SI231 - Matrix Computations, 2022 Fall

Homework Set #5

Prof. Yue Qiu

Acknowledgements:

1) Deadline: **2022-12-17 23:59:59**
2) **Late Policy details** can be found on piazza.
3) Submit your homework in **Homework 5** on **Gradescope**. Entry Code: **4V2N55**. **Make sure that you have correctly select pages for each problem.** If not, you probably will get 0 point.
4) No handwritten homework is accepted. You need to write LaTeX. (If you have difficulties in using LaTeX, you are allowed to use **MS Word or Pages** for the first and the second homework to accommodate yourself.)
5) Use the given template and give your solution in English. Solution in Chinese is not allowed.
6) Your homework should be uploaded in the PDF format, and the naming format of the file is not specified.

I. Power Iteration

Problem 1. (20 points) Work this problem by hand. Let $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Carry out the power method with starting vector $q_0 = \begin{bmatrix} a & b \end{bmatrix}^T$, where$\|q_0\|_2 = 1$, $a \geq 0, b \geq 0$, and $a \neq b$. Explain why the sequence fails to converge.

Solution:

We wish to compute the iterates in the power method by

$$q_{j+1} = \frac{Aq_j}{\|q_j\|_2},$$

Hence we compute

$$Aq_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} b \\ a \end{bmatrix},$$

Then

$$q_1 = \frac{Aq_0}{\|q_0\|_2} = \begin{bmatrix} b \\ a \end{bmatrix}.$$

For the next iteration we compute

$$Aq_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$$

Thus

$$q_2 = \frac{Aq_1}{\|q_1\|_2} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Similarly

$$Aq_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} b \\ a \end{bmatrix}$$

so we will continue to cycle between

$$q_{2n-1} = \begin{bmatrix} b \\ a \end{bmatrix} \quad \text{and} \quad q_{2n} = \begin{bmatrix} a \\ b \end{bmatrix}$$

for all $n \geq 1$.(12 points)

Hence the power method fails to converge to a single vector. Computing the eigenvalues of the matrix A at hand we get $\sigma(A) = \{-1, 1\}$, which demonstrates to us why the convergence argument for the power method fails: there is no dominating eigenvalue since the absolute value of both of them is 1 , so the error between q_n and some hypothetical eigenvector $c_1 v_1$ never decreases in modulus. (8 points)

II. Frobenius Norm

Problem 2. (8 points + 7 points + 5 points) The Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$, where a_{ij} is the entry in the i-th row and j-th column.

1) Prove that $\|A\|_F = \sqrt{\operatorname{tr}(A^T A)}$ with the denifition of Frobenius norm.
2) If U and V are square orthonormal matrices, prove that $\|UA\|_F = \|AV\|_F = \|A\|_F$.
3) Use the SVD of A to show that $\|A\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$, where $r = \min\{m,n\}$ and $\sigma_1, \cdots, \sigma_r$ are the singular values of A.

Solution:

1) $\operatorname{tr}\left(A^\top A\right) = \sum_{i=1}^n \left(A^\top A\right)_{ii} = \sum_{i=1}^n \left(\sum_{j=1}^m \left(A^\top\right)_{ij} A_{ji}\right) = \sum_{i=1}^n \left(\sum_{j=1}^m A_{ji}A_{ji}\right) = \sum_{i=1}^n \sum_{j=1}^m \left(A_{ji}^2\right) = \sum_{i=1}^n \sum_{j=1}^m \left(A_{ji}^2\right) = \|A\|_F^2$. (8 points)
2) $\|UA\|_F = \sqrt{\operatorname{tr}\left((UA)^\top (UA)\right)} = \sqrt{\operatorname{tr}\left(A^\top U^\top UA\right)} = \sqrt{\operatorname{tr}\left(A^\top A\right)} = \|A\|_F$,
   $\|AV\|_F = \left\|(AV)^\top\right\|_F = \left\|V^T A^T\right\|_F$. Since V is orthonormal, V^T is also orthonormal.
   $\left\|V^\top A^\top\right\|_F = \left\|A^\top\right\|_F = \|A\|_F$. Hence, $\|UA\|_F = \|AV\|_F = \|A\|_F$. (7 points)
3) $\|A\|_F = \left\|U\Sigma V^\top\right\|_F = \left\|\Sigma V^\top\right\|_F = \|\Sigma\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$ (5 points)

III. Singular Value Decomposition

Problem 3. (17 points + 8 points) Let $A \in \mathbb{R}^{m\times n}$, $m \geqslant n$, $\operatorname{rank}(A) = n$, with singular values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n > 0$. $A = U\Sigma V^T$, where U and V are orthogonal.

1) Find the SVD of $\left(A^TA\right)^{-1}$, $\left(A^TA\right)^{-1}A^T$ in term of the SVD of A. The singular values of $\left(A^TA\right)^{-1}$ and $\left(A^TA\right)^{-1}A^T$ should be in descending order.

   Hint: The permutation matrix $P \in \mathbb{R}^{n\times n}$ with 1 on the anti-diagonal and 0 elsewhere may be useful. And also $P^T = P$, $P^TP = P^2 = I$.
2) Prove that $\left\|\left(A^TA\right)^{-1}\right\|_2 = \sigma_n^{-2}$, $\left\|\left(A^TA\right)^{-1}A^T\right\|_2 = \sigma_n^{-1}$.

Solution:

1) $A = U\Sigma V^T$, then $A^T = V\Sigma U^T$ is SVD of A^T and $A^TA = V\Sigma^TU^TU\Sigma V^T = V\left(\Sigma^T\Sigma\right)V^T$ is SVD of A^TA, where $\Sigma^T\Sigma = \operatorname{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right) \in \mathbb{R}^{n\times n}$.

   Then we have $\left(A^TA\right)^{-1} = \left(V\Sigma^T\Sigma V^T\right) = \left(V^T\right)^{-1}\left(\Sigma^T\Sigma\right)^{-1}V^{-1} = V\left(\Sigma^T\Sigma\right)^{-1}V^T$,

   where $\left(\Sigma^T\Sigma\right)^{-1} = \operatorname{diag}\left(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_m^{-2}\right) \in \mathbb{R}^{n\times n}$.

   Since the singular values of $\left(A^TA\right)^{-1}$ are currently in ascending order, as opposed to descending. Let permutation matrix P be the $n\times n$ with 1 on the anti-diagonal and 0 elsewhere, then $P^T = P$, $P^TP = P^2 = I$.

   Therefore, $\left(A^TA\right)^{-1} = VPP\left(\Sigma^T\Sigma\right)^{-1}PPV^T = (VP)\left(P\left(\Sigma^T\Sigma\right)^{-1}P\right)(VP)^T$ is the SVD of $\left(A^TA\right)^{-1}$, where $P\left(\Sigma^T\Sigma\right)^{-1}P = \operatorname{diag}\left(\sigma_n^{-2}, \sigma_{n-1}^{-2}, \ldots, \sigma_1^{-2}\right)$ and VP is orthogonal since both V and P are orthogonal.

   (8 points)

   Analogously, $\left(A^TA\right)^{-1}A^T = \left(V\left(\Sigma^T\Sigma\right)^{-1}V^T\right)\left(V\Sigma^TU^T\right) = V\left(\Sigma^T\Sigma\right)^{-1}\Sigma^TU^T$,

   where $\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T = \operatorname{diag}\left(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_n^{-1}\right) \in \mathbb{R}^{n\times m}$.

   We adjust the order of the diagonal elements by the permutation matrix P above,

   then $\left(A^TA\right)^{-1}A^T = (VP)\left(P\left(\Sigma^T\Sigma\right)^{-1}\Sigma^TP\right)(UP)^T$ is the SVD of $\left(A^TA\right)^{-1}A^T$. (9 points)
2) Since the singular values of $\left(A^TA\right)^{-1}$ are $\sigma_n^{-2} \geqslant \sigma_{m-1}^{-2} \geqslant \ldots \geqslant \sigma_1^{-2}$, $\left\|\left(A^TA\right)^{-1}\right\|_2 = \sigma_n^{-2}$.

   Similarly, the singular values of $\left(A^TA\right)^{-1}A^T$ are $\sigma_n^{-1} \geqslant \sigma_{m-1}^{-1} \geqslant \ldots \geqslant \sigma_1^{-1}$, $\left\|\left(A^TA\right)^{-1}A^T\right\|_2 = \sigma_n^{-1}$

   (8 points)

IV. REGULARIZED LEAST SQUARE

Problem 4. (10 points + 5 points + 10 points + 10 points). Let matrix $A \in \mathbb{R}^{m\times n}$ have the singular value decomposition $A = U\Sigma V^\top$, and let its smallest singular value be $\sigma_{\min}(A) > 0$.

1) Consider the linear equation $x(\tilde{y}) = A^\dagger \tilde{y}$, where $\tilde{y} \in \mathbb{R}^n$ is a noisy measurement satisfying

$$\|\tilde{y} - y\|_2 \leq r$$

for some vector $y \in \mathbb{R}^n$ and $r > 0$. Let $x(y) = A^\dagger y$. Show that

$$\max_{\tilde{y}:\|\tilde{y}-y\|_2 \leq r} \|x(\tilde{y}) - x(y)\|_2 = \frac{r}{\sigma_{\min}(A)}$$

Hint: $A^\dagger$ is A's pseudo-inverse.

2) What happens if the smallest singular value of A is very close to zero?

3) we consider regularized least squares problem

$$x(y) = \arg\min_x ||Ax - y||_2^2 + \lambda||x||_2^2$$

then the optimal solution is

$$x_\lambda(y) = (A^T A + \lambda I)^{-1} A^T y,$$

Show that for all $\lambda > 0$,

$$\max_{\tilde{y}:\|\tilde{y}-y\|_2 \leq r} \|x_\lambda(\tilde{y}) - x_\lambda(y)\|_2 \leq \frac{r}{2\sqrt{\lambda}}$$

and explain that how does the value of λ affect the sensitivity of your solution $x_\lambda(y)$ to noise in y ?

Hint: For every $\lambda > 0$, we have

$$\max_{\sigma>0} \frac{\sigma}{\sigma^2 + \lambda} = \frac{1}{2\sqrt{\lambda}}.$$

(You need not show this; this optimization can be solved by setting the derivative of the objective function to 0 and solving for σ .)

4) let

$$A = \begin{bmatrix} 2 & 3 & 3 \\ -1 & 3 & 4 \\ -2 & 3 & 2 \\ -2 & 4 & 2 \end{bmatrix}, y = [1, 0, -1, 2]^T$$

since A is invertiable, we can have true solution $x^* = A^\dagger y$. We assumed that $\tilde{y} = y + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1 * I)$. Calculate the corresponding error, for different values of λ:

$$e_1 = \|x^* - x_\lambda(y)\|_2^2$$

$$e_2 = \|x_\lambda(y) - x_\lambda(\tilde{y})\|_2^2$$

Plot e_1 vs $\log \lambda$ and e_2 vs $\log \lambda$. Figures must show that how λ affect e_1 and e_2.

Hint: Submiting two images directly and the code to the PDF file. You must ensure that the code is clear. Otherwise, points will be deducted due to unclear code and will not be eligible for regrade.

Solution:

1) For any $y \in \mathbb{R}^n$, we have

$$x(y) = A^\dagger y = V\Sigma^{-1}U^\top y.$$

Note that

$$\{\tilde{y} \in \mathbb{R}^n : \|\tilde{y} - y\|_2 \leq r\} = \{y + u : u \in \mathbb{R}^n, \|u\|_2 \leq r\},$$

and therefore,

$$\max_{\tilde{y}:\|\tilde{y}-y\|_2\leq r} \|x(\tilde{y}) - x(y)\|_2 = \max_{\tilde{y}:\|\tilde{y}-y\|_2\leq r} \|x(y+u) - x(y)\|_2$$

Since we can write the differences between the estimates as

$$x(y+u) - x(y) = A^\dagger(y+u) - A^\dagger y = A^\dagger u = V\Sigma^{-1}U^\top u,$$

we obtain

$$\begin{aligned}\max_{u:\|u\|_2\leq r} \|x(y+u) - x(y)\|_2 &= \max_{u:\|u\|_2\leq r} \left\|V\Sigma^{-1}U^\top u\right\|_2 \\ &= \max_{u:\|u\|_2\leq r} \left\|\Sigma^{-1}u\right\|_2\end{aligned}$$

where the last equality follows from the fact that U and V are orthonormal matrices. The matrix Σ^{-1} is a diagonal matrix of entries are the inverse of those in Σ , and thus

$$\max_{u:\|u\|_2\leq r} \left\|\Sigma^{-1}u\right\|_2 = r\sigma_{\max}\left(\Sigma^{-1}\right) = \frac{r}{\sigma_{\min}(\Sigma)} = \frac{r}{\sigma_{\min}(A)}$$

2) In part 1, we showed that a perturbation of magnitude r on the measurement can change our estimate x by up to $\frac{r}{\sigma_{\min}(A)}$. If $\sigma_{\min}(A)$ is very small, the estimate can change by a large amount even if the measurements are only slightly perturbed.

3) By the above optimal solution $x_\lambda(y)$, we have that

$$x_\lambda(\tilde{y}) - x_\lambda(y) = \left(A^\top A + \lambda I\right)^{-1} A^\top(\tilde{y} - y)$$

and thus

$$\begin{aligned}\|x_\lambda(\tilde{y}) - x_\lambda(y)\|_2 &= \left\|\left(A^\top A + \lambda I\right)^{-1} A^\top(\tilde{y} - y)\right\|_2 \\ &\leq \left\|\left(A^\top A + \lambda I\right)^{-1} A^\top\right\|_2 \|\tilde{y} - y\|_2 \\ &= \sigma_{\max}\left(\left(A^\top A + \lambda I\right)^{-1} A^\top\right) \|\tilde{y} - y\|_2.\end{aligned}$$

This value is maximized when $\|\tilde{y}-y\|_2 = r$, so we can write new bound

$$\max_{\tilde{y}:\|\tilde{y}-y\|_2\le r} \|x_\lambda(\tilde{y}) - x_\lambda(y)\|_2 \le \sigma_{\max}\left(\left(A^\top A+\lambda I\right)^{-1} A^\top\right) r.$$

To simplify this bound further, we need to compute the largest singular value of $(A^\top A+\lambda I)^{-1}A^\top$. Plugging in our decomposition $A = U\Sigma V^\top$, we first compute

$$A^\top A+\lambda I = V\Sigma U^\top U\Sigma V^\top + \lambda I = V\Sigma^2 V^\top + \lambda I = V\Sigma^2 V^\top + \lambda V V^\top = V\left(\Sigma^2+\lambda I\right)V^\top,$$

which leads to

$$\left(A^\top A+\lambda I\right)^{-1} A^\top = V\left(\Sigma^2+\lambda I\right)^{-1} V^\top V\Sigma U^\top = V\left(\Sigma^2+\lambda I\right)^{-1}\Sigma U^\top.$$

This is the SVD decomposition of $\left(A^\top A+\lambda I\right)^{-1} A^\top$, and its singular values are the diagonal elements of $\left(\Sigma^2+\lambda I\right)^{-1}\Sigma$. Given the singular values $\{\sigma_i\}_{i=1}^n$ of A, the singular values of $\left(\Sigma^2+\lambda I\right)^{-1}\Sigma$ are

$$\left\{\frac{\sigma_i}{\sigma_i^2+\lambda}\right\}_{i=1}^n,$$

all of which we know are smaller than $\frac{1}{2\sqrt{\lambda}}$ from the given hint. We can then rewrite our bound above as

$$\max_{\tilde{y}:\|\overrightarrow{y}-y\|_2\le r} \|x_\lambda(\tilde{y}) - x_\lambda(y)\|_2 \le \sigma_{\max}\left(\left(A^\top A+\lambda I\right)^{-1} A^\top\right) r \le \frac{r}{2\sqrt{\lambda}}$$

as desired.

The larger we choose our λ, the tighter our bound on the deviation of our noisy solution $x_\lambda(\tilde{y})$ from our true solution $x_\lambda(y)$. In other words, if the regularization parameter λ is large enough, small perturbations in the measurement cannot change the estimate by a large amount.

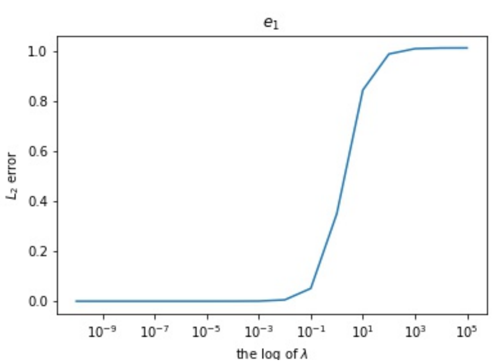

Figure 1: error1

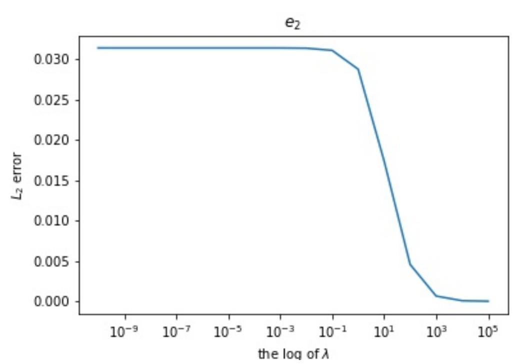

Figure 2: error2

---

4)

```
import numpy as np
import matplotlib.pyplot as plt
A=  np.array([[2,3,3],
```

```
            [-1,3,4],
            [-2,3,2],
            [-2,4,2]])

y = np.array([1,0,-1,2]).reshape(-1,1)

y_tilde = y+  np.random.randn(4,1)*0.1

x_true  =np.linalg.pinv(A).dot(y)



#lambda
l=1e-10
l_list = []
e1_res = []
e2_res = []
while l< 1e6:
    x_y = np.linalg.inv((A.T.dot(A)+l*np.eye(3))).dot(A.T).dot(y)
    x_y_tilde = np.linalg.inv((A.T.dot(A)+l*np.eye(3))).dot(A.T).dot(y_tilde)
    e2_res.append(np.linalg.norm(x_y-x_y_tilde))
    e1_res.append(np.linalg.norm(x_true-x_y))
    l_list.append(l)
    l *=10


plt.plot(l_list,e1_res)
plt.xscale('log')
plt.xlabel(r'the log of $\lambda$')
plt.ylabel(r'$L_2$ error')
plt.title(r'$e_1$')
plt.savefig("figure1.jpg")
plt.clf()
plt.plot(l_list,e2_res)
plt.xscale('log')
plt.xlabel(r'the log of $\lambda$')
```

```
plt.ylabel(r'$L_2$ error')
plt.title(r'$e_2$')
plt.savefig("figure2.jpg")
```