

1.1. SVD of  $A = [1, 1]$

$$U = [1]$$

$$\Sigma = \sqrt{2}$$

$$V^T = [1/\sqrt{2}, 1/\sqrt{2}]$$

1.2. 2. (15 pts) Let  $A \in \mathbb{R}^{m \times n}$  and let  $\sigma_1$  be the maximum singular value of  $A$ . For  $x \in \mathbb{R}^n \setminus \{0\}$  the spectral norm of  $A$  is defined as  $\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2}$ . Prove that

$$\|A\|_2 = \sigma_1.$$

Proof:  $\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$

$$\sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|U \Sigma V^T x\|_2}{\|x\|_2}$$

since  $U$  is orthonormal  
 $\|U\| = 1$

$$= \sup_{x \neq 0} \frac{\|\Sigma V^T x\|_2}{\|x\|_2}$$

let  $y = V^T x \Rightarrow Vy = x$

$$= \sup_{y \neq 0} \frac{\|\Sigma y\|_2}{\|Vy\|_2}$$

$$= \sup_{y \neq 0} \frac{(\sum_{i=1}^r \sigma_i^2 |y_i|^2)^{\frac{1}{2}}}{(\sum_{i=1}^r |y_i|^2)^{\frac{1}{2}}} \leq \sigma_1.$$

For  $y = [1 \ 0 \ \dots \ 0]^T$ ,  $\|\Sigma y\|_2 = \sigma_1$ , and the supremum is attained, which correspond to  $x = V_1$ . (Hence,  $Av_1 = \sigma_1 u_1$ .)

Therefore,  $\|A\|_2 \stackrel{\Delta}{=} \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1 = \sigma_{\max}(A)$

2.1.  $f(x) = \sin(x) + \cos(x)$

center around 0:

$$f'(x) = \cos(x) - \sin(x)$$

$$f(0) = 0 + 1 = 1$$

$$f''(x) = -\sin(x) - \cos(x)$$

$$f'(0) = 1 - 0 = 1 = f'(0)$$

$$f'''(x) = -\cos(x) + \sin(x)$$

$$f''(0) = -0 - 1 = -1$$

$$f^{(4)}(x) = \sin(x) + \cos(x)$$

$$f^{(3)}(0) = -1 + 0 = -1$$

$$f^{(5)}(x) = \cos(x) - \sin(x)$$

$$f^{(4)}(0) = 0 + 1 = 1$$

$$T(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

degree 5 Taylor polynomial center around 0:

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

2.2.  $f(x, y) = x^2 + y^2 + 2xy - 3x + 2y + 5$

$$\nabla f(x, y) = (2x + 2y - 3, 2y + 2x + 2)$$

$$H_f = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

$$Q_f(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T H_f(x_0)(x - x_0)$$

Given  $x_0 = (5, 10)$

$$Q_f(x, y) = f(5, 10) + \nabla f(5, 10) \cdot \begin{pmatrix} x-5 \\ y-10 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} x-5 & y-10 \end{pmatrix}^T \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} x-5 \\ y-10 \end{pmatrix}$$

$$f(5, 10) = 5^2 + 10^2 + 2 \times 5 \times 10 - 3 \times 5 + 2 \times 10 + 5 = 235$$

$$\nabla f(5, 10) = (10 + 20 - 3, 20 + 10 + 2) = (27, 32)$$

$$\begin{pmatrix} x-5 & y-10 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} x-5 \\ y-10 \end{pmatrix} = (2x-10 + 2y-20 \quad 2x-10 + 2y-20) \begin{pmatrix} x-5 \\ y-10 \end{pmatrix}$$

$$= (2x + 2y - 30 \quad 2x + 2y - 30) \begin{pmatrix} x-5 \\ y-10 \end{pmatrix} = 2x^2 + 2xy - 30x - 10x - 10y + 150 + 2xy + 2y^2 - 30y - 20x - 20y + 300$$

$$= 2x^2 + 4xy + 2y^2 - 60x - 60y + 450$$

Therefore,  $Q_f(x, y) = 235 + (27, 32) \begin{pmatrix} x-5 \\ y-10 \end{pmatrix} + \frac{1}{2}(2x^2 + 4xy + 2y^2 - 60x - 60y + 450)$

$$= 235 + 27x - 135 + 32y - 320 + x^2 + 2xy + y^2 - 30x - 30y + 225$$

$$= 5 - 3x + 2y + x^2 + 2xy + y^2$$

### 3. Derivatives.

(1)  $f(x) = \frac{1}{1+e^{-x}}$ ,  $x \in \mathbb{R}$ ,  $f: \mathbb{R}' \rightarrow \mathbb{R}'$   $(\nabla f)^{1 \times 1}$

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

$$\stackrel{(1 \times 1)}{=} e^{-x}(1+e^{-x})$$

(2)  $f(x) = \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$ ,  $x \in \mathbb{R}$ ,  $f: \mathbb{R}' \rightarrow \mathbb{R}'$   $(\nabla f)^{1 \times 1}$

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

$$\stackrel{(1 \times 1)}{=} -\frac{1}{\sigma^2}(x-\mu)^2 \cdot \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

(3)  $f(x) = \sin(x_1)\cos(x_2)$ ,  $x \in \mathbb{R}^2$ ,  $f: \mathbb{R}^2 \rightarrow \mathbb{R}'$   $(\nabla f)^{1 \times 2}$

$$\frac{\partial f}{\partial x_1} = \cos(x_1)\cos(x_2) \quad \frac{\partial f}{\partial x_2} = -\sin(x_1)\sin(x_2)$$

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \cos(x_1)\cos(x_2) & -\sin(x_1)\sin(x_2) \end{bmatrix}$$

$$\stackrel{(1 \times 2)}{}$$

(4)  $f(x) = x x^T$ ,  $x \in \mathbb{R}^n$ ,  $f: \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$   
 $n \times 1 \times 1 \times n^T$  Resulting  $\nabla f: n \times n \times n$

$$f(x) = \begin{bmatrix} x_1^2 & x_1 x_2 & \dots & x_1 x_n \\ x_2 x_1 & x_2^2 & & \vdots \\ \vdots & \boxed{x_i x_j} & \ddots & \vdots \\ x_n x_1 & \dots & \dots & x_n^2 \end{bmatrix}$$

$$f(x_1+h, x_2, \dots, x_n) = \begin{bmatrix} (x_1+h)^2 & (x_1+h)x_2 & \dots & (x_1+h)x_n \\ (x_1+h)x_2 & x_2^2 & & \vdots \\ \vdots & & \ddots & x_n^2 \\ (x_1+h)x_n & \dots & \dots & x_n^2 \end{bmatrix}$$

$$\frac{\partial f}{\partial x_1} = \lim_{h \rightarrow 0} \frac{f(x_1+h, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h} = \lim_{h \rightarrow 0} \frac{1}{h} \begin{bmatrix} (x_1+h)^2 - x_1^2 & h x_2 & \dots & h x_n \\ h x_2 & 0 & & 0 \\ \vdots & & \ddots & \\ h x_n & 0 & & 0 \end{bmatrix}$$

$$= \lim_{h \rightarrow 0} \begin{bmatrix} \frac{(x_1+h)^2 - x_1^2}{h} & x_2 & \dots & x_n \\ x_2 & 0 & & 0 \\ \vdots & & \ddots & \\ x_n & 0 & & 0 \end{bmatrix}$$

Similar process can be done to find  $\frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \dots, \frac{\partial f}{\partial x_n}$ , where each partial derivative of  $x_i$  with respect to  $f$  is a  $n \times n$  matrix.

Therefore, the derivative of  $f$  is a vector including  $n$   $n \times n$  matrix

$$\nabla f = \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right] = \left[ \lim_{h \rightarrow 0} \begin{bmatrix} \frac{(x_1+h)^2 - x_1^2}{h} & x_2 & \dots & x_n \\ \vdots & & & \\ x_n & 0 & & \end{bmatrix}, \dots, \lim_{h \rightarrow 0} \begin{bmatrix} 0 & \dots & x_n \\ \vdots & & \vdots \\ x_1 & \dots & \frac{(x_n+h)^2 - x_n^2}{h} \end{bmatrix} \right]$$

(5)  $f(x) = \sin(\log(x^T x))$ ,  $x \in \mathbb{R}^n$ ,  $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$   $(\nabla f)^{1 \times n}$

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

$(1 \times n)$

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

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

$$z: \mathbb{R}^n \rightarrow \mathbb{R}^1$$

$(1 \times n)$

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

$(\nabla f)^{1 \times n}$

(7)  $f(x) = x^T A x$ , where  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$ ,  $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$

$$|1 \times n \quad n \times n \quad n \times 1|$$

$\Rightarrow (\nabla f)^{1 \times n}$

$$\text{let } w = Ax \Rightarrow f(x) = x^T w$$

$$\frac{\partial f}{\partial x} = x^T \frac{dw}{dx} + w^T \frac{\partial x^T}{\partial x}$$

$(1 \times n)$

$$= x^T A + w^T$$

$$= x^T A + x^T A^T$$

$$= x^T (A + A^T)$$

#### 4. Optimization.

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

$$\frac{df}{dx} = 3x^2 + 12x - 3 = 0$$

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

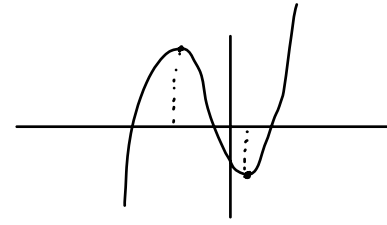
$$(x+2)^2 = 5$$

$$x+2 = \sqrt{5} \quad x+2 = -\sqrt{5}$$

$$x_1 = \sqrt{5} - 2 \quad x_2 = -2 - \sqrt{5} \quad \leftarrow \text{two stationary points.}$$

when  $x = \sqrt{5} - 2$   $f(x) \approx -5.361$ , which is a local minimum.

when  $x = -2 - \sqrt{5}$ ,  $f(x) \approx 39.361$ , which is a local maximum.



4.2. Given a linear system as  $y = Ax + e$ , where  $x$  is the input,  $y$  is the output and  $e$  is the noise term.  $A$  as the system parameter is a known  $p \times q$  matrix. To minimize the error term, we can minimize  $\|\hat{y} - Ax\|_2$ , where  $\hat{y}$  is the observation.

$$\hat{x} = \arg \min_{x: \hat{y} = Ax + e} \|e\|_2$$

(i) using gradient descent.

$$f(x) = \|e\|^2, \quad e = \hat{y} - Ax$$

$$\frac{\partial f}{\partial e} = 2e^T \quad \frac{\partial e}{\partial x} = -A$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial e} \frac{\partial e}{\partial x} = -2(\hat{y}^T - x^T A^T)A$$

$$\text{Set } \frac{\partial f}{\partial x} = 0 \Rightarrow -2(\hat{y}^T - x^T A^T)A = 0$$

$$\hat{y}^T A = x^T \underbrace{A^T A}$$

if  $A$  is full rank,

$A^T A$  has inverse

$$\hat{y}^T A (A^T A)^{-1} = x^T$$

$$x^* = (A^T A)^{-1} A^T \hat{y}$$

(2) using SVD. (same as HW8 3.2)

To find  $\min \|\hat{y} - Ax\|^2$

$$\begin{aligned}\|\hat{y} - Ax\|^2 &= \|\hat{y} - (U\Sigma V^T)x\|^2 \\&= \|U^T(\hat{y} - UV^Tx)\|^2 \quad \text{as } U^T \text{ is orthonormal, } \|U^T\|=1 \\&= \|U^T\hat{y} - U^T U \Sigma V^T x\|^2 \\&= \|U^T\hat{y} - \Sigma V^T x\|^2 \quad \text{let } z = V^T x \\&= \underbrace{\sum_{i=1}^r (u_i^T \hat{y}_i - \sigma_i z_i)^2}_{\text{minimize}} + \sum_{i=r+1}^p (u_i^T \hat{y}_i)^2\end{aligned}$$

To minimize  $\|\hat{y} - Ax\|^2$ , we only need to minimize  $\sum_{i=1}^r (u_i^T \hat{y}_i - \sigma_i z_i)^2$

Set it as zero  $\Rightarrow z_i = \frac{u_i^T \hat{y}_i}{\sigma_i}$ , for  $i=1, 2, \dots, r$ .

which make  $\|\hat{y} - Ax\|^2$  as minimum  $\sum_{i=r+1}^p (u_i^T \hat{y}_i)^2$ .

Therefore, we can find actual  $x^*$ .

$$z = V^T x^* = \frac{U^T \hat{y}}{\Sigma}$$

$$x^* = \frac{U^T \hat{y} V}{\Sigma} = \sum_{i=1}^r \left( \frac{u_i^T \hat{y}}{\sigma_i} \right) v_i$$

Discussion:

- When we have large data, for example matrix  $A$  is  $100,000 \times 1,000$  where there is 100,000 rows of observation, it is inefficient to calculate  $A^T A$  in results given by gradient descent as well as inverting a  $1000 \times 1000$  matrix. Therefore, using gradient is computational heavier.
- On the other hand, SVD use the idea of approximation, which calculates only first  $k$  singular value. This can make the computation faster when dealing with large data.