1. Consider the finite difference discretization of the equation

$$\begin{cases} -(a(x)u'(x))' = f(x), & \text{for } x \in (0,1) \\ u(0) = u(1) = 0. \end{cases}$$

The discretized linear system is $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$, where

$$\boldsymbol{A} = \begin{bmatrix} a_0+a_1 & -a_1 & & & & \\ -a_1 & a_1+a_2 & -a_2 & & & \\ & -a_2 & a_2+a_3 & -a_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -a_{n-2} & a_{n-2}+a_{n-1} & -a_{n-1} \\ & & & & -a_{n-1} & a_{n-1}+a_n \end{bmatrix}$$

and $a_0, a_1, \ldots, a_n$ are $a(x)$ on the grid points. Assume $C_1 \le a(x) \le C_2$ for all $x \in [0,1]$, where C_1 and C_2 are positive constants.

(a) Prove that $\boldsymbol{A}$ is symmetric positive definite.

(b) Show that both Jacobi and Gauss-Seidel converges for solving $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$.

(c) Prove that

$$4C_1 \sin^2\left(\frac{\pi}{2(n+1)}\right) \le \lambda_1 \le \lambda_n \le 4C_2,$$

where λ_1 and λ_n are the smallest and largest eigenvalues of $\boldsymbol{A}$ respectively.

(d) Since $\boldsymbol{A}$ is SPD, we may use the preconditioned conjugate gradient (PCG) to solve $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$. A candidate preconditioner will be

$$\boldsymbol{P} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

Estimate the number of iterations needed for a solution with ϵ precision. Your answer should be as tight as possible. *(This preconditioner will be practically useful for 2D case, because $\boldsymbol{P}$ is diagonalizable by discrete sine transform and inverted very efficiently by fast Fourier transform.)*

(a) It is obvious that A is symmetric.

Now we prove A is positive definite.

$\forall x \in \mathbb{R}^n$, $\quad x^T A x = \sum_{i=1}^{n} (a_{i-1}+a_i) x_i^2 - 2\sum_{i=1}^{n-1} a_i x_i x_{i+1}$

$= \sum_{i=0}^{n-1} a_i x_{i+1}^2 + \sum_{i=1}^{n} a_i x_i^2 - 2\sum_{i=1}^{n-1} a_i x_i x_{i+1}$

$= \sum_{i=1}^{n-1} a_i (x_{i+1}-x_i)^2 + a_0 x_1^2 + a_n x_n^2$

≥ 0

$x^T A x = 0 \Leftrightarrow \sum_{i=1}^{n-1} a_i (x_{i+1}-x_i)^2 + a_0 x_1^2 + a_n x_n^2$

$\Leftrightarrow x_i = 0 \quad i = 1, 2, \cdots, n$

So A is SPD.

(b) ① For Jacobi:

Let $D = \mathrm{diag}(\mathrm{diag}(A))$.

$$2D - A = \begin{pmatrix} a_0+a_1 & a_1 & & \\ a_1 & a_1+a_2 & \ddots & \\ & \ddots & \ddots & a_{n-1} \\ & & a_{n-1} & a_{n-1}+a_n \end{pmatrix}$$

$x^T(2D - A)x = \sum_{i=1}^{n-1} a_i (x_i + x_{i+1})^2 + a_0 x_1^2 + a_n x_n^2 \ge 0 \quad , \forall x \in \mathbb{R}^n$

$x^T(2D-A)x = 0 \Leftrightarrow x_i = 0 \quad , i = 1, 2, \cdots, n$

So $2D-A$ is SPD, indicating that Jacobi method converges.

② For Gauss-Seidel:

$A = D + L + U$

$x^{(k+1)} = (D+L)^{-1}b - (D+L)^{-1}Ux^{(k)}$

Now we show that $\rho((D+L)^{-1}U) < 1$.

By contradiction, if there exists an eigenvalue of $(D+L)^{-1}U$ such that $\lambda \geq 1$, and the corresponding eigenvector is x, then

$$(D+L)^{-1}Ux = \lambda x \ , \ \text{i.e.} \ [U - \lambda(D+L)]x = 0.$$

1) if $\lambda > 1$, $U - \lambda(D+L)$ is strictly diagonally dominant, indicating $U - \lambda(D+L)$ is invertible and x must be 0.

It contradicts with x is an eigenvector.

2) if $\lambda = 1$, $U - (D+L)$ is diagonally dominant and irreducible as

$a_i \neq 0$, $i = 0, 1, \cdots, n$. So $U - (D+L)$ is invertible, and x must be 0.

It contradicts with x is an eigenvector.

In conclusion, $\rho((D+L)^{-1}U) < 1$. So Gauss-Seidel method converges.

(c) $\lambda_n \leq \|A\|_1 \leq \max\limits_{1 \leq i \leq n} 2a_i + 2a_{i-1} \leq 4C_2$

$$\begin{aligned}
\lambda_1 &= \min_{\|x\|=1} x^T A x \\
&= \min_{\|x\|=1} \sum_{i=1}^{n-1} a_i (x_{i+1} - x_i)^2 + a_0 x_1^2 + a_n x_n^2 \\
&\geq C_1 \min_{\|x\|=1} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 + x_1^2 + x_n^2 \\
&= C_1 \min_{\|x\|=1} x^T P x \\
&= C_1 \lambda_1(P) = 4C_1 \sin^2\left(\frac{\pi}{2(n+1)}\right)
\end{aligned}$$

Hence, $4C_1 \sin^2\left(\frac{\pi}{2(n+1)}\right) \leq \lambda_1 \leq \lambda_n \leq 4C_2$

(d) Since it converges at $k = n$.

$k \geq n$, $\|x_k - x_*\|_A = 0$

$1 \leq k < n$, $\|x_k - x_*\|_A \leq 2\left(\frac{\sqrt{r}-1}{\sqrt{r}+1}\right)^k \|x_0 - x_*\|_A$, here $r = \dfrac{\lambda_{max}(P^{-1}A)}{\lambda_{min}(P^{-1}A)}$

$\|x_k - x_*\|_A \leq \varepsilon$, we can choose k such that

$$k \geq \frac{\log\left(\frac{\varepsilon}{2\|x_0 - x_*\|_A}\right)}{\log\left(\frac{\sqrt{r}-1}{\sqrt{r}+1}\right)}$$

2. The singular value decomposition (SVD) is a fundamental decomposition with numerous applications. In this question, we derive the SVD by the eigenvalue decomposition, and develop an algorithm for it. Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$. Since $\boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is square and symmetric positive semi-definite (SPSD), there exists an eigenvalue decomposition

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ are eigenvalues, and $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_n] \in \mathbb{R}^{n\times n}$ are corresponding eigenvectors.

(a) Prove that $\boldsymbol{A}\boldsymbol{A}^T \in \mathbb{R}^{m\times m}$ has at most n nonzero eigenvalues, which are also $\lambda_1, \dots, \lambda_n$.

(b) Therefore, $\boldsymbol{A}\boldsymbol{A}^T$ has an eigenvalue decomposition

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T,$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n] \in \mathbb{R}^{m\times n}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ corresponding to $\lambda_1, \dots, \lambda_n$ respectively. Assume all eigenvalues $\lambda_1, \dots, \lambda_n$ are all simple (though this assumption can be removed). Prove that there exists $\sigma_i \geq 0$, $i = 1, \dots, n$, such that

$$\begin{cases} \boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i, \\ \boldsymbol{A}^T\boldsymbol{u}_i = \sigma_i \boldsymbol{v}_i, \\ \sigma_i^2 = \lambda_i, \end{cases} \qquad i = 1, \dots, n.$$

(c) Define $\sigma_i = \sqrt{\lambda_i}$, $i = 1, \dots, n$. Prove that $\boldsymbol{A}$ has a decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$. This decomposition is SVD, and $(\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i)$ are called singular value, left and right singular vectors of $\boldsymbol{A}$ respectively.

(d) Similar to eigenvalues of symmetric matrices, singular values also have many nice variational properties. Prove

$$\sigma_1 = \max_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{v}\|_2=1} \boldsymbol{u}^T\boldsymbol{A}\boldsymbol{v},$$

where σ_1 is the largest singular value of $\boldsymbol{A}$. (There are other identities similar to the min-max theorem of eigenvalues.)

(e) Use (or not use) (d) to prove

$$\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T = \arg\min_{\mathrm{rank}(\boldsymbol{B})=1} \|\boldsymbol{A} - \boldsymbol{B}\|_F^2.$$

(That is, SVD gives the best rank-1 approximation. This can be extended to any best rank-r approximation, and this makes SVD a fundamental tool in many applications.)

(f) Propose a power iteration to compute the leading left and right singular vectors of $\boldsymbol{A}$. Your algorithm should use fewest possible matrix-vector products in each iteration. (All eigenvalue algorithms can be extended to SVD.)

(a) ① Since $\mathrm{rank}(AA^T) \le \mathrm{rank}(A) \le n$, AA^T has at most n nonzero eigenvalues.

② For any $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times m}$, we know that AB and BA have the same nonzero eigenvalues, we know the nonzero eigenvalues are also $\lambda_1, \lambda_2, \cdots, \lambda_n$.

(b) $A^TA = V\Lambda V^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T$.

Since $\lambda_1, \cdots, \lambda_n$ are simple, $V = (v_1, \cdots, v_n)$ is orthogonal.

Let $u_i = \dfrac{Av_i}{\sqrt{\lambda_i}} = \dfrac{Av_i}{\sigma_i}$, $i=1,\cdots,n$ (here $\lambda_i = \sigma_i^2$)

$A^Tu_i = \frac{1}{\sigma_i} A^TAv_i = \frac{1}{\sigma_i}\lambda_i v_i = \sigma_i v_i$, $i=1,\cdots,n$

Since $AA^Tu_i = \sigma_i Av_i = \sigma_i^2 u_i = \lambda_i u_i$, $i=1,\cdots,n$, u_i is the eigenvectors of AA^T.

In conclusion, we have

$$\begin{cases} Av_i = \sigma_i u_i \\ A^Tu_i = \sigma_i v_i \\ \sigma_i^2 = \lambda_i \end{cases} \quad i=1,\cdots,n$$

(c) From (b), we know that

$$AU = A(v_1, v_2, \cdots, v_n) = (Av_1, \cdots, Av_n)$$
$$= (u_1, u_2, \cdots, u_n)\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$
$$= U\Sigma .$$

i.e. $A = U\Sigma V^T$

(d) $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$

For $u \in \mathbb{R}^n$, $v \in \mathbb{R}^m$ such that $\|u\| = 1$, $\|v\| = 1$,

$$u = \sum_{j=1}^{m} c_j u_j , \quad v = \sum_{k=1}^{n} d_k v_k$$

$$u^T A v = \sum_{i=1}^{n} \sigma_i \Big(\sum_{j=1}^{m} c_j u_j^T \Big) u_i v_i^T \Big(\sum_{k=1}^{n} d_k v_k \Big)$$
$$= \sum_{i=1}^{n} \sigma_i c_i d_i$$
$$\leq \sigma_1 \sum_{i=1}^{n} c_i d_i$$
$$\leq \sigma_1 \Big(\sum_{i=1}^{n} c_i^2 \Big)^{\frac{1}{2}} \Big(\sum_{i=1}^{n} d_i^2 \Big)^{\frac{1}{2}}$$
$$\leq \sigma_1$$

The equation can be reached since $\sigma_1 = u_1^T A v_1$.

Thus we have $\sigma_1 = \max_{\|u\|_2=1, \|v\|_2=1} u^T A v$

(e) For any rank-1 matrix B, we have $B = c u v^T$, where $c \in \mathbb{R}$, $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, $\|u\|_2 = \|v\|_2 = 1$,

$$\|A - B\|_F^2 = tr[(A-B)^T(A-B)]$$
$$= tr(A^T A) - 2tr(B^T A) + tr(B^T B)$$
$$= c^2 - 2c\,tr(v u^T A) + tr(A^T A)$$
$$= c^2 - 2c\; tr(u^T A v) + tr(A^T A)$$

Since A is fixed,

$$\min_{rank(B)=1} \|A - B\|_F^2 \Leftrightarrow \min_{\substack{c \in \mathbb{R}, u \in \mathbb{R}^m, v \in \mathbb{R}^n \\ \|u\|_2 = \|v\|_2 = 1}} c^2 - 2c\,(u^T A v)$$

By (d), $(u_1, v_1) = \arg\max_{\|u\|_2=1, \|v\|_2=1} u^T A v$

then $\sigma_1 = \arg\min_{c} c^2 - 2\sigma_1 c$.

Hence $\sigma_1 u_1 v_1^T = \underset{rank(B)=1}{\arg\min} \|A - B\|_F^2$

(f) $M = AA^T$, $N = A^T A$.

For $k = 1, 2, \cdots$

$$u^{(k)} = M x^{(k-1)}$$
$$v^{(k)} = N y^{(k-1)}$$
$$x^{(k)} = \frac{u^{(k)}}{\|u^{(k)}\|_2}$$
$$y^{(k)} = \frac{v^{(k)}}{\|v^{(k)}\|_2}$$

end

3. We have used QR decompostion for solving the least squares (LS) problems in GMRES method

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2$$

where $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ is rank r with $m\geq n$. Actually, LS problems arises in many other applications in applied mathematics and engineering, and there are other solvers for them. We consider to use the singular value decomposition (SVD) to solve LS problems.

(a) If we take only the non-zero singular values, then we obtain the compact SVD of $\boldsymbol{A}$

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T,$$

where $\boldsymbol{U}\in\mathbb{R}^{m\times r}$ satisfies $\boldsymbol{U}^T\boldsymbol{U}=\boldsymbol{I}$, $\boldsymbol{V}\in\mathbb{R}^{n\times r}$ satisfies $\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}$, and $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_r)$ with $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_r>0$. Prove that

$$\mathrm{Ran}(\boldsymbol{A})=\mathrm{Ran}(\boldsymbol{U}),\qquad \mathrm{Ker}(\boldsymbol{A})=\mathrm{Ran}(\boldsymbol{V})^\perp,$$

where $(\cdot)^\perp$ stands for the orthogonal complementary.

(b) Assume $r=n$. Prove that the solution of LS is unique and is given by

$$\boldsymbol{x}=\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{b}.$$

(c) Continuing (b): Let $\tilde{\boldsymbol{x}}$ be the solution of the LS when the input $\boldsymbol{b}$ is perturbed to $\tilde{\boldsymbol{b}}$. Prove that

$$\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|_2}{\|\boldsymbol{x}\|_2}\leq C\frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|_2}{\|\boldsymbol{b}\|_2},$$

where $C=\frac{\sigma_1}{\sigma_n}$ if the LS is solved by formula in (b) and $C=\frac{\sigma_1^2}{\sigma_n^2}$ if the LS solution is obtained by solving the normal equation.

(d) Assume $r<n$. Prove that all solutions of LS are given by

$$\boldsymbol{x}=\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{b}+\boldsymbol{y},\quad \boldsymbol{y}\in\mathrm{Ker}(\boldsymbol{A}),$$

and $\boldsymbol{x}_0:=\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{b}$ is the solution of LS with the minimum 2-norm among all solutions.

(a) $\begin{cases} A=U\Sigma V^T & \Rightarrow \mathrm{Ran}(A)\subseteq \mathrm{Ran}(U) \\ U=AV\Sigma^{-1} & \Rightarrow \mathrm{Ran}(U)\subseteq \mathrm{Ran}(A)\end{cases} \Rightarrow \mathrm{Ran}(A)=\mathrm{Ran}(U)$

$\begin{cases} A=U\Sigma V^T & \Rightarrow \mathrm{Ran}(V)^\perp=\ker(V^T)\subseteq \ker(A) \\ V^T=\Sigma^{-1}U^TA & \Rightarrow \ker(A)\subseteq\ker(V^T)=\mathrm{Ran}(V)^\perp\end{cases} \Rightarrow \ker(A)=\mathrm{Ran}(V)^\perp$

(b) $\|Ax-b\|_2^2 = \|U\Sigma V^Tx-b\|_2^2$

$= \|\Sigma V^Tx-U^Tb\|_2^2$

$\overset{y=V^Tx}{=\!=} \|\Sigma y-U^Tb\|_2^2$

$\min\limits_x \|Ax-b\|_2^2 \Leftrightarrow \min\limits_y \|\Sigma y-U^Tb\|_2^2$

Let $\Sigma y-U^Tb=0$, we have $y=\Sigma^{-1}U^Tb$.

Then $x=Vy=V\Sigma^{-1}U^Tb$.

(c) ① Let $\tilde{x}=V\Sigma^{-1}U^T\tilde{b}$

$\|x-\tilde{x}\|_2 = \|V\Sigma^{-1}U^T(b-\tilde{b})\|_2$

$\leq \|V\Sigma^{-1}U^T\|_2\,\|b-\tilde{b}\|_2$

$= \|\Sigma^{-1}\|_2\,\|b-\tilde{b}\|_2$

$= \frac{1}{\sigma_n}\|b-\tilde{b}\|_2$

$x=V\Sigma^{-1}U^Tb \Rightarrow b=U\Sigma V^Tx$

$$\|b\|_2 = \| U\Sigma V^T x\|_2$$
$$\le \|U\Sigma V^T\|_2 \|x\|_2$$
$$= \|\Sigma\|_2 \|x\|_2$$
$$= \sigma_1 \|x\|_2$$

Thus $\dfrac{\|x-\tilde{x}\|_2}{\|x\|_2} \le \dfrac{\sigma_1}{\sigma_n} \dfrac{\|b-\hat{b}\|_2}{\|b\|_2}$

② For solving $A^TAx = A^Tb$, $x = (A^TA)^{-1}A^Tb$

$$\|x-\tilde{x}\|_2 \le \|(A^TA)^{-1}A^T\|_2 \|b-\tilde{b}\|_2$$
$$\le \|(A^TA)^{-1}\|_2 \|A^T\|_2 \|b-\tilde{b}\|_2$$
$$= \frac{\sigma_1}{\sigma_n^2} \|b-\tilde{b}\|_2$$

and by $\|b\|_2 \le \sigma_1 \|x\|_2$,

$$\frac{\|x-\tilde{x}\|_2}{\|x\|_2} \le \frac{\sigma_1^2}{\sigma_n^2} \frac{\|b-\hat{b}\|_2}{\|b\|_2}$$

(d) $\|Ax-b\|_2^2 = x^TA^TAx - 2x^TA^Tb + b^Tb$

Let its derivative be 0, i.e. $A^TAx = A^Tb$ (*)

Extending the columns of U and V to make it an orthonormal basis, we have $A = \tilde{U}\tilde{\Sigma}\tilde{V}^T$, where $\tilde{U} = (U, u_{r+1}, \cdots, u_m)$, $\tilde{V} = (V, v_{r+1}, \cdots, v_n)$, $\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}_{m\times n}$

Then (*) becomes $\tilde{\Sigma}^T\tilde{\Sigma}\tilde{V}^Tx = \tilde{\Sigma}^T\tilde{U}^Tb$.

Let $z = \tilde{V}^Tx$ and $z = \begin{pmatrix} z_r \\ z_{n-r} \end{pmatrix}$, then by solving the equation, we have $z_r = \Sigma^{-1}U^Tb$ and $z_{n-r} \in \mathbb{R}^{n-r}$.

$$x = \tilde{V}z = (V, V_{n-r})\begin{pmatrix} \Sigma^{-1}U^Tb \\ z_{n-r} \end{pmatrix}$$
$$= V\Sigma^{-1}U^Tb + V_{n-r}z_{n-r}$$

By the construction of V_{n-r}, we know that

$$\mathrm{Ran}(V_{n-r}) = \mathrm{Ran}(V)^{\perp} = \mathrm{Ker}(A).$$

Since $z_{n-r} \in \mathbb{R}^{n-r}$ is arbitrary, $y = V_{n-r}z_{n-r} \in \mathrm{Ker}(A)$.

In conclusion, all the solutions are

$$x = V\Sigma^{-1}U^Tb + y, \quad y \in \mathrm{Ker}(A).$$

Since $V\Sigma^{-1}U^Tb \in \mathrm{Ran}(V) = \mathrm{Ker}(A)^{\perp}$,

$$\|x\|_2^2 = \|V\Sigma^{-1}U^Tb + y\|_2^2$$
$$= \|V\Sigma^{-1}U^Tb\|_2^2 + \|y\|_2^2$$

In order to minimize 2-norm of x, we choose $y = 0$.

Thus $x_0 = V\Sigma^{-1}U^Tb$ is the solution of LS with the minimum 2-norm.