Instructions

- The homework is due on **Tuesday 4/25 at 5pm ET** before the lecture starts.
- There are 15 points available for extra credit.
- No extension will be provided, unless for serious documented reasons.
- Start early!
- Study the material taught in class, and feel free to do so in small groups, but the solutions should be a product of your own work.
- This is not a multiple choice homework; reasoning, and mathematical proofs are required before giving your final answer.

1 SVD again [20 points]

1. (5 pts) Find the SVD of $A = [1, 1]$ without the use of computing devices/software.
2. (15 pts) Let $A \in \mathbb{R}^{m\times n}$ and let σ_1 be the maximum singular value of A. For $x \in \mathbb{R}^n \backslash \{0\}$ the spectral norm of A is defined as $||A||_2 = \max_x \frac{||Ax||_2}{||x||_2}$. Prove that

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

2 Taylor polynomial approximation [10 points]

1. (5 pts) Let $f(x) = \sin(x) + \cos(x)$. Compute the degree 5 Taylor polynomial for f at $x = 0$.
2. (5 pts) Compute the quadratic approximation of the function $f(x, y) = x^2 + y^2 + 2xy - 3x + 2y + 5$ at the point $x = 5, y = 10$.

3 Derivatives [35 points]

Compute the derivative $\frac{df}{dx}$ for the following functions. It will be helpful to identify n, m where $f : \mathbb{R}^n \to \mathbb{R}^m$, and the dimensions of the derivative first.

(a) [5pts] $f(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$

(b) [5 pts] $f(x) = \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), x \in \mathbb{R}$

(c) [5 pts] $f(x) = \sin(x_1)\cos(x_2), x \in \mathbb{R}^2$. $f: \mathbb{R}^2 \to \mathbb{R}$

(d) [5 pts] $f(x) = xx^T, x \in \mathbb{R}^n$. $f: \mathbb{R}^n \to \mathbb{R}^{n\times n}$

(e) [5 pts] $f(x) = \sin(\log(x^T x)), x \in \mathbb{R}^n$.

(f) [5 pts] $f(z) = \log(1 + z)$ where $z = x^T x, x \in \mathbb{R}^n$

(g) [5 pts] $f(x) = x^T Ax$ where $x \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$.

4 Optimization [15 points]

1. (7.5 pts) Consider the univariate function $f(x) = x^3 + 6x^2 - 3x - 5$. Find its stationary points and indicate whether they are maximum, minimum or saddle points.

2. (7.5 pts) Explain how to solve the least squares loss in a linear model using (i) gradient descent and (ii) SVD. Discuss the pros and cons.

5 Coding [35 points]

Check the Jupyter notebook on our Git repo.

1.

1. (5 pts) Find the SVD of $A = [1, 1]$ without the use of computing devices/software.

$A^TA = \begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} = \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$

$\det(A^TA - \lambda I) = 0$

$\begin{bmatrix}1-\lambda & 1\\1 & 1-\lambda\end{bmatrix} = 0$

$(1-\lambda)^2 = 1$

$1-\lambda = \pm 1$

$\lambda = 0, 2$

$\lambda = 2$

$\begin{bmatrix}1-2 & 1\\1 & 1-2\end{bmatrix} \sim \begin{bmatrix}-1 & 1\\1 & -1\end{bmatrix} \sim \begin{bmatrix}1 & -1\\0 & 0\end{bmatrix}$

$\left[\begin{array}{cc|c}1 & -1 & 0\\0 & 0 & 0\end{array}\right] \Rightarrow \begin{array}{l}x_1 = x_2\\x_2 \text{ free}\end{array}$

$x = \begin{bmatrix}1\\1\end{bmatrix} x_2$

$V_1 = \begin{bmatrix}1\\1\end{bmatrix} \Rightarrow \begin{bmatrix}\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2}\end{bmatrix}$

$\lambda = 0$

$\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} \sim \begin{bmatrix}1 & 1\\0 & 0\end{bmatrix}$

$\left[\begin{array}{cc|c}1 & 1 & 0\\0 & 0 & 0\end{array}\right] \Rightarrow \begin{array}{l}x_1 + x_2 = 0\\x_2 \text{ free}\end{array}$

$x_1 = -x_2$

$x = \begin{bmatrix}-1\\1\end{bmatrix} x_2$

$V_2 = \begin{bmatrix}-1\\1\end{bmatrix} \Rightarrow \begin{bmatrix}-\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2}\end{bmatrix}$

$V = \begin{bmatrix}\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\end{bmatrix}$, $\Sigma = [\sqrt{2}, 0]$

$U_i = \frac{1}{\sigma} A V_i$

$U_1 = \frac{1}{\sqrt{2}}[1, 1]\begin{bmatrix}\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2}\end{bmatrix} = 1$

$A = [1] \cdot [\sqrt{2}, 0] \cdot \begin{bmatrix}\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\\-\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\end{bmatrix}$

2. (15 pts) Let $A \in \mathbb{R}^{m\times n}$ and let σ_1 be the maximum singular value of A. For $x \in \mathbb{R}^n \backslash \{0\}$ the spectral norm of A is defined as $||A||_2 = \max_x \frac{||Ax||_2}{||x||_2}$. Prove that

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

Proof:

$\because \; \|Ax\| = \sqrt{(Ax)^TAx} = \sqrt{x^TA^TAx} = \sqrt{x^T\lambda x}$, $\quad \sigma_{max} = \sigma_1 = \sqrt{\lambda_{max}}$

$\therefore \; \max \frac{\|Ax\|}{\|x\|} = \sqrt{\frac{x^T\lambda_{max}x}{x^Tx}} = \sqrt{\frac{\lambda_{max}\, x^Tx}{x^Tx}} = \sqrt{\lambda_{max}} = \sigma_1$ ∎

2.

1. (5 pts) Let $f(x) = \sin(x) + \cos(x)$. Compute the degree 5 Taylor polynomial for f at $x = 0$.

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

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

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

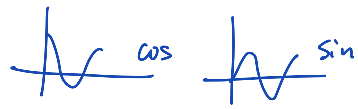
$f^3(0) = -\cos(0) + \sin(0) = -1$

$f^4(0) = \sin(0) + \cos(0) = 1$

$f^5(0) = 1$

$\sin' = \cos$

$\cos' = -\sin$



Since $x=0$, we have the special case of Taylor polynomial : Maclaurin.

$$\Rightarrow P_5(x) = f(0) + f'(0)\cdot x + \frac{f''(0)}{2!}x^2 + \frac{f^3(0)}{3!}x^3 + \frac{f^4(0)}{4!}x^4 + \frac{f^5(0)}{5!}x^5$$

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

2. (5 pts) Compute the quadratic approximation of the function $f(x, y) = x^2 + y^2 + 2xy - 3x + 2y + 5$ at the point $x = 5, y = 10$.

$$Q(x,y) = f(x,y) + \frac{df}{dx}(x - x_0) + \frac{df}{dy}(y - y_0)$$
$$+ \frac{1}{2}\frac{d^2f}{dx^2}(x-x_0)^2 + \frac{d^2f}{dxdy}(x-x_0)(y-y_0) + \frac{1}{2}\frac{d^2f}{dy^2}(y-y_0)^2$$

$f(x,y) = x^2+y^2+2xy-3x+2y+5 = 25+100+100-15+20+5 = 235$

$\frac{df}{dx} = 2x + 2y - 3 = 10 + 20 - 3 = 27$

$\frac{df}{dy} = 2y + 2x + 2 = 20 + 10 + 2 = 32$

$\frac{d^2f}{dx^2} = 2 \qquad \frac{d^2f}{dy^2} = 2 \qquad \frac{d^2f}{dxdy} = 2$

$$\Rightarrow Q(x,y) = 235 + 27(x-5) + 32(y-10) + (x-5)^2$$
$$+ 2(x-5)(y-10) + (y-10)^2$$

3.

a) $\frac{df}{dx} = \left(\frac{1}{1+e^{-x}}\right)' = \left[(1+e^{-x})^{-1}\right]'$

$= -1\,(1+e^{-x})^{-2} \cdot (-1)\, e^{-x}$

$= (1+e^{-x})^{-2} \cdot e^{-x}$

b) $f(x) = e^{-u} \qquad u = \frac{1}{2\sigma^2}(x-\mu)^2$

$\frac{df}{du} = -e^{-u}$

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

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

c) $f(x) = \sin(x_1)\cos(x_2), \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\frac{df}{dx} = \cos(x_1)\cos(x_2) - \sin(x_1)\sin(x_2)$

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

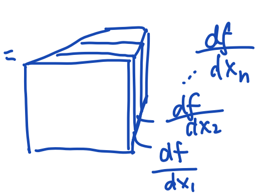
$$xx^T = \begin{bmatrix} x_1 x_1 & x_1 x_2 \cdots & x_1 x_n \\ x_2 x_1 & \ddots & \\ \vdots & & \\ x_n x_1 & & x_n x_n \end{bmatrix}$$

$\frac{df}{dx_1}$ ↳ $\lim_{h\to 0} \frac{f(x_1+h, x_2, x_3, \cdots) - f(x_1, x_2, \cdots)}{h} = \lim \begin{bmatrix} \frac{(x_1+h)^2 - x_1^2}{h} & x_2 \cdots & x_n \\ \vdots & & \\ x_n & & 0 \end{bmatrix}$

Similarly,

$\frac{df}{dx_2} = \lim_{h\to 0} \frac{f(x_1, x_2+h, x_3, \cdots) - f(x_1, x_2, x_3 \cdots)}{h} = \lim \begin{bmatrix} 0 & x_1 & \cdots & 0 \\ x_1 & \frac{(x_2+h)^2 - x_2^2}{h} & x_3, \cdots & x_n \\ \vdots & \vdots & & \\ & x_n & 0 & \end{bmatrix}$

Thus, each layer x_i in $(x_1, x_2 \cdots x_n)$ has $\frac{df}{dx_i} = \begin{bmatrix} \cdots & x_1 & \cdots \\ & \vdots & \\ x_1 \cdots & 2x_i & \cdots x_n \\ & \vdots & \\ & x_n & \end{bmatrix}$, other cells $= 0$.

$\frac{df}{dX} =$ 

e) $f(x) = \sin(\log(x^T x))$ $\quad f: \mathbb{R}^n \to \mathbb{R}, \quad x^T x = 2x \in \mathbb{R}$

$$\frac{df}{du} = (\sin u)' = \cos u$$

$$\frac{du}{dx} = [\log(x^T x)]' = \frac{1}{x^T x} \cdot (x^T x)'$$

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

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

f) $f(z) = \log(1+z)$ $\quad z = x^T x = x^2$

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

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

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

$x^T A x \in \mathbb{R}$

$$f(x) = [x_1 \cdots x_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$$

$$\frac{df}{dx} = \sum_{k=1}^{n} \Big(\sum_{j=1}^{n} \underset{\substack{\downarrow \\ \text{row}}}{a_{kj}} x_j + \sum_{i=1}^{n} \underset{\substack{\downarrow \\ \text{col}}}{a_{ik}} x_i \Big) = x^T A^T + x \cdot A = 2Ax$$

4 Optimization [15 points]

1. (7.5 pts) Consider the univariate function $f(x) = x^3 + 6x^2 - 3x - 5$. Find its stationary points and indicate whether they are maximum, minimum or saddle points.

2. (7.5 pts) Explain how to solve the least squares loss in a linear model using (i) gradient descent and (ii) SVD. Discuss the pros and cons.

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

$3(x^2 + 4x - 1) = 0$

$x^2 + 4x + 4 = 5$

$(x+2)^2 = 5$

$x = \pm\sqrt{5} - 2$

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

$f''(x_1) = 6\sqrt{5} - 12 + 12 = 6\sqrt{5} > 0$

$f''(x_2) = -6\sqrt{5} - 12 + 12 = -6\sqrt{5} < 0$

$x_1 = \sqrt{5} - 2 \quad x_2 = -\sqrt{5} - 2$

$f(x) \xrightarrow{x \to \infty} \infty \quad f(x) \xrightarrow{x \to -\infty} -\infty$

$\therefore x_1, x_2$ are not global.

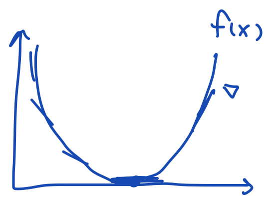
Since $f''(x_1) > 0$, it's a concave-up graph at x_1. $\Rightarrow$ x_1 is local minimum

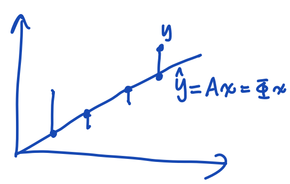
Similarly, $f''(x_2) < 0$, so x_2 is local maximum.

2)

i) gradient.

$\hat{x} = x - \eta(\nabla f(x))^T, \eta > 0$





We'll choose a proper η, η is the gap between each selection of x. We'll continue calculating the gradien of x we selected until we get a $\nabla_x f(\vec{x}) = 0$, which is a minimum.

ii) SVD. is basically doing calculation.

$\min \|Ax - b\|^2$, $A = U\Sigma V^T$

$\Rightarrow \quad A^TA\hat{x} = A^Tb$

$V\Sigma^{-1}U^TA\hat{x} = V\Sigma^{-1}U^Tb$

$\hat{x} = V\Sigma^{-1}U^Tb$ is the optimal solution.

Pro & Con
Next Page

	Pro	con
Gradient descent.	• If the sample is huge, it's faster than SVD, since SVD needs to compute an inverse matrix which takes time.	• It might get local minimum of x rather than global minimum. • We need to choose proper η, if η is too high, it'll cause drastic changes and we might miss the minimum. If η is too low, it'll take too many steps to reach a minimum.
SVD	• If the sample is not too huge, SVD is faster than Gradient descent. • It's more accurate than using Gradient descent.	• When sample is too large, it'll take a long time computing the inverse matrix.