Notes on Matrix Computation

University of Chicago, 2014

Vivak Patel

September 7, 2014

Contents

1 Introduction

1.1 Variations of solving $Ax = b$

1. Linear Regression. A is known and b is known but corrupted by some error unknown error r. Our goal is to find x such that:

$$x \in \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \arg\min\{\|r\|_2^2 : Ax + r = b\}$$

2. Data Least Squares. A is known but corrupted by some unknown error E. We want to determine:

$$x \in \arg\min_{x \in \mathbb{R}^n} \{\|E\|_F^2 : (A + E)x = b\}$$

3. Total least squares. A and b are both corrupted by errors E and r (resp.). We want to determine:

$$x \in \arg\min_{x \in \mathbb{R}^n} \{\|E\|_F^2 + \|r\|_2^2 : (A + E)x + r = b\}$$

4. Minimum norm least squares. Given any A and b, we want x such that:

$$x \in \arg\min\{\|z\|_2^2 : z \in \arg\min \|Az - b\|_2^2\} = \arg\min\{\|z\|_2^2 : A^T Az = A^T b\}$$

5. Robust Regression. The linear regression problem with a different norm for the error r.

6. Regularized Least Squares. Given a matrix Γ, we want to find:

$$x \in \arg\min\{\|Ax - b\|_2^2 + \|\Gamma x\|_2^2\}$$

7. Linear Programming. $x \in \arg\min\{c^T x : Ax \leq b\}$

8. Quadratic Programming. $x \in \arg\min\{0.5x^T Ax + c^T x : Bx = d\}$

1.2 Norms

1. Norm.

   Definition 1.1. *A norm is a real-valued function defined over a vector space, denoted by $\|\cdot\|$ such that:*

   (a) $\|x\| \geq 0$
   (b) $\|x\| = 0$ *if and only if* $x = 0$
   (c) $\|x + y\| \leq \|x\| + \|y\|$
   (d) $\|\alpha x\| = |\alpha| \|x\|$ *for any scalar* α

2. Vector Norms

   (a) p-norms

      i. For $p \geq 1$ the p-norm of $x \in V$ is $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$

ii. For $p = \infty$, the ∞-norm or Chebyshev norm is $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$

iii. The Chebyshev norm is the limit of p-norms

Lemma 1.1. *Let $x \in V$ then $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$.*

Proof. Let $\|x\|_\infty = 1$. Then $\exists j \in \{1, \dots, n\}$ such that $x_i = 1$. Then:

$$\lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \leq \lim_{p\to\infty} n^{1/p} = 1$$

Secondly,

$$\lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \geq \lim_{p\to\infty} |x_j| = 1$$

□

iv. Weighted p-norms: add a non-negative weight term to each component in the sum.

(b) Mahalanobis norm. Let A be a symmetric matrix.

$$\|x\|_A = \sqrt{x^* A x}$$

3. Matrix Norms

(a) Compatible. Submultiplicative/Consistent.

Definition 1.2. *Let $\|\cdot\|_M$ be a matrix norm and $\|\cdot\|_V$ be a vector norm.*

i. A matrix norm is compatible with a vector norm if:

$$\|Ax\|_V \leq \|A\|_M \|x\|_V$$

ii. A matrix norm is consistent or submultiplicative if:

$$\|AB\|_M \leq \|A\|_m \|B\|_M$$

(b) Holder Norms

i. The Holder p-norm of A is $\|A\|_{H,p} = \left(\sum_{j=1}^{m} \sum_{i=1}^{n} |a_{ij}|^p \right)^{1/p}$

ii. The Holder 2-norm is called the Frobenius norm

iii. The Holder ∞-norm is $\|A\|_{H,p} = \max_{i,j} |a_{ij}|$

(c) Induced Norms

i. Induced Norms. Spectral Norm

Definition 1.3. *Let $\|\cdot\|_\alpha, \|\cdot\|_\beta$ be vector norms. The matrix norm $\|\cdot\|_{\alpha,\beta}$ is the induced norm defined by:*

$$\|A\|_{\alpha,\beta} = \max_{x\neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\beta}$$

When $\alpha = \beta = 2$, the induced norm is called the spectral norm.

ii. Equivalent Definitions

Lemma 1.2. *The following are equivalent definitions for an induced norm:*

A. $\|A\|_{\alpha,\beta} = \sup\{\|Ax\|_\alpha : \|x\|_\beta = 1\}$

B. $\|A\|_{\alpha,\beta} = \sup\{\|Ax\|_\alpha : \|x\|_\beta \leq 1\}$

Proof. Let $x = \frac{v}{\|v\|_\beta}$. Using this in the definition, we see that the definition and first characterization are equivalent. For the second characterization, already have that the norm from the second characterization is necessarily less than or equal to the one from the definition. Assume that it is strictly less than the one from the definition. From the first characterization, we know that the norm is achieved and so let this point be x'. At $x'/2$ the norm is still maximized, so the definitions are equivalent. □

iii. Compatibility

Lemma 1.3. *Letting* $\|\cdot\| = \|\cdot\|_{\alpha,\beta}$*:*

$$\|Ax\| \leq \|A\| \, \|x\|$$

Proof. For any $x \neq 0$:

$$\|A\| =\geq \frac{\|Ax\|}{\|x\|}$$

When $x = 0$, the result holds simply by plugging in values. □

iv. Consistency

Lemma 1.4. *Letting* $\|\cdot\| = \|\cdot\|_{\alpha,\beta}$*:*

$$\|AB\| \leq \|A\| \, \|B\|$$

Proof.

$$\begin{aligned} \|AB\| &= \max \frac{\|A(Bx)\|}{\|x\|} \\ &\leq \|A\| \max \frac{\|Bx\|}{x} \\ &\leq \|A\| \, \|B\| \end{aligned}$$

□

v. Computing $\|\cdot\|_{1,1}$ norm. $A \in F^{m \times n}$.

Lemma 1.5.

$$\|A\|_{1,1} = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$$

Proof. There exists an x such that $\|x\|_1 = 1$ and $\|Ax\|_1 = \|A\|$. Therefore:

$$\begin{aligned}\|A\| &= \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}||x_j| \\ &= \sum_{j=1}^{n}|x_j|\sum_{i=1}^{m}|a_{ij}| \\ &\leq \left(\max_j \sum_{i=1}^{m}|a_{ij}|\right)\sum_{j=1}^{n}|x_j| \\ &\leq \max_j \sum_{i=1}^{m}|a_{ij}|\end{aligned}$$

For the other direction, suppose the maximum occurs at the kth column. Then $\|Ae_k\|_1 \leq \|A\|$ but the left hand side is right hand side of the array of equations above. □

vi. Computing $\|\cdot\|_\infty$.

Lemma 1.6.

$$\|A\|_\infty = \max_{i=1,\ldots,m}\sum_{j=1}^{n}|a_{ij}|$$

Proof. There is an x such that $\|x\|_\infty = 1$ and $\|Ax\|_\infty = \|A\|$. Therefore:

$$\begin{aligned}\|A\| &= \|Ax\|_\infty \\ &= \max_{i=1,\ldots,m}\sum_{j=1}^{n}|a_{ij}x_j| \\ &\leq \max_{i=1,\ldots,m}\sum_{j=1}^{n}|a_{ij}|\end{aligned}$$

For the other direction, let k be the index of the maximizing row. Let x be a vector such that $x_i = sgn(a_{ki})$. Then $\|x\| = 1$ and $\|A\| \geq \|Ax\|_\infty = \sum_{j=1}^{n}|a_{kj}|$. □

1.3 Error Analysis

1. Types of Error given for true value x and computed value $\hat{x}$
   (a) $\|\hat{x} - x\|$ is the absolute error, but it depends on units
   (b) $\|\hat{x} - x\| / \|x\|$ is the relative error, and it does not depend on units
   (c) Pointwise error: compute $\|y\|$ where $y_i = \frac{\hat{x}_i - x_i}{x_i}\mathbf{1}[x_i \neq 0]$.
2. Backwards Error Analysis
   (a) Notation

i. Suppose we want to solve $Ax = b$ and we denote $\Delta(A, b) = \hat{x}$ the algorithm which produces the estimate.
ii. The condition number of a matrix $\kappa(A) = \|A\| \left\|A^{-1}\right\|$.
iii. Let $\rho = \|\delta A\| \|A\|$ for some small perturbation matrix δA.

(b) Idea: View $\hat{x}$ as the solution to a nearby system $(A + \delta A)x = b + \delta b$.

(c) Error bound

Lemma 1.7. *Suppose A is an invertible matrix and we have a compatible norm. If $\frac{\|\delta A\|}{\|A\|} \leq \epsilon$ and $\frac{\|\delta b\|}{\|b\|} \leq \epsilon$ then*

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{2\epsilon}{1 - \rho}\kappa(A)$$

Proof. Note that:

$$(I + A^{-1}\delta A)(\hat{x} - x) = A^{-1}(\delta b - \delta A x)$$

Then:

$$(1 - \rho)\|\hat{x} - x\| \leq \left\|A^{-1}\right\| (\|\delta b\| + \|\delta A\| \|x\|)$$

Dividing both sides by $\|x\|$ and multiplying the right hand side by $1 = \|A\| / \|A\|$:

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \kappa(A) \left(\frac{\|\delta b\|}{\|A\| \|x\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

Noting that $\|b\| \leq \|A\| \|x\|$, the result follows. □

1.4 Floating Point Numbers

1. Motivation: computers do not have infinite memory and so they can only store numbers up to a certain precision

2. Floating Point Numbers, sign, mentissa, exponent, base

   Definition 1.4. *Floating Point Numbers are $F \subset \mathbb{Q}$, that have the following representation:*

   $$\pm a_1 a_2 \dots a_k \times b^{e_1 \dots e_l}$$

   (a) $\pm$ is called the sign
   (b) $a_1 \dots a_k$ is called the mentissa and a_i are values in some finite field
   (c) $e_1 \dots e_l$ is called the exponent and e_k are values in some finite field
   (d) b is the base

3. Floating Point Representation Standards

   (a) Floating Point Representation. Machine Precision.

Definition 1.5. *A floating point representation is a function* $fl : \mathbb{R} \to F$ *which is characterized by the machine precision denoted* ϵ_m *and defined as:*

$$\epsilon_m = \inf\{x \in \mathbb{R} : x > 0, fl(1+x) \neq 1\}$$

(b) Standard 1: $\forall x \in \mathbb{R}$, $\exists x' \in F$ such that $|x - x'| \leq \epsilon_m |x|$

(c) Standard 2: $\forall x, y \in \mathbb{R}$ and there is an $|\epsilon_1| \leq \epsilon_m$:

$$fl(x \pm y) = (x \pm y)(1 + \epsilon_1)$$

(d) Standard 3: $\forall x, y \in \mathbb{R}$ and there is an $|\epsilon_2| \leq \epsilon_m$:

$$fl(xy) = (xy)(1 + \epsilon_2)$$

(e) Standard 4: $\forall x, y \in \mathbb{R}$ such that $y \neq 0$, there is an $|\epsilon_3| \leq \epsilon_m$:

$$fl(x/y) = (x/y)(1 + \epsilon_3)$$

4. Floating Point in Computers

(a) Fields: $b = 2$ and $a_i, e_j \in \{0, 1\}$

(b) Storage

i. A floating point number requires $1 + l + k$ bits of storage using the following storage:

$$\pm |e_1|e_2| \cdots |e_l|a_1|a_2| \cdots |a_k$$

ii. 32-bit computers: 1 bit for sign, 8 bits for exponent, and 23 bits for the mentissa

iii. 64-bit computers: 1 bit for sign, 11 bits for exponent, and 52 bits for the mentissa

(c) Errors and (typical) Handling

i. Round-off Error is when the number is more precise than the mentissa allows, and is handled by cutting off lower priority values.

ii. Overflow Error is when the exponent is too large, and is handled by returning representations of the largest value allowed by the system (e.g. $1e-99$ or $-\infty$, etc.)

iii. Underflow Error is when the exponent is too negative, resulting in 0.

5. Examples

Example 1.1. *Computing* l^2 *norm. Suppose*

$$x = \begin{bmatrix} 10^{-49} & 10^{-50} & 10^{-50} & \cdots & 10^{-50} \end{bmatrix} \in \mathbb{R}^{101}$$

This can be stored exactly, but for $i = 2, \ldots, 101$, $fl(x_i^2) = fl(10^{-100}) = 0$. *Hence, using a naive algorithm would compute* $\|x\|_{l^2} = 10^{-49}$, *which has a* $\sim 12\%$ *(check) absolute error. An improved algorithm uses the naive algorithm on* $\hat{x} = x/\|x\|_\infty$. *This does not produce underflow errors.*

Example 1.2. *Sample Variance. There are two methods of computing sample variance.*

(a) First compute $\bar{x}$ *and then compute* $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

(b) Compute $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right]$ *The first method is more accurate but requires two passes over the data while the second only requires one.*

2 Eigenvalue Decomposition

2.1 Eigenvalues and Eigenvectors

1. Eigenvalue. Eigenvector.

   Definition 2.1. *An eigenvector, x, and eigenvalue, λ of a matrix A satisfy $Ax = \lambda x$.*

2. Basic Properties

   Lemma 2.1. *Let $A \in \mathbb{C}^{n\times n}$.*

   (a) Eigenvectors are scale invariant

   (b) The eigenspace of an eigenvalue λ, $V_\lambda := \{x \in \mathbb{C}^n : Ax = \lambda x\}$, is a vector space.

   (c) Any $n \times n$ matrix has n eigenvalues counted with multiplicity.

3. Eigenvalue Decompositions

   (a) Eigenvalue Decomposition

   Definition 2.2. *A matrix $A \in \mathbb{C}^{n\times n}$ admits an eigenvalue decomposition if there exists an invertible matrix X and diagonal matrix Λ of eigenvalues of A such that:*

   $$A = X\Lambda X^{-1}$$

   (b) Equivalent conditions

   Proposition 2.1. *The following three statements are equivalent for $A \in \mathbb{C}^{n\times n}$:*

   i. A is diagonalizable

   ii. A has an eigenvalue decomposition

   iii. A has n linearly independent eigenvectors

   Proof. We only prove equivalence between the last two. The first requires more machinery:

   $$X = \begin{bmatrix} x_1 & x_2 & \cdots x_n \end{bmatrix} \Leftrightarrow Ax_i = \lambda x_i$$

   X is invertible if and only if it has n linearly independent columns. □

2.2 Jordan Canonical Form

1. Every matrix $A \in \mathbb{C}^{n\times n}$ has a Jordan Canonical form $A = XJX^{-1}$ where:

   (a) J is a bidiagonal matrix where the main diagonal of J contains the eigenvalues of A with their algebraic multiplicity, and the super-diagonal of J is either 0 or 1. The remaining entries of J are 0.

   (b) The matrix X contains the eigenvectors of A and some other vectors

(c) Note that if the superdiagonal is all 0s then J is diagonalizable.

2. Jordan Blocks

(a) The matrix J can be written in block form with matrices $J_1, \ldots, J_k$:

$$\begin{bmatrix} J_1 & O & \cdots & O \\ O & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & J_k \end{bmatrix}$$

(b) The J_i are the Jordan blocks. Let $\lambda \in \Lambda(A)$ then a Jordan Block is of the form:

$$\begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{bmatrix}$$

(c) The number of Jordan blocks corresponding to a specific $\lambda_i \in \Lambda(A)$ is equal to its geometric multiplicity (that is, the dimension of the eigenspace)

(d) The sum of the dimensions of the Jordan blocks for any λ_i is equal to the algebraic multiplicity of the eigenvalue

3. Basic Properties

(a) Fundamental Lemma

Lemma 2.2. *Let J_r be a Jordan Block of a Jordan Canonical Form with eigenvalue λ_r. Let n_r be its dimension. Then for $k = 1, \ldots, n_r - 1$ $(J_r - \lambda_r I)^k e_k = 0$*

Proof. Let $diag_k(A)$ be the vector of values from A of the form $A_{i,i+k}$. Then, we note the following:

$$diag_1(J_r - \lambda_r I) = \mathbf{1}_{n_r - 1}$$

while all other elements in the matrix are 0. And in general:

$$diag_k((J_r - \lambda_r I)^k) = \mathbf{1}_{n_r - k}$$

while all other elements of the matrix are 0. So when we multiply by e_k we will get 0. □

(b) Powers of Jordan Canonical Form:

Lemma 2.3.

$$A^k = X J^k X^{-1}$$

4. Drawback: J cannot be computed.

2.3 Spectra

1. Spectra. Principal Eigenvalues.

   Definition 2.3. *Let $A \in \mathbb{C}^{n\times n}$. The set of eigenvalues of A is called the spectra of A and is denoted $\lambda(A)$. The maximum element of $\lambda(A)$ is called the Principal Eigenvalue of A.*

2. Spectral Theorems

   (a) Hermitian Adjoint. Normal Matrix. Hermitian Matrix.

   Definition 2.4. *The hermitian adjoint of a matrix A is the transpose of A with its elements conjugated. A matrix A is a normal matrix if $A^*A = AA^*$. A is a hermitian matrix if $A = A^*$.*

   (b) Spectral Theorem for Normal Matrices

   Theorem 2.1. *Let $A \in \mathbb{C}^{n\times n}$. Then the following are equivalent:*

   i. A is unitarily diagonalizable
   ii. A has an orthonormal eigenbasis
   iii. A is a normal matrix
   iv. A has an EVD of the form $A = V\Lambda V^*$ where V is unitary (i.e. $VV^* = V^*V = I$)

   (c) Spectral Theorem for Hermitian Matrices

   Theorem 2.2. *Let $A \in \mathbb{C}^{n\times n}$. Then the results of the Spectral Theorem for Normal Matrices applies and the eigenvalues of A are real.*

2.4 Spectral Radius

1. Spectral Radius

   Definition 2.5. *The spectral radius $\rho(A)$ of a matrix $A \in \mathbb{C}^{n\times n}$ is the largest absolute eigenvalue:*

   $$\rho(A) = \max\{|\lambda| : \lambda \in \Lambda(A)\}$$

   Note 2.1. *The spectral radius is not a norm. Consider $J \in \mathbb{R}^{2\times 2}$ where all entries are zero except $J_{12} = 1$. Then $\rho(J) = 0$ but $J \neq 0$, hence, ρ cannot be a norm.*

2. Minimality of ρ and norms

   Lemma 2.4. *If $A \in \mathbb{C}^{n\times n}$ and $\|\cdots\|$ is a compatible norm then $\rho(A) \leq \|A\|$*

   Proof. Let λ be an eigenvalue of A and x be its corresponding eigenvector. Then:

   $$\lambda \|x\| = \|Ax\| \leq \|A\| \|x\|$$

   Since λ is arbitrary in $\Lambda(A)$ the result follows.s □

3. Approximating Spectral Radius

Theorem 2.3. *Let $A \in \mathbb{C}^{n\times n}$ and $\epsilon > 0$. There exists an operator norm $\|\cdot\|_\alpha$ depending on A and ϵ such that $\|A\|_\alpha \leq \rho(A) + \epsilon$.*

4. A limiting property

Lemma 2.5. *Let $A \in \mathbb{C}^{n\times n}$. $A^m \to O$ as $m \to \infty$ if and only if $\rho(A) < 1$.*

Proof. Let λ be the largest eigenvalue and x be an eigenvector. Then, $A^m x = \lambda^m x$. If $A^m \to O$ as $m \to \infty$ then taking the limit on both sides implies that $\lambda^m \to 0$ as $m \to \infty$ and so $\lambda < 1$. For the other direction, let $\epsilon > 0$ such that $\rho(A) + 2\epsilon = 1$. Then, we can find an α such that:

$$\|A^m\|_\alpha \leq \|A\|_\alpha^m \leq (\rho(A) + \epsilon)^m < 1$$

Hence, $A^m \to 0$. □

2.5 Diagonal Dominance and Gerschgorin's Disk Theorem

1. Diagonally Dominant

Definition 2.6. *A matrix $A \in \mathbb{C}^{n\times n}$ is strictly diagonally dominant if $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for all $i \in 1, \ldots, n$. It is diagonally dominant if the inequality is not strict.*

2. Diagonal dominance and non-singularity

Lemma 2.6. *A strictly diagonally dominant matrix is nonsingular.*

Proof. Suppose A is a strictly diagonally dominant matrix and it is singular. That is $\exists x \neq 0$ such that $Ax = 0$. Let k be the index of the absolute largest element of the vector x. Then:

$$0 = \sum_{i=1}^{n} a_{ki}x_i \implies -a_{kk}x_k = \sum_{i\neq k} a_{ki}x_i$$

Then:

$$|a_{kk}||x_k| \leq \sum_{i\neq k} |a_{ki}||x_i| \leq |x_k| \sum_{i\neq k} |a_{ki}| < |x_k||a_{kk}|$$

Hence, A cannot be singular. □

3. Gerschgorin's Disk Theorem

   (a) Gerschgorin's Disks

   Definition 2.7. *For $A \in \mathbb{C}^{n\times n}$, define for $i = 1, \ldots, n$ the disks $G_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i|$ where $r_i = \sum_{j\neq i} |a_{ij}|$. The G_i are called Gerschgorin's Disks.*

   (b) Gerschgorin's Disk Theorem

Theorem 2.4. *Let $A \in \mathbb{C}^{n\times n}$ and G_i be its disks and $\Lambda(A)$ be its spectra.*

i. $\Lambda(A) \subset \bigcup_{i=1}^{n} G_i$

ii. The number of eigenvalues (with multiplicity) in each connected component of $\bigcup_{i=1}^{n} G_i$ is the number of G_i constituting that component.

Proof. Suppose there is a $\lambda \in \Lambda(A)$ such that $\lambda \notin G_i$ for any i. Then $\forall i,\ |\lambda - a_{ii}| > r_i$. Hence, $A - \lambda I$ is a strictly diagonally dominant matrix which implies that $\det(A - I\lambda) \neq 0$, but this is a contradiction since λ is an eigenvalue. For the second part, we use a bit of a trick. Let $t \in [0, 1]$. And let $A(t)$ be the matrix whose off diagonals of A multiplied by t. Note that $G_i(t) \subset G_i$. Moreover, since eigenvalues are a continuous function of t, $A(0)$ and $A(1)$ have the same number of eigenvalues in each connected component of $\bigcup_i G_i$. □

3 Singular Value Decomposition

3.1 Theory

1. Singular Value Decomposition (SVD)

   Definition 3.1. *Let $A \in \mathbb{C}^{m\times n}$. The singular value decomposition (SVD) of A is $U\Sigma V^*$, where:*

   (a) $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary matrices

   (b) The columns of U are the left singular vectors of A

   (c) The columns of V are the right singular vectors of A

   (d) $\Sigma \in \mathbb{R}^{m\times n}_{\geq 0}$ is a diagonal matrix with values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ on the diagonal where r is the rank of A

   (e) $A = U\Sigma V^*$

2. Properties of Singular Values and Singular Vectors

   Lemma 3.1. *The left singular vectors of A are the eigenvectors of AA^*. The right singular vectors of A are the eigenvectors of A^*A. The square of singular values of A are the eigenvalues of AA^* and A^*A.*

   Proof. From $Ay = \sigma x$ and $A^*x = \sigma y$. Then:

   (a) $A^*Ay = \sigma A^*x = \sigma^2 y$

   (b) $AA^*x = \sigma Ay = \sigma^2 x$

   □

3. Other Forms of SVD

   (a) Compact/Reduced SVD

   Definition 3.2. *The compact or reduced SVD of a matrix $A \in \mathbb{C}^{m\times n}$ can be written as $A = U\Sigma V^*$ where, if $r = rank(A)$:*

   i. $U \in \mathbb{C}^{m\times r}$, whose columns are the left singular vectors of A corresponding to non-zero singular values, and $U^*U = I_r$

   ii. $V \in \mathbb{C}^{n\times r}$, whose columns are the right singular vectors of A corresponding to non-zero singular values, and $V^*V = I_r$

   iii. $\Sigma \in \mathbb{R}^{r\times r}_{\geq 0}$ is a diagonal matrix whose diagonal values are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$.

   (b) Rank-1 SVD

   Definition 3.3. *Letting $u_1, \ldots, u_r$ and $v_1, \ldots, v_r$ be the left and right singular vectors, and $\sigma_1, \ldots, \sigma_r$ be the singular values. then, the Rank-1 SVD is:*

   $$A = \sigma_1 u_1 v_1^* + \cdots + \sigma_r u_r v_r^*$$

4. Existence of SVD

Theorem 3.1. *Every matrix has a condensed SVD*

Proof. There are three steps:

(a) Constructing the Wielandt Matrix and Characterizing its Eigenvalues

$$W = \begin{bmatrix} O & A \\ A^* & O \end{bmatrix} = W^*$$

Since W is Hermitian, by the spectral theorem, it has an eigenvalue decomposition $W = Z\Lambda Z^*$ with real eigenvalues. Suppose A has rank r, then W has rank $2r$, so it has $2r$ eigenvalues (since it is Hermitian). Let $z^T = \begin{bmatrix} x & y \end{bmatrix}$ be the transpose of a column of Z and σ be the corresponding eigenvalue. Then, $Wz = \sigma z$ which implies $Ay = \sigma x$ and $A^*x = \sigma y$. Moreover, we see that $\begin{bmatrix} x & -y \end{bmatrix}$ is also an eigenvector of W with eigenvalue $-\sigma$. Hence, $\Lambda = diag(\sigma_1, \sigma_2, \ldots, \sigma_r, -\sigma_r, \ldots, \sigma_1, 0, \ldots, 0)$.

(b) Normalizing columns of Z and showing $A = Y\Sigma_r X^*$. Let $\Sigma_r = diag(\sigma_1, \ldots, \sigma_r)$.And normalize the eigenvectors of W so that $z^*z = 2$. Then, $x^*x + y^*y = 2$. Also, since the eigenvectors are orthogonal and we have from the first part that $x^*x - y^*y = 0$. This implies that $x^*x = y^*y = 1$. Rewriting Z in block notation with X,Y, and $-Y$ blocks, we have that $A = Y\Sigma_r X^*$.

(c) Orthonormality of X and Y. We now show that the columns of X are orthonormal to themselves and similarly with Y. This follows from the orthonormality of z's and specifically considering $\begin{bmatrix} x & -y \end{bmatrix}$.

□

3.2 Applications

1. Solving Linear Systems

   Example 3.1. *Suppose we want to solve $Ax = b$, and the system is consistent (that is, $b \in \Im(A)$). If $A = U\Sigma V^*$ is the full SVD of A then $\Sigma V^*x = U^*b$. Letting $y = V^*x$ and $c = U^*b$, $\Sigma y = c$ can be solved using back-solve and $y_{r+1}, \ldots, y_n$ are free parameters. Then $Vy = x$.*

2. Inverting Non-singular Matrices

   Example 3.2. *If A is non-singular, then Σ has no zeros on its diagonal. Therefore:*

   $$A^{-1} = (U\Sigma V^*)^{-1} = (V^*)^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^*$$

   Moreover, if $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ then $\Sigma^{-1} = diag(\sigma_1^{-1}, \ldots, \sigma_n^{-1})$

3. Computing the 2-norm

   (a) SVD and EVD of Hermitian, psd matrices

Lemma 3.2. *If M is Hermitian, positive semi-definite then its EVD and SVD coincide.*

Proof. By the spectral theorem $M = X\Lambda X^*$ be the EVD of M where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$. Since M is positive semi-definite, $\lambda_i \geq 0$. Therefore, the EVD is an SVD. □

(b) The 2-norm is unitary invariant

Lemma 3.3. *Let U, V be unitary matrices. Then $\|UAV\|_2 = \|A\|_2$*

Proof. First:

$$\|UAV\|_2^2 = \max_{\|x\|_2=1} \|UAVx\|_2^2 = \max_{\|x\|=1} x^*V^*A^*AVx = \max_{\|x\|=1} \|AVx\|_2^2$$

Second: let x be any vector and v_i be the columns of V. Then, $\|Vx\|_2^2 = \sum_{i=1}^n \bar{x}_i v_i^* v_i x = \sum_{i=1}^n |x_i|^2$. And since V is invertible:

$$\|AV\|_2^2 = \max_{\|Vx\|_2^2=1} \|AVx\|_2^2 = \max_{\|y\|=1} \|Ay\|_2^2 = \|A\|_2^2$$

□

(c) 2-norm and Singular Values

Corollary 3.1. *Let the SVD of A be $U\Sigma V^*$. Then $\|A\|_2 = \sigma_1$.*

Proof. $\|A\|_2^2 = \|\Sigma\|_2^2$. Let x be such that $x_1 = 1$ and all other indices are zero, then this maximizes $\|\Sigma x\|_2^2 = \sigma_1^2$. □

4. Computing the Frobenius Norm

(a) Frobenius Norm is unitary invariant

Lemma 3.4. *Let U, V be unitary matrices. Then $\|UAV\|_F = \|A\|_F$.*

Proof.

$$\begin{aligned}
\|UAV\|_F^2 &= tr(V^*A^*U^*UAV) \\
&= tr(V^*A^*AV) \\
&= tr(VV^*A^*A) \\
&= tr(A^*A) \\
&= \|A\|_F^2
\end{aligned}$$

□

(b) Frobenius Norm and Singular Values

Corollary 3.2. *Let the SVD of A be $U\Sigma V^*$. Then*

$$\|A\|_F^2 = \sum_{i=1}^{rank(A)} \sigma_i^2$$

Proof. $\|A\|_F^2 = \|\Sigma\|_F^2 = \sum_{i=1}^{rank(A)} \sigma_i^2$ □

5. Schatten and KyFan Norms

   (a) Schatten p-norm

   Definition 3.4. *For $p \in [1, \infty)$, the Schatten p-norm is*

   $$\|A\|_{\sigma,p}^p = \sum_i \sigma_i(A)^p$$

   When $p = \infty$, $\|A\|_{\sigma,\infty}$ $s = \max_i \sigma_i(A)$.

   (b) Examples of Schatten p-norm

   Example 3.3. *From the definition:*

   i. $\|A\|_{\sigma,1} = \sum_i |\sigma_i(A)|$. *This is also denoted* $\|A\|_*$ *and called the nuclear norm.*

   ii. $\|A\|_{\sigma,2} = \|A\|_F$

   iii. $\|A\|_{\sigma,\infty} = \|A\|_2$

   (c) Ky Fan (p, k)-norm

   Definition 3.5. *The Ky Fan (p, k)-norm of A is* $\|A\|_{\sigma,p,k}^p = \sum_{i=1}^k \sigma_i(A)^p$.

6. Computing the Magnitude of the Determinant

   (a) The Eigenvalues of Unitary Matrices are 1.

   Lemma 3.5. *The eigenvalues of unitary matrices are all* 1.

   Proof. Let U be unitary and x be an eigenvector with eigenvalue λ. Then $Ux = \lambda x$. Then:

   $$\|x\|_2 = \|Ux\|_2 = \|\lambda x\|_2 = \|x\|_2 |\lambda|$$

   □

   (b) Let A be a matrix with SVD $U\Sigma V^*$ then:

   $$|\det(A)| = |\det(U)\det(\Sigma)\det(V)| = |\det(\Sigma)| = \prod_{i=1}^n \sigma_i(A)$$

7. Existence and Computing of Pseudo Inverses

   Theorem 3.2. *For any $A \in \mathbb{C}^{m\times n}$, $\exists A^\dagger \in \mathbb{C}^{n\times m}$ such that:*

   (a) Symmetries: $(AA^\dagger)^* = AA^\dagger$ and $(A^\dagger A)^* = A^\dagger A$

   (b) "Identity": $AA^\dagger A = A$ and $A^\dagger AA^\dagger = A^\dagger$

   Proof. Let $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0) \in \mathbb{C}^{m\times n}$. Then

   $$\Sigma' = diag(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0)$$

   satisfy these conditions. So we denote $\Sigma^\dagger = \Sigma'$. Now let A have SVD $U\Sigma V^*$. Letting $A' = V\Sigma^\dagger U^*$:

(a) Symmetries: $(AA^\dagger)^* = (U\Sigma\Sigma^\dagger U^*)^* = U\Sigma\Sigma^\dagger U^* = AA^\dagger$. The second one follows similarly.

(b) "Identity": $AA^\dagger A = U\Sigma V^* V\Sigma^\dagger U^* U\Sigma V^* = U\Sigma\Sigma^\dagger\Sigma V^* = A$. The second one follows similarly.

Hence, $A^\dagger$ exists and $A^\dagger = V\Sigma^\dagger U^*$. □

8. Fredholm Alternative

(a) General Linear Group, Kernels and Images

Lemma 3.6. *If $A \in \mathbb{C}^{m\times n}$ and $S \in GL(n)$ and $T \in GL(m)$ then $\ker(TA) = \ker(A)$ and $im(AS) = im(A)$.*

Proof. If $y \in im(A)$ then $\exists x$ such that $Ax = y$. Since S is invertible, $\exists z$ such that $Sz = x$. In the other direction, if $\exists z$ such that $ASz = y$ then letting $x = Sz$, $Ax = y$. For the kernels, if $x \in \ker(A)$ then $Ax = 0$ so $TAx = 0$, which implies $x \in \ker(TA)$. If $x \in \ker(TA)$, then $TAx = 0$ and since T is invertible, $Ax = 0$. □

(b) Co-kernel. Co-image.

Definition 3.6. *Let $A \in \mathbb{C}^{m\times n}$. $\ker(A^*)$ is the co-kernel of A and $im(A^*)$ is its co-image.*

(c) SVD and Spans of Kernel, Co-Kernel, Image and Co-Image.

Proposition 3.1. *Let A be a complex valued $m \times n$ matrix. Let $u_1, \cdots, u_m$ and $v_1, \ldots, v_n$ be the left and right singular vectors of A. Then:*

i. $\ker(A) = span\{v_{r+1}, \ldots, v_n\}$
ii. $\ker(A^*) = span\{u_{r+1}, \ldots, u_m\}$
iii. $im(A) = span\{u_1, \ldots, u_r\}$
iv. $im(A^*) = span\{v_1, \ldots, v_r\}$

Proof. Let $U\Sigma V^*$ be the SVD of A and let $r = rank(A)$.

$$\begin{aligned}\ker(A) &= \ker(U\Sigma V^*) = \ker(\Sigma V^*)\\ &= V\ker(\Sigma)\\ &= V span\{e_{r+1}, \ldots, e_n\}\\ &= span\{v_{r+1}, \ldots, v_n\}\end{aligned}$$

Similar reasoning gives $\ker(A^*) = span\{u_{r+1}, \ldots, u_m\}$. For the image of A:

$$\begin{aligned}im(A) &= im(U\Sigma V^*) = im(U\Sigma)\\ &= U im(\Sigma)\\ &= U span\{e_1, \ldots, e_r\}\\ &= \{u_1, \ldots, u_r\}\end{aligned}$$

Similar reasoning gives $im(A^*) = span\{v_1, \ldots, v_r\}$. □

(d) Fredholm Alternative

Corollary 3.3. *Let $A \in \mathbb{C}^{m \times n}$. Then:*

i. $\ker(A) \perp im(A^*)$
ii. $\ker(A^*) \perp im(A)$
iii. $\ker(A) \oplus im(A^*) = \mathbb{C}^n$
iv. $\ker(A^*) \oplus im(A) = \mathbb{C}^m$

9. Projections

(a) Projection. Orthogonal Projection.

Definition 3.7. *Let $P \in \mathbb{C}^{n \times n}$.*

i. P is a Projection if it is idempotent (i.e. $P^2 = P$).
ii. P is an Orthogonal Projection if it is idempotent and Hermitian (i.e. $P^* = P$).

(b) Orthogonal Projections onto Kernel, Co-Kernel, Image and Co-Image

Lemma 3.7. *The following are orthogonal projections onto the respective subspaces:*

i. $P_{im(A)} = AA^\dagger$
ii. $P_{im(A^*)} = A^\dagger A$
iii. $P_{\ker(A^*)} = (I - AA^\dagger)$
iv. $P_{\ker(A)} = (I - A^\dagger A)$

Proof. First we check that the mappings have the correct target spaces. Let $U\Sigma V^*$ be the SVD of A. Then $AA^\dagger = U\Sigma\Sigma^\dagger U^*$. We now note two facts:

i. Since Σ is a diagonal matrix:

$$\Sigma\Sigma^\dagger = \begin{bmatrix} I_r & O \\ O & O \end{bmatrix}$$

ii. Secondly:

$$im(AA^\dagger) = im(U\Sigma\Sigma^\dagger U^*) = im(U\Sigma\Sigma^\dagger) = span\{u_1, \ldots, u_r\}$$

Similarly, $im(A^\dagger A) = span\{v_1, \ldots, v_r\}$. By the Fredholm alternative, the other two manipulations map to the kernel and co-kernel. Idempotence follows from the identity property of the Moore-Penrose Inverse. Orthogonality follows from the symmetric property of the Moore Penrose Pseudo inverse. □

10. Least Square Problem

(a) Problem: Find x which minimizes $\|b - Ax\|_2^2$.

Solution. Let $U\Sigma V^*$ be the SVD of A. Letting $y = V^*x$ and $c = U^*b$, we can restate the problem as finding y which minimizes:

$$\begin{aligned}\|b - U\Sigma y\|_2^2 &= \|U^*b - \Sigma y\|_2^2 \\ &= \|c - \Sigma y\|_2^2 \\ &= \sum_{i=1}^{rank(A)} (c_i - \sigma_i y_i)^2 + \sum_{i=rank(A)+1} s^n c_i^2\end{aligned}$$

This is minimized when $y_i = c_i/\sigma_i$ for $i = 1, \ldots, rank(A)$ and the other y_i are free to be whatever they choose. We can recover any solution of $x = Vy$. □

(b) It is clear that unless A is non-singular (i.e. has full rank), that the minimizers are not unique.

11. Minimum Length Least Squares Problem

(a) Problem: Find $x \in \arg\min\{\|x\|_2 : \|b - Ax\|_2^2 \leq \|b - Ay\|_2^2 \ \forall y\}$.

Solution. Again, using the fact that $\|x\| = \|Vy\| = \|y\|$, we see that $y_i = c_i/\sigma_i$ for $i = 1, \ldots, rank(A)$ and $y_i = 0$ for all other i recovers the minimum $\|x\|_2$. □

(b) Pseudo Inverse and Minimum Length Least Square Problem

Lemma 3.8. *The minimum length least squares solution $z = A^\dagger b$*

Proof. Using the Fredholm Alternative, $\exists b_1, b_2$ such that $b = b_1 + b_2$ and $b_1 \in im(A)$ and $b_2 \in \ker(A)$. Therefore $\|b - Ax\|_2^2 = \|b_1 - Ax\|_2^2 + \|b_2\|_2^2$. Moreover, $\exists x$ such that $b_1 = Ax$. Using the projections, $AA^\dagger b = b_1$ and so $AA^\dagger b = Ax$. Letting $z \in \ker(A)$ that is $z = (I - A^\dagger A)y$, we guess the solution $x = A^\dagger b + (I - A^\dagger A)y$. Plugging this in, we see that we do indeed have all of the solutions. Finally, we want to minimize $\|x\|_2^2$:

$$\begin{aligned}\|x\|_2^2 &= \left\|A^\dagger b\right\|_2^2 + 2\langle A^\dagger b, (I - A^\dagger A)y\rangle + \left\|(I - A^\dagger A)y\right\|_2^2 \\ &= \left\|A^\dagger b\right\|_2^2 + 2\langle (I - A^\dagger A)A^\dagger b, y\rangle + \left\|(I - A^\dagger A)y\right\|_2^2 \\ &= \left\|A^\dagger b\right\|_2^2 + \left\|(I - A^\dagger A)y\right\|_2^2\end{aligned}$$

This is minimized when $y = 0$. Therefore, $A^\dagger b = x$ is the minimum length least squares solution. □

12. Rank and Numerical rank

(a) Because of floating point errors, matrices are almost always of full rank on computers, even if they are not analytically

(b) We can use the decay rate of singular values to approximate the numerical rank. If the decay slows too much then we have likely reached the true rank of the matrix.

(c) Numerical Ranks

Definition 3.8. *Let $A \in \mathbb{C}^{m\times n}$ and $\tau > 0$ be a tolerance. Numerical ranks are:*

i. $\rho - rank(A) = \min\{r \in \mathbb{N} : \sigma_{r+1} \leq \tau\sigma_r\}$

ii. $\mu - rank(A) = \min\{r \in \mathbb{N} : \sum_{i\geq r+1} \sigma_i^2 \leq \tau \sum_{i\geq r} \sigma_i^2\}$

iii. $\nu - rank(A) = \|A\|_F^2 / \|A\|_2^2$

13. Finding Closest Unitary/Orthonormal Matrices

(a) Let $U(n) \subset GL(n)$ be all $n \times n$ unitary matrices, and $O(n) \subset U(n)$ be all $n \times n$ orthogonal matrices

(b) Closest Unitary Approximation

Lemma 3.9. *Let $A \in \mathbb{C}^{n\times n}$, then $\min_{X\in U(n)} \|A - X\|_F$ can be $X = UV^*$ where $U\Sigma V^*$ is the SVD of A.*

Proof. By unitary invariance, let $Z = U^*XV$. So we want to find:

$$\min_{Z\in U(n)} \|\Sigma - Z\|_F^2 = n + \min_{|z_i|=1} \sum_{i=1}^{r} (\sigma_i^2 - 2\sigma_i(Re(z_i) + iIm(z_i)))$$

since, Z must be diagonal to minimize the problem and the moduli of its elements must be 1. Using Lagrange multipliers, we have the Lagrangian:

$$\sigma_i^2 - 2\sigma_i Re(z_i) - i\sigma_i Im(z_i) + \lambda(Re(z_i)^2 + Im(z_i)^2 - 1)$$

Taking derivatives with respect to the real and imaginary parts of z_i and λ, we conclude that $Im(z_i) = 0$ and $Re(z_i) \in \{-1, 1\}$. The minimum occurs when $z_i = 1$ since $\sigma_i > 0$. So $z_1 = \cdots = z_r = 1$ and $z_{r+1}, \ldots, z_n$ are complex valued numbers with modulus 1. If they are required to be real, then the solution is unique. Hence, $UZV^* = X$ and in the real case $UV^* = X$. □

(c) Procrustes Problem (?)

Lemma 3.10. *$X = V_B U_B^* U_A V_A^*$ is a minimizer of $\min_{X\in U(n)} \|A - BX\|_F^2$ given A and B.*

Note 3.1. *This does not look to be true.*

14. Best r-rank approximation

(a) Problem: find $\arg\min_{X:rank(X)\leq r} \|A - X\|_2$ given A

(b) Eckart-Young Theorem:

Theorem 3.3. *Let the SVD of A be $\sum_{i=1}^{rank(A)} \sigma_i u_i v_i^*$. Then for any r the solution to the problem is $X = \sum_{i=1}^{r} \sigma_i u_i v_i^*$ and $\min \|A - X\|_2 = \sigma_{r+1}$*

Proof. Suppose $\exists B \in \mathbb{C}^{m\times n}$ such that $rank(B) \leq r$ and $\|A - B\|_2 < \sigma_{r+1}$.

i. By the rank-nullity theorem, the $nullity(B) \geq n - r$. Let $w \in \ker(B)$. Then $Bw = 0$. So:

$$\|Aw\|_2 = \|(A - B)w\|_2 \leq \|A - B\|_2 \|w\|_2 < \sigma_{r+1} \|w\|_2$$

ii. Let $v \in P := span(v_1, \ldots, v_{r+1})$ where $U\Sigma V^*$ is the SVD of A, and v_i are the columns of V. Then, $\exists z$ such that $v = V_{r+1}z$.

$$\|Av\|_2^2 = \|U\Sigma V^* V_{r+1} z\|_2^2 = \sum_{i=1}^{r+1} \sigma_i^2 |\alpha_i|^2 \geq \sigma_{r+1}^2 \|v\|_2^2$$

iii. Since $\dim(P) = r + 1$ and $nullity(B) \geq n - r$, $P \cap \ker(B) \neq \emptyset$. Hence, $\exists av \in P \cap \ker(B)$ such that $\sigma_{r+1} \|v\|_2 \leq \|Av\|_2^2 < \sigma_{r+1} \|v\|_2$ which is a contradiction. Thus $\|A - X\|_2 \geq \sigma_{r+1}$, and $\|A - X\|_2 = \sigma_{r+1}$ when X is given as in the theorem.

□

15. Least Squares with Quadratic Constraints

(a) Problem: For $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\alpha < \left\|A^\dagger b\right\|_2$, find

$$\arg\min\{\|b - Ax\|_2 : \|x\|_2 = \alpha\}$$

Solution. Let $U\Sigma V^*$ be the SVD of A. Let $U^*b = c$ and $V^*x = z$. Then we can restate the problem as:

$$\arg\min\{\|c - \Sigma z\| : \|z\|_2 = \alpha\{$$

The Lagrangian for this problem is then:

$$\mathcal{L}(z, \mu) = \|c - \Sigma z\|_2^2 + \mu(\|z\|_2^2 - \alpha^2)$$

Taking derivatives with respect to z and μ, we have the following system:

$$\begin{cases} -2\Sigma(c - \Sigma z) + 2\mu z = 0 \\ \|z\|_2^2 = \alpha^2 \end{cases}$$

The first equation gives that $(\Sigma^2 + \mu I)^{-1}\Sigma c = z$. More explicitly:

$$\sum_{j=1}^{rank(A)} \frac{\sigma_j c_j}{\sigma_j^2 + \mu} e_j = z$$

As long as $\mu > 0$ this matrix is invertible (since it is diagonalized). Now we can use the second equation to solve for μ.

$$\alpha^2 = \sum_{i=1}^{rank(A)} \left(\frac{\sigma_j c_j}{\sigma_j^2 + \mu} \right)^2$$

Using some numerical method, we can solve for μ. And use this to compute z. □

(b) Newton-Raphson is one option for solving for μ.

16. Generalized Condition Number

    (a) Generalized Condition Number

    Definition 3.9. *Given $A \in \mathbb{C}^{m\times n}$, its Moore-Penrose Inverse $A^\dagger$, and a matrix norm $\|\cdot\|$, the Generalized Condition Number of A is $\kappa(A) = \|A\| \left\|A^\dagger\right\|$*

    (b) If $\|\cdot\|$ is the 2-norm or Frobenius Norm, we can compute these values using the SVD. For example:

$$\kappa_2(A) = \|A\|_2 \left\|A^\dagger\right\|_2 = \frac{\sigma_1}{\sigma_{rank(A)}}$$

17. Solving Total Least Squares Problems Problem: Find $\arg\min\{\|E\|_F^2 + \|r\|_2^2 : (A+E)x = b + r$.

    Solution. Let $C = [A\ b]$, $F = [E\ r]$ and $z^T = [x^T\ -1]$. Then

$$\arg\min\{\|F\|_F^2 : (C+F)z = 0\}$$

    Since $z \neq 0$, we want to find F so that $rank(C+F) < n$ and $\|F\|_F$ is minimized. If the SVD of $C = \sum_{i=1}^{n+1} \sigma_i u_i v_i^*$ then we can reduce the rank of $C+F$ by removing one of the singular values, but we choose the smallest singular values to minimize the norm of F. So $F = -\sigma_{n+1} u_{n+1} v_{n+1}^*$. Now we want to find z such that:

$$\left(\sum_{i=1}^{n} \sigma_i u_i v_i^*\right) z = 0$$

    A simple choice is v_{n+1}, but we need the last element, $v_{n+1,n+1} = -1$, so we let

$$z = \frac{-1}{v_{n+1,n+1}} v_{n+1}$$

    □

4 Rank Retaining Factorization

4.1 Theory

1. Rank Retaining Factorization

   Definition 4.1. *Let $A \in \mathbb{C}^{m\times n}$ have rank r. A rank retaining factorization (RRF) of $A = GH$ where*

   (a) $G \in \mathbb{C}^{m\times r}$ *and* $H \in \mathbb{C}^{r\times n}$

   (b) $rank(G) = rank(H) = r$

2. Properties

   (a) Non-Singularity

   Lemma 4.1. *If an RRF of A is GH and $rank(A) = r$ then G^*G and HH^* are non-singular.*

   Proof. Let $U\Sigma V^* = G$ be its SVD. Note that $\Sigma^* = \begin{bmatrix} \Sigma_r & O \end{bmatrix}$ where Σ_r is an $r \times r$ diagonal matrix of full rank. Then $G^*G = V\Sigma_r^2 V^*$, which is the SVD of G^*G. Hence $rank(G^*G) = r$. □

   (b) Kernel. Co-kernel. Image. Co-image.

   Lemma 4.2. *Let the RRF of $A = GH$ and $rank(A) = r$. Then:*

   i. $Im(A) = Im(G)$
   ii. $\ker(A) = \ker(H)$
   iii. $Im(A^*) = Im(H^*)$
   iv. $\ker(A^*) = \ker(G^*)$

   Proof. If $A = GH$ is a RRF for A then $A^* = H^*G^*$ is a RRF for A^*. So we need only prove the first two points. If $y \in Im(A)$, $\exists x$ s.t. $Ax = y$. Let $z = Hx$. Then $Gz = y$. So $y \in Im(G)$. If $y \in Im(G)$ then $\exists z$ such that $Gz = y$. $H : \mathbb{C}^n \to \mathbb{C}^r$ is onto since it has rank r. Therefore, $\exists x$ such that $Hx = z$. So $Im(A) = Im(G)$. For the kernel, we note that $nullity(G) = 0$ by the rank nullity theorem. Therefore, only $G0 = 0$. Hence, $\ker(A) = \ker(H)$. □

4.2 Applications

1. Suppose we want to solve $Ax = b$ where the system is consistent and $A \in \mathbb{C}^{m\times n}$ has full column rank n.

   Solution. By the Fredholm alternative, $\mathbb{C}^n = \ker(A) \oplus im(A^*)$. Therefore, $x = x_0 + x_1$ where $x_0 \in \ker(A)$ and $x_1 \in im(A^*)$. Therefore:

   $$b = Ax = GH(x_0 + x_1) = GHx_1$$

   since $\ker(A) = \ker(H)$. Since $x_1 \in H^*$ there is a z such that $H^*z = x_1$. So:

   $$b = GHH^*z$$

Finally, $G \in \mathbb{C}^{n\times n}$ and has full rank, and so

$$(HH^*)^{-1}G^{-1}b = z$$

Multiplying through then:

$$H^*(HH^*)^{-1}G^{-1}b = x_1$$

□

2. Suppose now we have that A has full column rank and we want to find:

$$\arg\min\{\|x\|_2^2 : \|Ax-b\|_2 \leq \|Az-b\|_2^2 \ \forall z\}$$

Solution. Since $b \in \mathbb{C}^m$, by Fredholm's alternative, $b = b_0 + b_1$ where $b_0 \in \ker(A^*)$ and $b_1 \in im(A)$. Since $\ker(A^*) = \ker(G^*)$ and since $Ax = b_1$ is consistent, $G^*Ax = G^*b_1 = G^*b$ is consistent. Moreover, this leaves that $Hx = (G^*G)^{-1}G^*b$. Now we can split x as we did above to see that

$$x = H^*(HH^*)^{-1}(G^*G)^{-1}G^*b$$

□

5 QR & Complete Orthogonal Factorization

5.1 Theory

1. Gram-Schmidt Orthogonalization

Lemma 5.1. *Let $A \in \mathbb{C}^{m\times n}$. Then there $\exists Q \in U(m)$ and $R \in \mathbb{C}^{m\times n}$ such that R is upper triangular.*

Proof. Suppose $rank(A) = s$. Let q_i be the columns of a matrix Q which we will construct, and r_{ij} be the entries of a matrix R which we will construct. Let $R = rank(A)$. First, let a_1 be the first column of A and let $r_{11} = \|a_1\|$, and

$$q_1 = \frac{1}{r_{11}} a_1$$

Let P_2 be a permutation matrix such that AP_2 has first column a_1 and second column a_2 that is linearly independent of the first column. Such a permutation exists as long as $s \geq 2$. Let $j \leq s$ and let P_j be a permutation matrix such that $AP_2P_3 \cdots P_j$'s first $j-1$ columns are the same as $AP_2 \cdots P_{j-1}$ and its j^{th} column is linearly independent of the first $j-1$ columns. We can continue this process up to $AP_2 \cdots P_s$. Let the columns of this matrix be $a_1, \ldots, a_n$. Then for the first s columns:

$$q_j = \frac{a_j - \sum_{i=1}^{j-1} r_{ij} q_i}{r_{jj}}$$

where

$$r_{ij} = \langle a_j, q_i \rangle$$

$$r_{jj} = \left\| a_j - \sum_{i=1}^{j-1} r_{ij} q_i \right\|$$

Note that $a_{s+1}, \ldots, a_n \in span\{q_1, \ldots, q_s\}$. Hence, we can finish population R by finding the coefficients for these columns in a matrix S. Unfortunately, we have only computed $A\Pi = Q'M'$ where $Q' \in \mathbb{C}^{m\times s}$ and $[R\ S] = M' \in \mathbb{C}^{s\times n}$ (where Π is a product of permutation matrices so its inverse is Π^T). To get the remaining columns, we can complete the basis with the vectors in Q' of $\mathbb{R}^m$ and make them orthogonal by this process. To complete M' we need only add rows of zeros until the right dimension is achieved. Hence:

$$A\Pi = \begin{bmatrix} Q' & Q_{s+1,\cdots,m} \end{bmatrix} \begin{bmatrix} R' & S \\ O & O \end{bmatrix}$$

□

2. Versions of QR Factorization

   (a) Full QR Decomposition. This is the version stated in the Lemma. Note when $rank(A) < m \wedge n$ then Q is not unique.

(b) Reduced QR Factorization. This version is simply $Q'[R'\ S]$ calculated in the proof. It not unique given that the permutations can occur in several ways.

(c) Complete Orthogonal Factorization. Consider $[R'\ S]^T$, which has full column rank. Then it has a QR decomposition

$$Z \begin{bmatrix} U \\ O \end{bmatrix}$$

. The complete orthogonal factorization is then:

$$A = Q \begin{bmatrix} R' & S \\ O & O \end{bmatrix} \Pi^T = \begin{bmatrix} U^T & O \\ O & O \end{bmatrix} Z^T \Pi^T = Q \begin{bmatrix} L & O \\ O & O \end{bmatrix} \Omega^T$$

where L is lower triangular.

5.2 Applications

1. **Full Rank Least Squares.** Suppose $A \in \mathbb{C}^{m \times n}$ has full column rank and $n \leq m$. Find

$$\arg\min\{\|Ax - b\|_2\}$$

Solution 1. The solution is unique since $x = (A^*A)^{-1}A^Tb$ is the analytic solution. We use the full QR decomposition. So we have that:

$$\|Ax - b\|_2^2 = \left\| \begin{bmatrix} R \\ O \end{bmatrix} x - Q^*b \right\|$$

We can partition Q^*b into c and d so that:

$$\|Ax - b\|_2^2 = \|Rx - c\|_2^2 + \|d\|_2^2$$

Therefore, $x = R^{-1}c$ which we can computed by back solve (back substitution). □

Solution 2. Now we solve the normal equations $A^*Ax = A^*b$, which comes from taking the derivative of $\|Ax - b\|_2^2$. In this case we can do QR on A^*A and since A^*A is of full rank, $Rx = Q^*A * b$ can be solved by back-solve. Note that, $\kappa_2(A^*A) = \kappa_2(A)^2$ and so it is less numerically stable, but also $A^*A \in \mathbb{C}^{n \times n}$. So this method may be beneficial if $n \ll m$. □

2. **Least Squares with Linear Contraints**. Find $\arg\min \|Ax - b\|_2^2$: $C^Tx = d$.

Solution 1. Note that the Lagrangian is:

$$\mathcal{L}(\lambda, x) = \|b - Ax\|_2^2 + 2\lambda^T(C^Tx - d)$$

Differentiating returns the system:

$$\begin{cases} -2A^T(b - Ax) + 2C\lambda = 0 \\ C^Tx = d \end{cases}$$

The first equation can be re-written as $A^TAx + C\lambda = A^Tb$. Writing this as an augmented system:

$$\begin{bmatrix} A^TA & C \\ C^T & O \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A^Tb \\ d \end{bmatrix}$$

If there is sparsity in A and C, this system can be solved with sparse methods. □

Solution 2. Let $QR = C$. Let $y = Q^Tx$ and partition y into u and v. Then we can solve $Ru = d_1$, and solve for u. Using this, we can do a normal minimization of $\|b - AQy\| = \left\|b - \tilde{A}_1u - \tilde{A}_2v\right\|$, where only v is unknown. □

5.3 Givens Rotations

1. A Givens Rotation, $G^{(i,j)}$ is a unitary matrix of the form:

$$G^{(i,j)}_{lk} = \begin{cases} 1 & l = k \notin \{i,j\} \\ \lambda & l = k \in \{i,j\} \\ \sigma & l = i, k = j \\ -\sigma & l = j, i = k \\ 0 & \text{otherwise} \end{cases}$$

   where $\sigma^2 + \lambda^2 = 1$

2. If σ and λ are selected correctly for a vector v, $(G^{(i,j)}v)_i$ or $(G^{(i,j)}v)_j$ can be set to 0.

   Example 5.1. *Suppose $i < j$. We can find λ and σ as follows if we want to set $(G^{(i,j)}v)_j = 0$. Note that:*

$$(G^{(i,j)}v)_k = \begin{cases} v_k & k \notin \{i,j\} \\ \lambda v_i + \sigma v_j & k = i \\ -\sigma v_i + \lambda v_j & k = j \end{cases}$$

   Hence, we need to solve for σ and λ which satisfy:

$$\begin{cases} 0 & = \lambda v_j - \sigma v_i \\ 1 & = \lambda^2 + \sigma^2 \end{cases}$$

   Taking the positive options:

$$\lambda = \frac{v_i}{\sqrt{v_i^2 + v_j^2}} \qquad \sigma = \frac{v_j}{\sqrt{v_i^2 + v_j^2}}$$

3. For a matrix A, we can easily find $G_1, \ldots, G_n$ such that $G_n \cdots G_1 A = R$ where R is upper triangular. And, letting $Q^T = G_n \cdots G_1$, we have the QR decomposition of A.

4. Givens Rotations are beneficial when A is sparse.

5. Pivoting (Partial or Complete) can be implemented to ensure that the element on the diagonal of A is the largest in the row and column

5.4 Householder Reflections

1. A Householder reflection, H, is a matrix of the form $I + \tau vv^T$ where $\|v\| = 1$ and $\tau \in \mathbb{C}$. It reflects any vector over the line $\{tv : t \in \mathbb{R}\}$.

2. For use in QR decomposition, we require that $HH^T = 1$ and $H^T H = 1$. Since $H^T = H$, we need to check only one:

$$\begin{aligned} H^T H &= I + 2\tau vv^T + \tau^2 \|v\|^2 vv^T \\ &= I + \tau vv^T(2 + \tau \|v\|^2) \\ &= I + \tau vv^T(2 + \tau) \end{aligned}$$

   This implies $\tau = -2$

3. Moreover, given a vector v, we want $Hz = \alpha e_1$. Since H causes a reflection, we choose

$$v = \frac{z + \alpha e_1}{\|z + \alpha e_1\|}$$

   Substituting this in, we have that:

$$(I - 2vv^T)z = z - 2\frac{(z + \alpha e_1)(\|z\|^2 + \alpha w_1)}{\|z + \alpha e_1\|^2} = \alpha e_1$$

   which holds if $2(\|z\| + \alpha w_1) = \|z\|^2 + 2\alpha w_1 + \alpha^2$. This implies $\pm \|w\| = \alpha$.

4. Taking the negative case, we have that the appropriate Householder Reflection is:

$$I - 2\frac{(z - \|z\| e_1)(z - \|z\| e_1)^T}{\|z - \|z\| e_1\|^2}$$

5. Letting $A_p = \begin{bmatrix} a_p & \cdots & a_{mp} \end{bmatrix}^T$ of a matrix $A \in \mathbb{C}^{m \times n}$ and permutation matrices Π:

   (a) Let $A^{(1)} \in \mathbb{C}^{m \times n}$ be the matrix we want to QR factorize

   (b) Let H_1 be the householder matrix such that $H_1 A_1^{(1)} = \left\| A_1^{(1)} \right\| e_1 \in \mathbb{C}^n$. Let $A^{(2)} = H_1 A^{(1)}$.

   (c) Let $\tilde{H}_2$ be the householder matrix such that $\tilde{H}_2 A_2^{(2)} = \left\| A_2^{(1)} \right\| e_1 \in \mathbb{C}^{n-1}$. Let

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H}_2 \end{bmatrix}$$

   and $A^{(3)} = H_2 A^{(2)}$

   (d) Then, $H_n \cdots H_1 A^{(1)} = R$ an upper right triangular matrix. Letting $Q^T = H_n \cdots H_1$ gives us Q.

   (e) If necessary, we can pivot the rows to achieve non zeros along the diagonals.

6 LU, LDU, Cholesky and LDL Decompositions

1. Both LU and LDU Factorization are based on Gaussian Elimination.

2. Given a system of equation $Ax = b$, we add multiples of the first row to the subsequent rows to set them equal to zero.

   Example 6.1. *Let* $v = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T$. *To set all* $v_i = 0$ *for* $i \neq 1$, *we multiply by:*

   $$L = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -v_2/v_1 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ -v_n/v_1 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

3. This requires that the diagonal elements of a matrix are non-zero. Hence, a sequence of pivots (partial or complete) resulting in:

   $$M_n\Pi_n M_{n-1}\Pi_{n-1}\cdots M_1\Pi_1 A = U \qquad M_n\Pi_n M_{n-1}\Pi_{n-1}\cdots M_1\Pi_1 A\Pi_0 = U$$

   where U is upper triangular.

4. To recover A or at least $A\Pi_0$ (in the partial case at least):

   $$\begin{aligned} A &= (M_n\Pi_n M_{n-1}\Pi_{n-1}\cdots M_1\Pi_1)^{-1}U \\ &= \Pi_1^T M_1^{-1}\Pi_2^T\cdots\Pi_n^T M_n^{-1}U \\ &= \Pi_1^T\Pi_2^T(\Pi_2 M_1^{-1}\Pi_2^T)\cdots\Pi_n^T M_n^{-1}U \\ &= \Pi_1\Pi_2^T\cdots\Pi_n^T L_1\cdots L_n U \end{aligned}$$

5. If A has non-singular principal submatrices, then LDU can be recovered by performing LU decomposition on $U^T = (U')^T D$

6. If A is symmetric, LDL^T is recovered.

7. It is difficult to check if A has non-singular principal submatrices (unless A is positive definite)

8. Suppose A is symmetric, positive definite. Consider L to be lower triangular, we want to find $A = LL^T$, and so we can compute the terms of L directly from this relationship.

9. Similarly, we can find the algorithm form $A = LDL^T$ if A is symmetric positive definite.

7 Iterative Methods

7.1 Overview

1. Iterative methods can be used to compute solutions to linear systems, least squares, eigenvalue problems, and singular value problems
2. Suppose we want to solve $Ax = b$. Iterative methods compute a sequence x_k where $x_k \to x = A^{-1}b$, and we can control the accuracy to stop the process
3. Classes of Iterative Methods
   (a) Splitting/One-Step Stationary Methods
      i. Split $A = M - N$ where $Mx = c$ is easy to solve for some c.
      ii. Solve $Mx_k = Nx_{k-1} + b$
   (b) Semi-Iterative Methods
      i. Generate, for a choice of B, $y_k = By_{k-1} + c$
      ii. Then $x_k = \sum_{j=0}^{k} a_{jk} y_j$
   (c) Krylov Subspace Methods
      i. Find iterates $x_k \in \{b, Ab, A^2b, \cdots, A^{k-1}b\}$
      ii. Uses the fact that eventually, $\mathcal{K}_r = \{b, Ab, A^2b, \cdots, A^{r-1}b\}$ will be linearly dependent as r increases.

7.2 Splitting Methods

1. Overview
   (a) Strategy: Suppose A is invertible in $Ax = b$. We find M such that $Mx = c$ for some c is easy to solve and let $N = M - A$. We then have the following iteration scheme:

$$Mx_k = Nx_{k-1} + b$$

   (b) General Convergence of Errors

   Proposition 7.1. *Let $e_k = x - x_k$ where x solves $Ax = b$. $\|e_k\| \to 0$ for any initial x_0 if and only if $\rho(M^{-1}N) < 1$.*

   Proof. Note that:

$$x_k = M^{-1}Nx_{k-1} + M^{-1}b$$

   Since $x = M^{-1}Nx + M^{-1}b$ (since $(M - N)x = Ax = b$):

$$e_k = M^{-1}Ne_{k-1} =: Be_{k-1}$$

   Therefore, $e_k = B^k e_0$. By **Lemma 2.5**, both directions follow. □

2. Jacobi Method

(a) $M = diag(A)$. As long as the diagonal elements of A are non-zero, then iterates can be explicitly computed as:

$$x_k^i = \frac{1}{a_{ii}} \left[b^i - \sum_{j \neq i} a_{ij} x_{k-1}^j \right]$$

(b) Jacobi Convergence of Errors

Corollary 7.1. *A is strictly diagonally dominant then $e_k \to 0$.*

Proof. We need only show that $\rho(M^{-1}N) < 1$. This matrix is:

$$M^{-1}N = \begin{bmatrix} 0 & -a_{12}/a_{11} & \cdots & -a_{1n}/a_{11} \\ -a_{21}/a_{22} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1}/a_{nn} & \cdots & -a_{n,n-1}/a_{nn} & 0 \end{bmatrix}$$

Suppose A is strictly diagonally dominant, then for any row:

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|$$

Using the fact that $\rho(A) \leq \|A\|_\infty$, we then have that

$$\rho(M^{-1}N) \leq \left\|M^{-1}N\right\|_\infty < 1$$

Applying **Proposition 7.1**, the result follows. □

3. Gauss-Seidel Method

(a) Notice that in the Jacobi Method, we can compute $x_k^1, \ldots, x_k^{i-1}$ before we compute x_k^i. Gauss-Seidel uses these updated values to compute x_k. This yields:

$$x_k^i = \frac{1}{a_{ii}} \left[b^i - \sum_{j<i} a_{ij} x_k^j - \sum_{j>i} a_{ij} x_{k-1}^j \right]$$

(b) Gauss-Seidel Convergence of Errors

Corollary 7.2. *If A is strictly diagonally dominant then $e_k \to 0$.*

Proof. Let L be the sub-diagonal entries of A and U be the super diagonal entries of A. Then:

$$Mx_k = b - Lx_k - Ux_{k-1}$$

Therefore, $x_k = -(M+L)^{-1}Ux_{k-1} + (M+L)^{-1}b$. Hence, the errors are:

$$e_k = -(M+L)^{-1}Ue_{k-1}$$

or explicitly:

$$e_k^i = -\frac{1}{a_i i}\left[\sum_{j<i} a_{ij}e_k^j + \sum_{j>i} a_{ij}e_{k-1}^j\right]$$

Let $r_i = \sum_{j\neq i}\frac{|a_{ij}|}{|a_{ii}|}$. By diagonal dominance, $\max_i r_i =: r < 1$. We now proceed inductively to show that $|e_k^i| \leq r\left\|e_{k-1}\right\|_\infty$. When $i = 1$:

$$|e_k^1| \leq r\left\|e_{k-1}\right\|_\infty$$

Suppose this holds up to $i-1$. Then:

$$\begin{aligned}
|e_k^i| &\leq \sum_{j<i}\frac{|a_{ij}|}{|a_{ii}|}|e_k^j| + \sum_{j>i}\frac{|a_{ij}|}{|a_{ii}|}|e_{k-1}^j| \\
&\leq r\left\|e_{k-1}\right\|_\infty \sum_{j<i}\frac{|a_{ij}|}{|a_{ii}|} + \left\|e_{k-1}\right\|_\infty \sum_{j>1}\frac{|a_{ij}|}{|a_{ii}|} \\
&\leq r\left\|e_{k-1}\right\|_\infty
\end{aligned}$$

Therefore, $\left\|e_k\right\|_\infty \leq r^k\left\|e_0\right\|_\infty \to 0$ as $k \to \infty$. □

(c) Number of iterations

i. From the proof of the corollary, we have that $\|e_k\| \leq r^k\|e_0\|$.

ii. Given a tolerance $\epsilon > 0$, and $x_0 = 0$, we need only

$$k = \left\lceil \frac{\log\epsilon}{\log r}\right\rceil$$

to bound the relative error by ϵ.

4. Successive Over Relaxation (SOR)

(a) SOR is a generalization of Gauss-Seidel. Let $A = D - L - U$ where D is diagonal, L has only non-zero subdiagonal and U has only non-zero superdiagonals. The iteration is derived as follows:

$$\begin{aligned}
\omega Dx &= \omega b - \omega Lx - \omega Ux \\
Dx &= \omega b - \omega Lx - \omega Ux + (1-\omega)Dx
\end{aligned}$$

Then, using the most recent estimates to compute the iteration as in Gauss-Seidel:

$$Dx_k = \omega b - \omega Lx_k - \omega Ux_{k-1} + (1-\omega)Dx_{k-1}$$

and, explicitly:

$$x_k^i = (1-\omega)x_{k-1}^i + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j<i} a_{ij}x_k^j - \sum_{j>i} a_{ij}x_k^j\right]$$

(b) When $\omega = 1$ this is the Gauss-Seidel method, and when $\omega > 1$ this is Successive Over Relaxation

(c) Ostrowki's Lemma:

Lemma 7.1. *Suppose A is symmetric positive definite. $\omega \in (0,2)$ if and only if $e_k \to 0$.*

7.3 Semi-Iterative Methods

1. Richardson's Method

   (a) Richardson's Method is an unstable numerical method with a parameter α which updates iterates using:

   $$x_k = (I - \alpha A)x_{k-1} + \alpha b$$

   Note 7.1. *Effectively, Richardson's Method is a line search method for minimizing $\frac{1}{2}x^TAx - x^Tb$. Given a starting point x_{k-1}, then the direction of steepest descent is $b - Ax_{k-1}$ (by taking derivatives). Therefore,*

   $$x_k = x_{k-1} + \alpha(b - Ax_{k-1})$$

   where α is the step length parameter.

   (b) Richardson Convergence of Errors

   Corollary 7.3. *Suppose A is symmetric positive definite. $e_k \to 0$ if and only if $0 < \alpha < \frac{1}{\mu_{\min}}$ where $\mu_{\min}$ is the smallest eigenvalue of A. Moreover, $e_k \to 0$ optimally if $\frac{2}{\mu_{\min}+\mu_{\max}}$.*

   Proof. We have that the errors are:

   $$e_k = (I - \alpha A)e_{k-1}$$

   Hence, from **Proposition 7.1**, $e_k \to 0$ if and only if $\rho(I - \alpha A) < 1$. Since A is symmetric positive definite, its eigenvalues are positive. Letting μ be a vector of the eigenvalues and $\mu_{\max}$ be the largest and $\mu_{\min}$ be the smallest. Letting $A = Xdiag(\mu)X^*$ be the EVD of A, we have that the EVD of $I - \alpha A$ is:

   $$X(I - \alpha diag(\mu))X^*$$

   Hence, $0 < \rho(I - \alpha A) = 1 - \alpha\mu_{\min} < 1$. This implies the first result.

   For optimality, note that the SVD and EVD of A coincide. Hence, $\|I - \alpha A\|_2 = \max|1 - \alpha\mu_i| = \|\mathbf{1} - \alpha\mu\|_\infty$. We want to minimize this over α, which gives us the result. □

2. Steepest Descent

   (a) This method is akin to Richardson's method, except at every step, we optimize α_k so that we minimize the norm of $b - Ax_{k+1}$ (i.e. we bring the slope closer to zero and hence closer to the stationary point of $0.5x^TAx + b^Tx$).

   (b) Optimal Choice of α_k

   Lemma 7.2. *Letting $r_k = b - Ax_k$, the optimal choice of α_k to minimize r_{k+1} is:*

   $$\alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}$$

Proof. Note that norms are equivalent in finite dimensional space. Hence, we can minimize the $\|\cdot\|_2$, which, in the computation, will require determining A^2, or we can minimize $\|\cdot\|_{A^{-1}}$ which overcomes this cost. First:

$$r_{k+1} = b - Ax_{k+1} = b - A(x_k + \alpha_k r_k) = r_k - \alpha_k A r_k$$

Second:

$$r_{k+1}^T A^{-1} r_{k+1} = r_k^T r_k - 2\alpha r_k^T r_k + \alpha^2 r_k^T A r_k$$

Taking derivatives and noting that the quadratic coefficient is positive, we have the result. □

3. Chebyshev's Method

   (a) Notice that in the Steepest Descent Method:

   $$e_k = (I - \alpha_k A)(I - \alpha_{k-1} A) \cdots (I - \alpha_0) e_0 =: P_k(A) e_0$$

   Instead of optimizing over α_k stepwise, Chebyshev's method tries to optimize over all $\alpha_0, \ldots, \alpha_k$ at each k.

   (b) Since $\|e_k\| \leq \|P_k(A)\| \, \|e_0\|$ we want to minimize $\|P_k(A)\|$

   (c) This is solved using Chebyshev's Polynomials.

7.4 Krylov Space Methods

1. Overview

   (a) Suppose we want to solve $Ax = b$ and A is invertible. Hence, $x \in im(A^{-1})$. Computing the inverse is expensive, but we can more easily compute $A^k c$ for some c implicitly.

   (b) Krylov Subspace

   Lemma 7.3. *There is an r such that the solution to $Ax = b$, when A is invertible, is in $\mathcal{K}_r(A)$.*

   Proof. From the Cayley-Hamilton theorem, there is a minimal polynomial of degree r,

   $$P_r(x) = \sum_{i=0}^{r} \alpha_i x^i$$

   such that $P_r(A) = 0$. Hence:

   $$A^{-1} = \frac{1}{\alpha_0} \sum_{i=1}^{r} \alpha_i A^{i-1}$$

   Therefore, $A^{-1} b \in span\{b, Ab, Ab^2, \ldots, Ab^{r-1}\}$. □

2. Conjugate Gradients

(a) Our goal is to minimize $\frac{1}{2}x^TAx - b^Tx$, or, equivalently, solve $Ax = b$ when $A \in \mathbb{R}^{n\times n}$ is symmetric, positive definite.

(b) Conjugacy.

i. Conjugated Vectors.

Definition 7.1. *Let $\langle u, v\rangle_A = u^TAv$ and $\|v\|_A^2 = \langle v, v\rangle_A$. A set of vectors $\{p_1, \ldots, p_r\} \subset \mathbb{R}^n$ are Conjugated with respect to A if for all $i \neq j$:*

$$\langle p_i, p_j\rangle = 0$$

ii. Conjugated Vectors form a basis.

Lemma 7.4. *Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $p_1, \ldots, p_r$ are conjugated with respect to A. Then $p_1, \ldots, p_r$ are linearly independent.*

Proof. Suppose this is not true. Then there are $\alpha_1, \ldots, \alpha_r$ not all zero such that:

$$\alpha_1 p_1 + \cdots + \alpha_r p_r = 0$$

Then:

$$\begin{aligned} 0 &= (\alpha_1 p_1 + \cdots + \alpha_r p_r)^T A(\alpha_1 p_1 + \cdots + \alpha_r p_r) \\ &= \alpha_1^2 p_1^T Ap_1 + \cdots + \alpha_r^2 p_r^T Ap_r \end{aligned}$$

Since $\alpha_i^2 \geq 0$ and $p_i^T Ap_i > 0$, we have a contradiction. □

(c) Conjugated Gradients

i. Let $x_0, \ldots, x_k$ be a sequence of iterates

ii. Their steepest descent directions are $r_0 = b - Ax_0, \ldots, r_k = b - Ax_k$

iii. We then create vectors $p_0, \ldots, p_k$ which are conjugated with respect to A from the descent directions $r_0, \ldots, r_k$ using a Gram-Schmidt type approach:

$$p_k = r_k - \sum_{j<k} \frac{\langle r_k, p_j\rangle_A}{\|p_j\|_A^2} p_j$$

iv. We update $x_{k+1} = x_k + \alpha_k p_k$ where α_k is the

$$\arg\min \frac{1}{2}x_{k+1}^T Ax_{k+1} - b^T x_{k+1}$$

(d) Relation to Krylov Spaces: The new axes system is an orthonormal basis for the Krylov Subspace

(e) Geometric Interpretation: we rotate/dilate the axes system with respect to A and b, and each iterate minimizes along each "coordinate" of this new axis system. Since there are n coordinates, this requires at most n iterates.