h(x) = \frac{d(x)\nabla n(x) - n(x)\nabla d(x)}{(d(x))^2}. \tag{62}$$

Hint: Start with the expression of $f(x)$ from equation 51 in part a) and reduce the problem to that of computing the gradient of a function of the form $h(x) = \frac{n(x)}{d(x)}$ with $d(x) = 1 + x^\top x$.

Solution: Using the hint we start from equation 51,

$$\nabla f(x) = \nabla \left(\|z\|_2^2 - \frac{(z^\top x)^2}{1 + x^\top x} \right) \tag{63}$$

$$= \nabla \left(\|z\|_2^2 - \frac{x^\top (zz^\top) x}{1 + x^\top x} \right) \tag{64}$$

$$= -\nabla \left(\frac{x^\top Qx}{1 + x^\top x} \right), \quad Q \doteq zz^\top. \tag{65}$$

$$= -\nabla h(x), \quad h(x) \doteq \frac{x^\top Qx}{1 + x^\top x}. \tag{66}$$

Observe that $h(x) = \frac{n(x)}{d(x)}$ with $n(x) = x^\top Qx$ and $d(x) = 1 + x^\top x$.
Using the quotient rule and the fact that $Q = Q^\top$ we have,

$$\nabla h(x) = \frac{(1 + x^\top x)\nabla(x^\top Qx) - (x^\top Qx)\nabla(1 + x^\top x)}{(1 + x^\top x)^2} \tag{67}$$

$$= \frac{(1 + x^\top x)2Qx - (x^\top Qx)2x}{(1 + x^\top x)^2} \tag{68}$$

$$= \frac{2}{1 + x^\top x} Qx - \frac{2x^\top Qx}{(1 + x^\top x)^2} x \tag{69}$$

$$= \frac{2}{1 + x^\top x} (Qx - \lambda(x)x), \quad \lambda(x) \doteq \frac{x^\top Qx}{1 + x^\top x}. \tag{70}$$

From equation 66 this means,

$$\nabla f(x) = \frac{2}{1 + x^\top x} (-Qx + \lambda(x)x). \tag{71}$$

(c) Describe the set of points for which the gradient is zero.

Solution: The gradient vanishes if and only if $Qx = \lambda(x)x$. Hence the set of points for which the gradient vanishes are the eigenvectors of Q with corresponding eigenvalue $\lambda(x)$. In the case when $\lambda(x) = 0$, x must be orthogonal to z.

(d) Let $Q = zz^\top$. Compute the Hessian of f at a point x. Again, we expect a symbolic answer in terms of Q and x. Don't consider partial derivatives with respect to components of x.

Hints:

i. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ denote the gradient map, $g(x) = \nabla f(x)$. First find the Jacobian of g denoted as $Dg(x)$.

Recall the product rule for finding the Jacobian of $g(x) = v(x)s(x)$, when $g(x)$ is a product of a vector-valued function $v(x)$ and a scalar valued function $s(x)$. We have,

$$Dg(x) = Dv(x)s(x) + v(x)Ds(x).$$

Note that the first term is a product of a matrix and a scalar while the second term is the product of a column vector with a row vector.

ii. Recall that the Hessian of f is related to the Jacobian of the gradient of f as $\nabla^2 f(x) = (Dg(x))^\top$. Additionally if f is twice differentiable, the Hessian will be symmetric.

Solution: We start out with the expression for gradient from part b. We have the gradient $g(x)$ given by,

$$g(x) = \frac{2}{1+x^\top x}(-Qx + \lambda(x)x), \quad \lambda(x) \doteq \frac{x^\top Qx}{1+x^\top x}. \tag{72}$$

We will compute the Jacobian of $g(x)$ denoted by $Dg(x)$. Recall that the Jacobian is linear so $Dg(x) = Da(x) + Db(x)$ with $a(x) \doteq \frac{-2}{1+x^\top x}Qx$ and $b(x) = \frac{2\lambda(x)}{1+x^\top x}x$. Next we describe how to compute $Da(x)$.

Observe that $a(x)$ can be expressed as the product of a vector-valued function $v(x)$ and a scalar-valued function $s(x)$ as,

$$a(x) = v(x)s(x), \quad v(x) = Qx, \quad s(x) = \frac{-2}{1+x^\top x}. \tag{73}$$

Using the identity from Hint i. we have,

$$\begin{aligned} Da(x) &= Dv(x)s(x) + v(x)Ds(x) && (74)\\ &= (Q)s(x) + v(x)\left(\frac{2}{(1+x^\top x)^2}2x^\top\right) && (75)\\ &= -\frac{2}{1+x^\top x}Q + \frac{4}{(1+x^\top x)^2}Qxx^\top. && (76)\end{aligned}$$

We can use a similar approach to compute $Db(x)$ by considering $b(x) = v(x)s(x)$ with $v(x) = x$ and $s(x) = \frac{2\lambda(x)}{1+x^\top x} = \frac{2x^\top Qx}{(1+x^\top x)^2}$. This will yield,

$$Db(x) = \frac{2x^\top Qx}{(1+x^\top x)^2}I_n + \frac{4}{(1+x^\top x)^2}xx^\top Q^\top - \frac{8x^\top Qx}{(1+x^\top x)^3}xx^\top. \tag{77}$$

Putting everything together we have,

$$\begin{aligned} Dg(x) &= Da(x) + Db(x) && (78)\\ &= -\frac{2}{1+x^\top x}Q + \frac{2x^\top Qx}{(1+x^\top x)^2}I_n + \frac{4}{(1+x^\top x)^2}Qxx^\top + \frac{4}{(1+x^\top x)^2}xx^\top Q^\top - \frac{8x^\top Qx}{(1+x^\top x)^3}xx^\top. && (79)\end{aligned}$$

Finally we obtain the Hessian of f, $\nabla^2 f(x)$ as $\nabla^2 f(x) = (Dg(x))^\top$. Note that since Q is symmetric the Jacobian of g is symmetric and is equal to the Hessian of f.

5. Homework Process

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

NOTE: If you didn't work with anyone, you can put "none" as your answer.