①

① This question asks to find the SVD of $A = [1, 1]$ without computation. → $A = [1, 1] = U\Sigma V^T$

From the previous homework, we defined U as the orthonormal basis for the column space of A. This is simply $[1]$.

Σ is the eigenvalue matrix. The rank of the matrix A is 1, so there is only one eigenvalue. To calculate it, in the last homework assignment, we figured that AA^T is diagonalizable, giving

$$AA^T x = \lambda^2 x$$

$$[1\ 1]\begin{bmatrix}1\\1\end{bmatrix} x = \lambda^2 x$$

$$2x = \lambda^2 x$$

$$\lambda^2 = 2$$

$$\lambda = \sqrt{2}$$

Since there are more columns than rows, we need to have Σ be a rectangular matrix of size 1×2. The other element in Σ, therefore is 0.

$$\Sigma = [\sqrt{2}, 0]$$

↳ $\left[\begin{matrix}\sigma_1 & 0\\ 0 & \sigma_n\end{matrix}\middle|\begin{matrix}0 & \cdots & 0\\ 0 & \cdots & 0\end{matrix}\right] \rightarrow [\sigma_1, 0]$

orthonormal basis of

Next, we need to find V^T. V is the row space and nullspace of matrix A because there are more columns than rows. The rowspace of A is $[1\ 1]$ and making it orthonormal, we get $\frac{1}{\sqrt{2}}[1\ 1] = [\sqrt{2}/2\ \sqrt{2}/2]$.

The nullspace of A is calculated by

$$Ax = 0$$

$$x_1 + x_2 = 0$$

$$x_1 = -x_2$$

$$[1 \ -1]$$

To make it orthonormal, we get $\frac{1}{\sqrt{2}}[1 \ -1] = [\sqrt{2}/2 \ \ -\sqrt{2}/2]$

Therefore, V is $\begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}$.

Combining it all together, we get

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

$$= [1][\sqrt{2} \ 0] \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}^T$$

$$= [1][\sqrt{2} \ 0] \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}$$

①②

This question asks to prove that $\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \theta_1$, the maximum singular value.

To prove that $\|A\|_2 = \theta_1$, we need to show that $\|A\|_2 \leq \theta_1$ and plug in θ_1 to the equation.

First, we will prove that $\|A\|_2 \leq \theta_1$. In other words, $\max_x \frac{\|Ax\|_2}{\|x\|_2} \leq \theta_1$.

$\|Ax\|_2 = \|(Ax)^T(Ax)\|_2 = \|x^T A^T A x\|_2$ by definition. From the previous assignment, we proved that A^TA matrix can be written as $U\Sigma^2 U^T$, where U is the orthonormal eigenvector matrix and Σ^2 is the eigenvalue squared diagonal matrix. Therefore,

$$\|A\|_2 = \max_x \frac{\|x^T U\Sigma^2 U^T x\|_2}{\|x\|_2} = \max_x \frac{\|x^T U\Sigma \cdot \Sigma U^T x\|_2}{\|x\|_2}$$

$$= \max_x \frac{\|(\Sigma^T U^T x)^T (\Sigma U^T x)\|_2}{\|x\|_2}$$

Since $\Sigma^T = \Sigma$ (diagonal matrix),

$$\|A\|_2 = \max_x \frac{\|\Sigma U^T x\|_2}{\|x\|_2}$$

Once again, Σ is simply eigenvalue matrix, we can take it out of the norm.

$$\|A\|_2 = \max_x \left(\theta_i \frac{\|U^T x\|_2}{\|x\|_2} \text{ for } i = 1, \ldots, n \right)$$

By definition, U^T is a unitary matrix, and unitary matrix multiplied by a vector x only changes the shape of the vector, leaving the vector length unchanged. Therefore, $\|U^T x\|_2 = \|x\|_2$.

$$\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \max_x \left(\theta_i \frac{\|U^T x\|_2}{\|x\|_2} \text{ for } i = 1, \ldots, n \right)$$

$$= \max_x \left(\theta_i \frac{\|x\|_2}{\|x\|_2} \text{ for } i = 1, \ldots, n \right)$$

Since $\|x\|_2$ is nothing but scalar value, we can cancel them out.

$$\|A\|_2 = \max_x (\theta_i \text{ for } i = 1, \ldots, n)$$

Now, by definition of SVD, θ_1 is the largest diagonal element of Σ (largest eigenvalue), which means that all elements (eigenvalues) are less than or equal to θ_1.

$$\|A\|_2 = \max_x (\theta_i \text{ for } i = 1, \ldots, n) \leq \theta_1.$$

Therefore, we proved that $\|A\|_2 \leq \theta_1$.

Now, we need to plug in the eigenvector of θ_1 into the vector x to complete the proof that $\|A\|_2 = \theta_1$.

Denote the eigenvector of θ_1 as u_1.

$$\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \max_x \frac{\|x^T A^T A x\|_2}{\|x\|_2} = \max_x \frac{\|x^T (A^T A) x\|_2}{\|x\|_2}$$

Since $A^T A$ is diagonalizable, if we plug in θ_1 and corresponding eigenvector u_1,

$$\|A\|_2 = \max_x \frac{\|u_1^T a_1^T a_1 u_1\|_2}{\|u_1\|_2} = \max_x \frac{\|u_1^T u_1 \theta_1\|_2}{\|u_1\|_2}$$

↳ $a_1^T a_1 = u_1 \theta_1 u_1^T$

$a_1^T a_1 u_1 = u_1 \theta_1 u_1^T \cdot u_1$

Once again, σ_1 is simply a scalar value, which means we can take it out.

$$\|A\|_2 = \max_x \sigma_1 \frac{\|u_1^T u_1\|_2}{\|u_1\|_2}$$

Since u_1 and u_1^T are orthonormal vectors, $\|u_1\|_2 = 1$ and $\|u_1^T u_1\|_2 = 1$

Therefore, $\|A\|_2 = \max_x \sigma_1 \cdot \frac{1}{1} = \max_x \sigma_1 = \sigma_1$ ✓

By plugging in the eigenvector of σ_1, we proved that $\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1$.

② ①

$f(x) = \sin(x) + \cos(x)$ at $x=0$ for degree 5 taylor polynomial

$$T_5(x) = f(0) + \frac{f'(0)(x-0)}{1!} + \frac{f''(0)(x-0)^2}{2!} + \frac{f'''(0)(x-0)^3}{3!}$$

$$+ \frac{f''''(0)(x-0)^4}{4!} + \frac{f'''''(0)(x-0)^5}{5!}$$

$f(0) = \sin(0) + \cos(0) = 1$

$f'(0) = \cos(0) - \sin(0) = 1$

$f''(0) = -\sin(0) - \cos(0) = -1$

$f'''(0) = -\cos(0) + \sin(0) = -1$

$f''''(0) = \sin(0) + \cos(0) = 1$

$f'''''(0) = \cos(0) - \sin(0) = 1$

$$T_5(x) = 1 + 1(x) + \frac{(-1)(x)^2}{2} + \frac{(-1)(x)^3}{6} + \frac{(1)(x)^4}{24} + \frac{(1)(x)^5}{120}$$

$$= 1 + x - \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}$$

②

②

$f(x,y) = x^2 + y^2 + 2xy - 3x + 2y + 5$ at $x=5, y=10$

Let $a=5, b=10$

$f(x,y) \approx f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-10)$
$+ f_{xx}(a,b)(x-a)^2/2 + f_{xy}(a,b)(x-5)(y-10)$
$+ f_{yy}(a,b)(y-b)^2/2$

$f(5,10) = 25 + 100 + 100 - 15 + 20 + 5$
$= 235$

$f_x(5,10) = 2x + 2y - 3$
$= 2(5) + 2(10) - 3$
$= 27$

$f_{xx}(5,10) = 2$

$f_y(5,10) = 2y + 2x + 2$
$= 2(10) + 2(5) + 2$
$= 32$

$f_{xy}(5,10) = f_{yx}(5,10) = 2$

$f_{yy}(5,10) = 2$

$f(x,y) \approx 235 + 27(x-5) + 32(y-10) + \frac{2(x-5)^2}{2} + 2(x-5)(y-10)$
$+ 2(y-10)^2/2$
$\approx 235 + 27x - 135 + 32y - 320 + x^2 - 10x + 25$
$+ 2(xy - 10x - 5y + 50) + y^2 - 20y + 100$
$\approx x^2 + y^2 + 2xy + 27x - 10x - 20x + 32y - 10y - 20y$
$+ 235 - 135 - 320 + 25 + 100 + 100$
$\approx x^2 + y^2 + 2xy - 3x + 2y + 5$

↳ same as the original function

The quadratic approximation of function $f(x,y) = x^2 + y^2 + 2xy - 3x + 2y + 5$ at $x=5$ and $y=10$ is itself.

③ (a)

$$f(x) = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1}, \quad x \in \mathbb{R}$$

$f: \mathbb{R} \to \mathbb{R}$

$$\frac{df}{dx} = \frac{-1}{(1+e^{-x})^2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2}$$

(b)

$$f(x) = \exp(-1/2\theta^2 (x-\mu)^2) = e^{\frac{-1}{2\theta^2}(x-\mu)^2}, \quad x \in \mathbb{R}$$

$f: \mathbb{R} \to \mathbb{R}$

$$f(u) = e^u, \quad u = \frac{-1}{2\theta^2}(x-\mu)^2$$

$$\frac{df}{du} = e^u \qquad \frac{du}{dx} = \frac{-2}{2\theta^2}(x-\mu), \qquad \frac{df}{dx} = \frac{df}{du} \cdot \frac{du}{dx}$$

$$\frac{df}{dx} = e^u \cdot \frac{-1}{\theta^2}(x-\mu)$$

$$= \exp(-1/2\theta^2 (x-\mu)^2) \cdot \frac{-1}{\theta^2}(x-\mu)$$

(c) $f(x) = \sin(x_1)\cos(x_2)$, $x \in \mathbb{R}^2$

$f: \mathbb{R}^2 \to \mathbb{R}^{1\times 2}$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \cos(x_1)\cos(x_2) & -\sin(x_1)\sin(x_2) \end{bmatrix} \in \mathbb{R}^{1\times 2}$$

(d) $f(x) = xx^T$, $x \in \mathbb{R}^n$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_1x_2 & x_2x_2 & \cdots & x_2x_n \\ \vdots & & & \vdots \\ x_1x_n & x_2x_n & \cdots & x_nx_n \end{bmatrix}$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

$$\begin{bmatrix} 2x_1 & x_2 & \cdots & x_n \\ x_2 & & & \\ \vdots & & \bigcirc & \\ x_n & & & \end{bmatrix}^{n\times n} \quad \begin{bmatrix} 0 & x_1 & & \\ x_1 & 2x_2 & \cdots & x_n \\ 0 & \vdots & & \bigcirc \\ & x_n & & \end{bmatrix}^{n\times n} \quad \begin{bmatrix} & x_1 & \\ & x_2 & \\ & \vdots & \\ x_1 & x_2 & \cdots \; 2x_n \end{bmatrix}^{n\times n}$$

Each $\frac{\partial f}{\partial x_i}$ where $i = 1, \ldots, n$ has 0s other than the i^{th} row and i^{th} column since there are no terms regarding x_i in other areas in the matrix. This means the derivative of that value respect to x_i is 0.
Regarding the elements in the i^{th} row and i^{th} column, the element a_{ij} (element at i^{th} row and j^{th} column) or a_{ji}, when $i \neq j$, the values are simply x_j.
When $i = j$, the value is $2x_i$.
Ex) when $i = 1$, $a_{11} = 2x_1$, $a_{12} = x_2$, ..., $a_{1n} = x_n$
, $a_{21} = x_2$, ..., $a_{n1} = x_n$ $\rightarrow$ $\begin{bmatrix} 2x_1 & x_2 & \cdots & x_n \\ x_2 & & & \\ \vdots & & \bigcirc & \\ x_n & & & \end{bmatrix}$
and other areas are 0s

(e) $f(x) = \sin(\log(x^T x)), x \in \mathbb{R}^n$

$= \sin(u)$

$f(x) = \sin(u)$

$u = \log(z)$

$z = x^T x$

we first look at z $\quad z: \mathbb{R}^n \to \mathbb{R}^{1\times n}$

$\frac{dz}{dx} = \frac{d}{dx}(x_1^2 + x_2^2 + \dots + x_n^2)$ Let $\Delta = x_1^2 + \dots + x_n^2$

$= \left[\frac{\partial \Delta}{\partial x_1} \quad \frac{\partial \Delta}{\partial x_2} \dots \frac{\partial \Delta}{\partial x_n}\right]$

$= [2x_1 \; 2x_2 \cdots 2x_n]$

$= 2x^T$

$\frac{du}{dz} = \frac{1}{z}, \; z = x^T x$

$\frac{df}{du} = \cos(u), \; u = \log(x^T x)$

$\frac{df}{dx} = \frac{df}{du}\frac{du}{dz}\frac{dz}{dx}$

$= \frac{\cos(\log(x^T x))}{x^T x} \cdot 2x^T$

Since $x^T x = \|x\|_2^2$ (scalar value),

$= \frac{2\cos(\log(\|x\|_2^2))}{\|x\|_2^2} \cdot x^T \in \mathbb{R}^{1\times n}$

(f) $f(z) = \log(1+z)$ where $z = x^T x$, $x \in \mathbb{R}^n$

$f(z) = \log(1+z)$

$z = x^T x$

We first look at z $\quad z: \mathbb{R}^n \to \mathbb{R}^{1\times n}$

$\frac{dz}{dx} = 2x^T$ (from previous section)

$\frac{df}{dz} = \frac{1}{1+z}, \quad z = x^T x$

$$\frac{df}{dx} = \frac{df}{dz}\frac{dz}{dx} = \frac{1}{1+z}\cdot 2x^T$$

$$= \frac{1}{1+x^T x}\cdot 2x^T$$

$$= \frac{2}{1+\|x\|_2^2}\cdot x^T \in \mathbb{R}^{1\times n}$$

↳ scalar value

(g) $f(x) = x^T A x$ where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$

$$x^T = [x_1 \cdots x_n]^{1 \times n}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}^{n \times n}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^{n \times 1}$$

Now, we calculate $x^T A x$ in terms of $x_{1 \ldots n}$ and a.
First, we multiply x^T by A

$$x^T A = [x_1 a_{11} + \ldots + x_n a_{n1} \quad \cdots \quad x_1 a_{1n} + \ldots + x_n a_{nn}]$$

$$= \left[\sum_{i=1}^{n} x_i \cdot a_{i1} \quad \cdots \quad \sum_{i=1}^{n} x_i \cdot a_{in}\right]^{1 \times n}$$

Now, we calculate $x^T A x$

$$x^T A \cdot x = \left[x_1 \sum_{i=1}^{n} x_i \cdot a_{i1} + \ldots + x_n \sum_{i=1}^{n} x_i \cdot a_{in}\right]^{1 \times 1}$$

Now, we have to take derivative of $x^T A x = f(x)$

$$\nabla f(x) = \left[\frac{\partial x^T A x}{\partial x_1} \quad \frac{\partial x^T A x}{\partial x_2} \quad \cdots \quad \frac{\partial x^T A x}{\partial x_n}\right] \in \mathbb{R}^{1 \times n} \text{ where}$$

$$\frac{\partial x^T A x}{\partial x_i} = \sum_{j=1}^{n} x_j a_{ij} + 2 x_i a_{ii} + \sum_{j=1}^{n} x_j a_{ji}$$

when $i \neq j$ when $i \neq j$

For example, $\frac{\partial x^T A x}{\partial x_2} = \sum_{j=1}^{n} x_j a_{2j} + 2 x_2 a_{22} + \sum_{j=1}^{n} x_j \cdot a_{j2}$

④①

$f(x) = x^3 + 6x^2 - 3x - 5$

To find stationary points, take derivative and set equal to 0.

$f'(x) = 3x^2 + 12x - 3 = 0$

$= x^2 + 4x - 1 = 0$

$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

$x = \frac{-4 \pm \sqrt{16+4}}{2} = \frac{-4 \pm \sqrt{20}}{2} = \frac{-4 \pm 2\sqrt{5}}{2} = -2 \pm \sqrt{5}$

$= -4.236,\ 0.236$

$f''(x) = 6x + 12$

To check whether it is maximum, minimum, or saddle point, we need to plug in the x stationary points into the second derivative and check whether they are 0, positive, or negative.

$f''(0.236) = 13.416 > 0 \rightarrow$ min

$f''(-4.236) = -13.416 < 0 \rightarrow$ max

The stationary points are $x = -4.236$ and 0.236.
$x = -4.236$ is when the function is at maximum, and
$x = 0.236$ is when the function is at minimum.

④ ② (i)

When solving least squares loss in a linear model using gradient descent...

we want to solve for $\min_{\theta \in \mathbb{R}^D} (\|y - X \cdot \theta\|^2)$

where θ is the parameter vector of length D, X is $n \times D$ input feature matrix, and y are corresponding observations of length n.

We find the minimum θ since we want to minimize the error. Let $e = y - X \cdot \theta$ and $L = \|e\|^2$. Therefore, we want to calculate $\min_{\theta \in \mathbb{R}^D} (L)$.

We can now use gradient descent and calculate partial derivatives

$L = \|e\|^2$

$= e^T e = (e_1^2 + e_2^2 + \dots + e_n^2)$ Let $\Delta = e^T e$

$$\frac{\partial L}{\partial e} = \left[\frac{\partial \Delta}{\partial e_1}, \dots \frac{\partial \Delta}{\partial e_n}\right] = [2e_1 \dots 2e_n] = 2e^T$$

Now, we calculate $\frac{\partial e}{\partial \theta} = \left[\frac{\partial e}{\partial \theta_1} \cdots \frac{\partial e}{\partial \theta_n}\right]$, where each

$\frac{\partial e}{\partial \theta_i}$ for $i = 1, \dots, n$ is $-X_1, \dots, -X_n \rightarrow n^{th}$ column of matrix X

↓ first columns of matrix X

Therefore, $\frac{\partial e}{\partial \theta} = [-X_1 \dots -X_n]$, which is simply $-X$

↓ matrix X.

Now, $\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial e} \frac{\partial e}{\partial \theta} = -2e^T \cdot X$ where $e = y - X \cdot \theta$

$= -2(y^T - \theta^T X^T) X$

Setting it equal to 0 and moving the variables, we get

$$0 = -2(y^T - \theta^T X^T)X$$
$$y^T X = \theta^T X^T X$$
$$y^T X (X^T X)^{-1} = \theta^T$$

Now, we know this is the minimum because the Hessian of $\|y - X\cdot\theta\|^2$ equals $X^T X$ is positive semi definite

(ii)

By using SVD, just like how we did in last homework assignment, we can do it by

$$\min_x \|Ax - b\|_2 = \|AVV^T x - b\|_2 \rightarrow \text{since multiplying by } VV^T = I \text{ is just A}$$
$$= \|U^T AVV^T x - U^T b\|_2 \rightarrow \text{since } U^T \text{ is unitary matrix, doesn't change the length}$$
$$= \|U^T U \Sigma V^T V V^T x - U^T b\|_2 \rightarrow A = U\Sigma V^T$$
$$= \|\Sigma V^T x - U^T b\|_2$$
$$= \sum_{i=1}^{r} (\theta_i v_i^T x_i - u_i^T b)^2 + \sum_{i=r+1}^{m} (u_i^T b)^2$$

↳ since $m > n$

Now, make $\sum_{i=1}^{r} (\theta_i v_i^T x_i - u_i^T b) = 0$, which will equal $\sum_{i=r+1}^{m} (u_i^T b)^2$

$$\theta_i v_i^T x_i = u_i^T b \text{ for } i = 1, \ldots, r$$
$$v_i^T x_i = \frac{u_i^T b}{\theta_i} \text{ for } i = 1, \ldots, r$$
$$x_i = \frac{v_i u_i^T b}{\theta_i} \text{ for } i = 1, \ldots, r$$

↳ since v_i^T is orthonormal

Therefore, x^* equals $\sum_{i=1}^{r} \left(\frac{v_i u_i^T b}{\theta_i}\right)$

Some pros & cons for SVD and gradient descent

- gradient descent
  - pros
    - can handle large datasets that don't fit memory
    - can handle non-linear models
  - cons
    - can sometimes get stuck on local minima or saddle points
    - computation is very expensive and timely if dataset is large
    - requires to take inverse
- SVD
  - pros
    - more computationally efficient for small number of datasets
    - can handle cases where it is rank deficient
    - do not need to calculate for all i but first k and change accuracy
  - cons
    - if the dataset is large, it may not scale well

Generally, if the dataset is large, it is efficient to use gradient descent and when dataset isn't large enough, it is efficient to use SVD