1 Function Approximations

Consider learning the approximation of a function f in an interval $[a, b]$. One approach is data fitting or also called optimized regression. Here the values of the function f_i are given only at finitely many points $x_i \epsilon [a, b]$, $i = 1, ..., m$. The input data pairs are (x_i, f_i). For a best solution to the regression, one chooses a model or a parametric form of the corresponding approximation function $F(a, x)$, with an unknown vector of parameters $a = (a_1, a_2, \dots, a_n)^T$, and then sets up an objective criterion for the optimal quality of approximation, i.e. minimizing a distance function $d(f(x), F(a, x))$ in a metric space, or minimizing a norm $\|F(a) - f\|$ in a normed linear space. First we shall consider various optimizing criteria, and then several methods for solving such optimized approximations.

1.1 Definition of Vector Spaces

Definition. *Formally, a field is a set $\mathbb{F}$ together with two operations called addition and multiplication, $\mathbb{F}$ with addition forms an Abelian group with identity element "0" while $\mathbb{F}$ with multiplication forms an Abelian group with identity element "1". An Abelian group is commutative and generalizes arithmetic on integers.*

Definition. *A vector space is defined as $V = \{X, +, *, \mathbb{F}\}$, where X is a set and $\mathbb{F}$ is a Field. X and $+ : X \times X \to X$ forms an Abelian group and $* : \mathbb{F} \times X \to X$ satisfying the following:*

- $\alpha * (a + b) = \alpha * a + \alpha * b$
- $(\alpha + \beta) * a = \alpha * a + \beta * a$
- $\alpha * (\beta * a) = (\alpha\beta)a$
- $1 * a = a$
- $0 * a = 0$

where $\alpha, \beta \in \mathbb{F}$, $a, b \in X$

Definition. *Let V be a (linear) vector space over $\boldsymbol{F}$. A set of vectors $v_1, ..., v_n \subset V$ is said to be **linearly dependent** if there are scalars $a_1, ..., a_n \in F$, not all zero, such that*

$$\sum_{j=1}^{n} a_j v_j = 0$$

*If there are no such scalars, the set $v_1, ..., v_n$ is said to be **linearly independent**.*

The set X of polynomials $p(x) = a_n x^n + ... + a_1 x + a_0$ of degree $\leq n$ and having *rational* coefficients is a vector space over the field $\mathbf{Q}$ of rational numbers.

1.2 Norms and Normed Spaces

A linear vector space V equipped with a norm $\|.\|$ is called a normed linear space.

Definition. *V is a vector space, $N : V \to \mathbb{R}$ is a **norm** of V if :*

- $N(v) \geqslant 0$, *and* $N(v) = 0$ *if and only if* $v = 0$
- $N(\alpha v) = |\alpha| N(v)$
- $N(u + v) \leqslant N(u) + N(v)$

We always denote $\|u\| := N(u)$.

remark. You can verify that $\|u - v\|$ is a metric on V, thus every normed space is a metric space.

Here are some examples of norms.

- The L_q $(1 \leq q < \infty)$ norm for a vector $\mathbf{v} \in R^p$ is given by $\|\mathbf{v}\| = (\sum_{i=1..p} |\mathbf{v}|^q)^{\frac{1}{q}}$. In particular, we have the least absolute deviation L_1 norm : $\|x\|_1 := \sum_{i=1}^{n} |x_i|$, and the least squares deviation L_2 norm : $\|x\|_2 := \sum_{i=1}^{n} |x_i|^2$
- The L_∞ norm can be defined as $\|x\|_\infty := \max_i |x_i|$; Similarly $\|x\|_{-\infty} := \min_i |x_i|$
- For metric spaces the norm of an element $\mathbf{v}$ is just the distance of $\mathbf{v}$ from the null vector 0.
- Define $supp(x) := \{i \mid x_i \neq 0, x \in \mathbb{R}^n\}$, an vector $\mathbf{x}$ is **s-sparse** if $|supp(x)| \leqslant s$. Denote $\|x\|_0 = |supp(x)|$ as L_0 norm.
- Balls with respect to norms/seminorm are defined as $B_q(r) = \{\mathbf{v} \in R^p \mid \|\mathbf{v}\|_q \leq r\}$. Note, $B_0(s) = \{$set of s-sparse vectors $\}$. See also figure below.

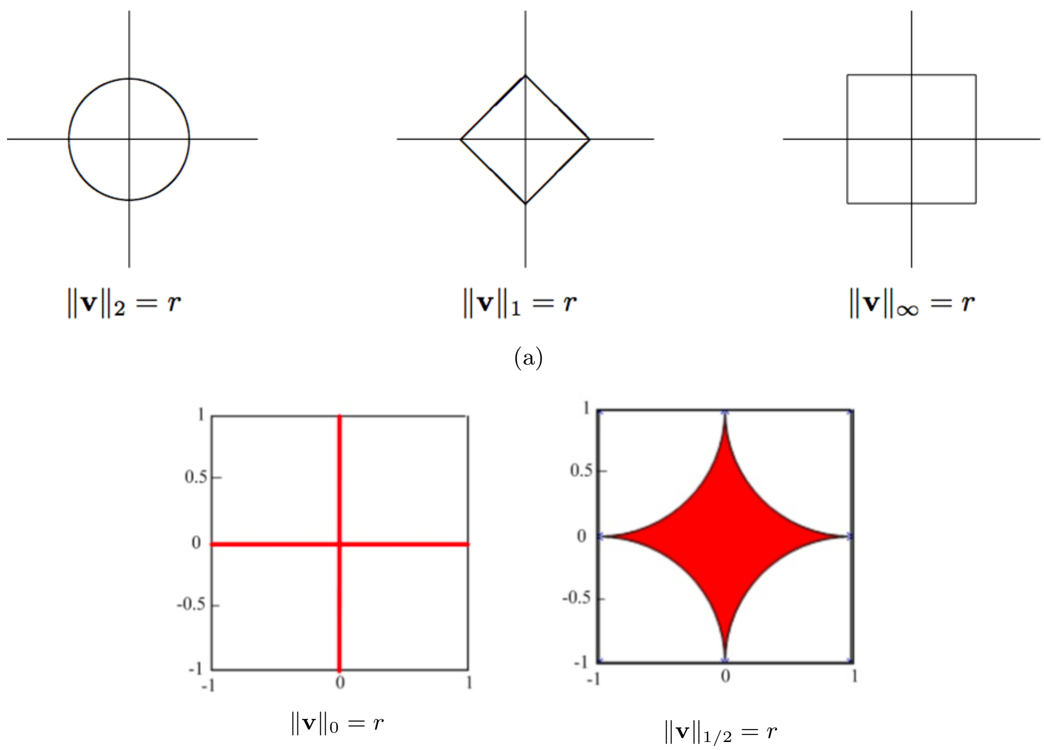

Figure 1: B_q with different value of q in $2D$, note when $q = 1/2$,$\| \cdot \|_q$ is not a norm (check by yourself!)

1.3 Matrix Norms

Starting with an arbitrary but fixed norm $\|.\|$ on vectors, we will define a related norm on matrices, so that the following property holds: whenever

$$b = Ax, \quad \text{then } \|b\| \leq \|A\|\|x\| \tag{1}$$

Here A is a matrix, $\|A\|$ is its norm and b and a are arbitrary vectors related only by $Ax = b$. Let us think of $\|A\|$ as some measure of the size of measure of the size of the matrix A.
Consider the equation $A^{-1}(b - b') = (x - x')$. Apply (1), one obtain:

$$\|x - x'\| \leq \|A^{-1}\|\|b - b'\| \tag{2}$$

Consider the equation (1):

$$\|b\| \leq \|A\|\|x\| \tag{3}$$

or equivalently,

$$\frac{1}{\|x\|} \leq \|A\| \frac{1}{\|b\|} \tag{4}$$

Multiply inequality (2) by inequality (4) to get:

$$\frac{\|x - x'\|}{\|x\|} \leq \|A^{-1}\|\|A\| \frac{\|b - b'\|}{\|b\|} \tag{5}$$

Definition The number

$$cond(A) = \|A^{-1}\|\|A\|$$

is called the condition number of the matrix A (in the norm $\|.\|$)
The condition number plays a fundamental role in the analysis of algorithms for the solution of linear systems. If $\|A^{-1}\|\|A\|$ is small then the solution $Ax = b$ is considered to be stable under perturbations of the right side: that is, the content of (5).

Lemma. *Let $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$. Let $C \in \mathbb{R}^{n \times n}$ and $d \in \mathbb{R}^n$ be small perturbations to A and b respectively. Then the error on the solution of the perturbed system*

$$(A + \epsilon C)x(\epsilon) = (b + \epsilon d)$$

is given by

$$\frac{\|x(\epsilon) - x\|}{\|x\|} \leq cond(A)(\frac{|\epsilon|\|C\|}{\|A\|} + \frac{|\epsilon|\|d\|}{\|x\|}) + O(\epsilon^2)$$

where $x = x(0)$.

Proof. First differentiating the linear system:

$$Cx(\epsilon) + (A + \epsilon C)\frac{dx(\epsilon)}{d\epsilon} = d$$

This yields:

$$\frac{dx(\epsilon)}{d\epsilon}\Big|_{\epsilon \to 0} = A^{-1}(d - Cx)$$

Next, apply Taylor's expansion around $\epsilon = 0$.

$$x(\epsilon) = x + \epsilon \frac{dx(\epsilon)}{d\epsilon}\Big|_{\epsilon \to 0} + O(\epsilon^2) = x + \epsilon A^{-1}(d - Cx) + O(\epsilon^2)$$

Hence

$$\frac{\|x(\epsilon) - x\|}{\|x\|} \leq |\epsilon|\|A^{-1}\|(\frac{\|d\|}{\|x\|} + \|C\|)$$

$$\frac{\|x(\epsilon) - x\|}{\|x\|} \leq cond(A)(\frac{|\epsilon|\|d\|}{\|b\|} + \frac{|\epsilon|\|C\|}{\|A\|})$$

$\square$

If on the other hand, $\|A^{-1}\|\|A\|$ is large (much bigger than 1), one should be prepared for the worst-case scenario. Except that the inequality in (5) is an equality for some choices of b, b': there maybe vectors b, b' with small relative difference $\|b - b'\|/\|b\|$ for which the the relative solution error $\|x - x'\|/\|x\|$ will be large,

$$\frac{\|x - x'\|}{\|x\|} = cond(A) \times \frac{\|b - b'\|}{\|b\|}$$

We will see that $\|A^{-1}\|\|A\|$ is always ≤ 1 "Small" therefore means: the condition number is not too much bigger than 1. How big can it be and still be small is a question left to numerical analysis.
The set of $n \times n$ is a linear space of dimension n^2 - the n^2 coordinates are simply arranged in a square, rather than in the customary single column. One could therefore define a norm on matrices by taking one of the standard l^p norms on $\mathbb{R}^n$ (or C^n). For a matrix A, define

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

(It is standard practice to use the same symbol, $\|.\|$, for the norms of A and x). The matrix norm $\|.\|$ is called the **matrix norm induced by the vector norm** $\|.\|$ on $\mathbb{R}^n$.

$$\|A\| = \sup(\frac{\|Ax\|}{\|x\|})$$

Consider $A = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$

The l^2 norm of A is the square root of the largest eigenvalue of $A^T A$. Since eigenvalues are difficult to compute except for 2×2 matrices - the norm $\|A\|_2$ can in general not be found explicitly. The norms $\|A\|_1$ and $\|A\|_\infty$ however can be read off from the entries of A. Hence it is more convenient to use these norms, and $\|A\|_2$ can be approximated from these if necessary.

Example. *Start with the l^∞ norm on $\mathbb{R}^2$. By definition,*

$$\begin{aligned}
\|A\|_\infty &= \sup_{\|x\|_\infty=1} \|Ax\|_\infty \\
\|x\|_\infty &= 1 \ \textit{means} \ \max\{|x_1|, |x_2|\} = 1 \\
\|Ax\|_\infty &= \max\{|Ax_1|, |Ax_2|\} \\
&= \max\{|x_1 + 2x_2|, |x_1|\}
\end{aligned}$$

Maximization of a function of two variables subject to constraints is a standard problem in the calculus of several variables: it is usually solved by the method of Lagrange Multipliers (Which will be introduced later).
The following exercises rely on the various properties of the appropriate matrix norm.

Exercise. *Let A be an $n \times n$ matrix. Show that*

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$$

Lemma. *The function $x \to \|Ax\|$ from $\mathbb{R}^n$ to [0, ∞) is continuous.*

Lemma. *A continuous function on the set $\{\|x\| = 1\} \subset \mathbb{R}^n$ is bounded.*

Proposition. *Let I be the $n \times n$ identity matrix, Let $\|.\|$ be a vector norm on $\mathbb{R}^n$ and denote the corresponding natural matrix norm by the same symbol. Then,*

- $\|I\| = 1$

- $\|Ax\| \leq \|A\|\|x\|$ *for all* A *and* x
- *if* A *and* B *are two* $n \times n$*matrices, then* $\|AB\| \leq \|A\|\|B\|$

Corollary. *The condition number cond(A)* $= \|A\|\|A^{-1}\|$ *of a matrix satisfies*

$$\|A\|\|A^{-1}\| \geq 1$$

Proof. Using the above Proposition 1.3, $\|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| = 1$ □

1.3.1 Eigenvalues and Singular Values

As an application of matrix norm, we will now examine eigenvalues and singular values of a matrix.

Let $A \in \mathbb{R}^{n \times n}$. An **eigenvalue** and an **eigenvector** of A is a pair $(\lambda, \mathbf{x})$ so that $A\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{x} \neq \mathbf{0}$.

$$\begin{aligned} & A\mathbf{x} - \lambda\mathbf{x} = 0 \\ \iff & (A - \lambda I)\mathbf{x} = 0 \\ \iff & \det(A - \lambda I) = 0 \end{aligned}$$

The characteristic polynomial $\det(A - \lambda I)$ of A is a degree n polynomial in λ.

$$\det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = (a_{11} - \lambda) \begin{vmatrix} a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} + \cdots$$

The Fundamental Theorem of Algebra guarantees that any real polynomial

$$p(\lambda) = \sum_{i=0}^{n} a_i \lambda^i \qquad (a_i \in \mathbb{R})$$

of degree n has exactly n (possibly complex) roots counting multiplicities.

Two matrices A and B are *similar* if $A = PBP^{-1}$ for some invertible P. If A and B are similar, then A and B have the same eigenvalues (but not necessarily the same eigenvectors).

A square matrix A is *diagonalizable* if A is similar to a diagonal matrix D. In other words, $A = PDP^{-1}$ for some P. A matrix $A \in \mathbb{R}^{n \times n}$ is diagonalizable iff it has n linearly independent eigenvectors. Columns of P are the eigenvectors of A and diagonal elements of A are the corresponding eigenvalues.

A matrix P is *orthogonal* if it is invertible and $P^{-1} = P^{\mathrm{T}}$.

A matrix A is *orthogonally diagonalizable* if there exists an orthogonal matrix P such that $P^{-1}AP = D$ is diagonal.

Theorem (Real Spectral Theorem [BHK, Theorem 12.7])**.** *Let* A *be a symmetric* $n \times n$ *matrix with real entries. Then,* A *has real eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *with corresponding eigenvectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$*. Furthermore,* A *is orthogonally diagonalizable*

$$A = VDV^{\mathrm{T}} = \sum_{i=1}^{m} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}},$$

where V *is the matrix of column vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ *and* D *is the diagonal matrix with* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *as diagonal entries.*

When A is not a square matrix, the spectral theorem for A to be an $n \times d$ matrix leads to the general singular value decomposition (SVD). Similar as Real Spectral Theorem, we will see that any matrix $\mathbb{R}^{m\times n}$ (w.l.o.g. $m \leq n$) can be written as:

$$A = \sum_{i=1}^{m} \sigma_i u_i v_i^T$$

where $\sigma_i \geq 0$, $\{u_i\}, \{v_i\}$ are orthonormal basis of $\mathbb{R}^m$,$\mathbb{R}^n$, respectively. How can we deduce the formula? Consider the matrix $AA^T \in \mathbb{R}^{m\times m}$. Let us set u_i to be the i-th eigenvector of AA^T. By definition of eigenvalue, we have $AA^T u_i = \lambda_i u_i$. Since AA^T is positive semidefinite matrix we have $\lambda_i \geq 0$. Since AA^T is symmetric we have

$$u_j^T AA^T u_i = \lambda_i u_j^T u_i, \quad (u_j^T AA^T u_i)^T = u_i^T AA^T u_j = \lambda_j u_i^T u_j$$

This implies $u_i^T u_j = \delta_{ij}$. Let $\sigma_i = \sqrt{\lambda_i}$ and $v_i = \frac{1}{\sigma_i} A^T u_i$. Now we can compute that

$$v_i^T v_j = \frac{1}{\sigma_i \sigma_j} u_i^T AA^T u_j = \frac{\lambda_j}{\sigma_i \sigma_j} u_i^T u_j = \delta_{ij}$$

We have constructed σ_i, u_i, v_i. We are only left to show that $A = \sum_{i=1}^m \sigma_l u_i v_i^T$. To achieve that we examine the norm of the difference induced by any test vector $w = \sum_{i=1}^m \alpha_i u_i$.

$$\begin{aligned}
\|w^T(A - \sum_{i=1}^{m} \sigma_i u_i u_i^T)\| &= \|(\sum_{i=1}^{m} \alpha_i u_i^T)(A - \sum_{i=1}^{m} \sigma_i u_i v_i^T)\| \\
&= \|\sum_{i=1}^{m} \alpha_i u_i^T A - \sum_{i=1}^{m}\sum_{j=1}^{m} \delta_{ij}\alpha_i \sigma_j v_j^T\| \\
&= \|\sum_{i=1}^{m} \alpha_i \sigma_i v_i^T - \sum_{i=1}^{m} \alpha_i \sigma_i v_i^T\| \\
&= 0
\end{aligned} \tag{6}$$

Hence we have the definition of such the decomposition for arbitrary matrix.

Definition. *The **singular value decomposition** (SVD) of an arbitrary matrix $A \in \mathbb{R}^{m\times n}$ is denoted as:*

$$A = U\Sigma V^T$$

*Where U is a m x m orthogonal matrix , V is a n x n orthogonal matrix and Σ is a n x n matrix with $\Sigma_{ii} := \sigma_i \geq 0, \Sigma_{ij} = 0$ if $i \neq j$. σ_i are called **singular value** of this matrix. Usually, without loss of generality, we assume that $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n$. Or, denote $\sigma_i(A)$ to be the i-th largest singular value of matrix A.*

[Note: The singular values of a real symmetric matrix are known as absolutie value of its eigenvalues.]

There are several applications of the Singular Value Decomposition (SVD).

- Determining range, null space and rank of matrices.
- Matrix approximation .
- Least Squares approximation, and Inverse, Psuedo-Inverse
- Denoising
- Data Compression

The SVD and the eigen-decompositions are related but also there are differences between them.

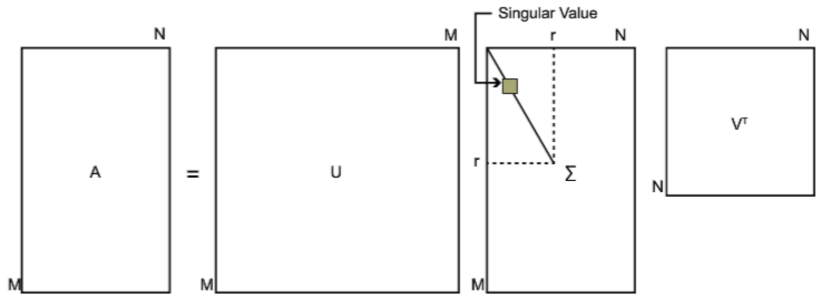

Figure 2: Singular Value Decomposition

- Not every (square) matrix has an eigen-decomposition, however any matrix has an SVD.
- In eigendecomposition the eigen-basis is not always orthogonal. The basis of singular vectors are always orthogonal.
- SVD has two singular spaces (defined by the left U and right singular vectors V).
- Computing the SVD is more numerically stable.

Relation with matrix norms
Let $A \in \mathbb{R}^{m\times n}$. For $1 \le p \le \infty$, the (induced) p-norm (also called the p-spectral norm) of A is defined as the solution of the optimization problem

$$\|A\|_p = \max_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \|\mathbf{x}\|_p=1}} \|Ax\|_p.$$

The Frobenius norm of a matrix

$$A = \begin{pmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{pmatrix}.$$

is given by

$$\|A\|_{\mathrm{F}} = \sqrt{\mathbf{trace}(A^{\mathrm{T}}A)} = \sqrt{\mathbf{trace}(AA^{\mathrm{T}})} = \sqrt{\sum_{i=1}^{n} \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i} = \sqrt{\sum_{j=1}^{m} \mathbf{a}_j^{\mathrm{T}}\mathbf{a}_j},$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ denote the columns and rows of A respectively. (Note, a matrix is just a tensor of order 2.) Note we additionally allow A to have complex entries. There are several properties and inequalities between matrix norms. Some of them (as examples) are

Lemma. $\|AB\|_2 \le \|A\|_2\|B\|_2$.

Lemma. $\|A\|_2 \le \|A\|_{\mathrm{F}} \le \sqrt{min(m,n)}\|A\|_2$.

The Transpose of $A \in \mathbb{R}^{m\times n}$ is given by $A^{\mathrm{T}} \in \mathbb{R}^{n\times m}$ with $A^{\mathrm{T}}(i,j) = A(j,i)$. If A is a complex matrix then we use the *adjoint* which is the complex conjugate of the transpose, namely, $A^* = \overline{A^{\mathrm{T}}}$.

Example. *These matrix norms relates with the matrix's singular values :*

- *The L_2 norm of a matrix has a strong relationship with singular values. A matrix's L_2 norm equals to the maximum singular value of the matrix. In fact:*

$$\|A\|_2 := \sup_{\|x\|_2\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{\|x\|_2=1} \sqrt{x^TA^TAx} = \sqrt{\lambda_{max}(A^TA)} = \sigma_{max}(A)$$

- *The **Frobenius** form is defined as:*

$$\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$$

And you can verify that :

$$\|A\|_F = \sqrt{tr(A^T A)} = \sqrt{\sum_{i=1}^{n} \sigma_i^2(A)}$$

by using the SVD decomposition of A.

- *Nuclear norm is defined as:*

$$\|A\|_* = \sum_i \sigma_i(A)$$

The singular value decomposition (SVD) of a matrix yields valuable geometric information about the matrix. For example, note

Lemma. *A square matrix A is invertible $\iff$ all of A's singular (eigen) values are non-zero.*

Proof. From the SVD of A we have $A = U\Sigma V^T$. If Σ has no zeros on the diagonal, then B $= V\Sigma^{-1}U^T$ exists, and $AB = U\Sigma V^T \Sigma V \Sigma^{-1} U^T =$ I, and similarly $BA = I$. Thus A is invertible.
If A is invertible, then we can construct Σ^{-1}, which is sufficient to show that all the singular values of A are all non-zero:

$$\mathrm{A} = \mathrm{U}\ \Sigma\ V^T$$
$$\Sigma = U^T \mathrm{AV}$$
$$\Sigma^{-1} = V^T\ A^{-1}\ \mathrm{U}$$

□

Exercise. *Let A be an invertible symmetric $n \times n$ matrix. Let $|\lambda|_{max}$ be the maximum of the absolute values of its eigenvalues, and let $|\lambda|_{min}$ be the minimum of the absolute values of its eigenvalues. Show that*

$$cond(A) = |\lambda|_{max}/|\lambda|_{min}$$

Computing the SVD : Power Iteration
We give out a simple algorithm for computing the SVD of a matrix $A \in \mathbb{R}^{m\times n}$. We start by computing the first singular value σ_1 and left and right singular vectors u_1, v_1 of A, for which $min_{i<j} \log(\frac{\sigma_i}{\sigma_j}) \geq \lambda$ (λ is a threshold value):

1. Generate x_0 such that $x_0(i) \sim \mathcal{N}(0,1)$.
2. $s \longleftarrow \log(\frac{4\log(\frac{2n}{\delta}}{\epsilon\delta}/2\lambda$.
3. for i in $[1, 2, \ldots, s]$:
4. $\quad x_i \leftarrow A^T A x_{i-1}$.
5. $v_1 \longleftarrow \frac{x_i}{\|x_i\|}$
6. $\sigma_1 \longleftarrow \|Av_1\|_2$
7. $u_1 \longleftarrow \frac{1}{\sigma_1} Av_1$
8. return (σ_1, u_1, v_1)

2 Normed, Metric and Inner Products Spaces

In this section we will introduce normed, metric and inner products Spaces. Vector spaces show us to speak linear transformations, summation, subspace and duality.

2.1 Topological Space

Definition. *X is an nonempty set, $\mathcal{X}$ is the class of subsets of X such that:*

- $X \in \mathcal{X}$
- $\emptyset \in \mathcal{X}$
- $X_1, X_2, \ldots, X_n \in \mathcal{X} \implies \bigcap_{i=1}^{n} X_i \in \mathcal{X}$ *(finite intersection)*
- $\bigcup_{i \in \mathcal{I}} X_i \in \mathcal{X}$ *(any union)*

*Then $\mathcal{X}$ defines the topology on X, $\forall x \in \mathcal{X}$ is called an open set in X, and $\mathcal{V} = (X, \mathcal{X})$ forms a **topological space**.*

Definition. *$x_1 \in X$, B_{x_1} is defined as neighborhood of x_1 if B_{x_1} is a subset at X and there exists an open set $U \in \mathcal{X}$ containing x_1 s.t. $U \subset B_{x_1}$*

Definition. *$\mathcal{V} = (X, \mathcal{X})$ is **Hausdorff** if and only if, $\forall$ pair of points $x_1, x_2 \in X$, $\exists$ neighborhood B_{x_1},B_{x_2} such that:*

$$B_{x_1} \cap B_{x_2} = \emptyset$$

(Point) Topological spaces allow us to speak of open sets, closed sets, compactness, convergence of sequences, continuity of functions, etc.

Example. *Let $\{X, \mathcal{X}\}$,$\{Y, \mathcal{Y}\}$ be two topological spaces. $F : X \to Y$ is a continuous mapping at $x_0 \in X$ if and only if : $\forall$ open set $Y_0 \in \mathcal{Y}$ containing $F(x_0)$ contains an open set B that is the image of an open set containing x_0. (An open set's original image is an open set)*

2.2 Metric Space

Definition. *A metric space is an ordered pair (X, d) where X is the set and d is a function defined on $X \times X$:*

$$d : X \times X \to R$$

such that for $\forall x, y, z \in X$, the following holds:

- $d(x, y) \geqslant 0$
- $d(x, y) = 0 \implies x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leqslant d(x, y) + d(y, z)$

Example. *Every **metric space** (denoted as (X, d)) is a topological space. Since we can define open sets*

$$B_r(x_0) = \{y \in X : d(x_0, y) = r\}$$

like the balls on metric space. In this case:

- $x_n \to x_0 \iff \forall \epsilon > 0, \exists n \in \mathbb{N}$ *such that* $d(x_0, x_n) < \epsilon$ *for all* $m > n$
- F *is continuous* $\iff \forall \epsilon > 0, \exists \delta > 0$ *such that* $d(F(x), F(x_0)) < \epsilon$ *whenever* $d(x, x_0) < \delta$

2.3 Topological Vector Space

Definition. *$\mathcal{V}$ is called a topological vector space if and only if:*

- *$\mathcal{V}$ is a vector space*
- *The underlying set V of vectors in $\mathcal{V}$ is endowed with a topology $\mathcal{U}$ such that:*
  - *$(V,\mathcal{U})$ is a Hausdorff topological space*
  - *vector addition is continuous: $u + v \in V$ if $u, v \in V$*
  - *scalar multiplication is continuous: $\alpha u \in V$ if $\alpha \in F, u \in V$*

2.4 Normed Space

Definition. *V is a vector space, $N : V \to \mathbb{R}$ is a **norm** of V if :*

- *$N(v) \geqslant 0$, and $N(v) = 0$ if and only if $v = 0$*
- *$N(\alpha v) = |\alpha| N(v)$*
- *$N(u + v) \leqslant N(u) + N(v)$*

We always denote $\|u\| := N(u)$.

remark. You can verify that $\|u - v\|$ is a metric on V, thus every normed space is a metric space.

remark. You can also verify that every normed space is a Topological Vector Space:

- let we assume there are two convergent sequence $\{u_n\}, \{v_n\} \subset V$:

$$u_n \to u, v_n \to v$$

  where $u, v \in V$, then we can verify that:

$$\|(u_n + v_n) - (u + v)\| \leqslant \|u - u_n\| + \|v - v_n\| \to 0 \text{ as } n \to 0$$

- Suppose $\alpha_n \to \alpha$ in $\mathbb{F}$, then:

$$\|\alpha_n u_n - \alpha u\| \leqslant |\alpha - \alpha_n| \|u_n\| + |\alpha| \|u - u_n\| \to 0 \text{ as } n \to \infty$$

Therefore, in normed spaces, we have the concept that adapted both from linear spaces and topological spaces. Next is the definition for a Banach Space.

Definition. *A complete normed space is a **Banach** space, or a B space. Here complete means : every Cauchy sequence in a metric space converges in that metric space.*

Here are some properties pertinent to normed spaces:

- $A : U \to V$, U, V are underlying sets of normed spaces with norms $\|\cdot\|_U, \|\cdot\|_V$, respectively.
- A is linear if and only if

$$A(\alpha u_1 + \beta u_2) = \alpha A(u_1) + \beta A(u_2) \forall u_1, u_2 \in U$$

- A is bounded if and only if A maps a bounded sets in U into bounded sets in V:

$$\|u\|_U \leqslant C_1 \implies \exists C_2 \text{ such that } \|Au\|_V \leqslant C_2$$

- A is continuous if and only if $\forall \epsilon > 0, \exists \delta > 0$ such that:

$$\|u - v\|_U < \delta \implies \|Au - Av\|_V < \epsilon$$

or if and only if , whenever $u_n \to u$ ($\|u - u_n\|_V \to 0$ as $n \to \infty$), we have:

$$\|Au - Av\|_V \to 0 \text{ as } n \to \infty$$

Theorem. *Let $(U, \|\cdot\|_U), (V, \|\cdot\|)_V$ be normed spaces over the same field. Let $A : U \to V$ be a linear function. Then the following are equivalent:*

1) A is continuous

2) A is continuous at $u = 0$

3) A is bounded

4) $\exists C > 0$ such that:

$$\|Au\|_V \leqslant C\|u\|_U \qquad \forall u \in U$$

Proof.
1) $\Rightarrow$ 2) is obvious.
2) $\Rightarrow$ 3):
Let $\|u\|_U < r$. Since A is continuous at 0, $\forall \epsilon > 0, \exists \delta > 0$ such that

$$\|Au\|_V < \epsilon \implies \|u\|_U < \delta$$

Pick $\epsilon = 1$, then $\exists\ \delta$ such that $\|u\|_U < \delta \Rightarrow \|Au\|_V < 1$.
If $\|u\|_U < r$,

$$\|\frac{\delta}{r}u\|_U = \frac{\delta}{r}\|u\|_U \leqslant \delta$$

Thus

$$\|A(\frac{\delta}{r}u)\|_V \leqslant 1 \implies \|Au\|_V \leqslant \frac{r}{\delta} = \text{constant}$$

Hence, A is bounded.
3) $\Rightarrow$ 4):
Since A is bounde, $\exists C > 0$ such that $\|Au\|_V \leqslant C$ whenever $\|u\|_U \leqslant 1$.
Thus, $\forall u \neq 0$,

$$\|A(\frac{u}{\|u\|_U})\|_V \leqslant C$$

and therefore:

$$\|Au\|_V \leqslant C\|u\|_U$$

4)$\Rightarrow$ 1):
If $u_n \to u$, then:

$$\|Au - Au_n\|_V \leqslant C\|u - u_n\|_V \to 0 \text{ as } n \to \infty$$

$\square$

2.5 Inner Product Space

Definition. *Let V be a vector space, and define $p : V \times V \to \mathbb{F}(\mathbb{C}$ or $\mathbb{R})$. Then p is an **inner product** on V if it satisfies the following:*

- $\forall u \in V, p(u, u) \geqslant 0$; $p(u, u) = 0 \iff u = 0$

- $\forall u, v \in V, p(u, v) = \overline{p(v, u)}$ *(Conjugate Symmetry)*
- $\forall u_1, u_2, v \in V, \forall \alpha_1, \alpha_2 \in \mathbb{F}, p(\alpha_1 u_1 + \alpha_2 u_2, v) = \alpha_1 p(u_1, v) + \alpha_2 p(u_2, v)$

*Denote as $(u, v) = p(u, v)$. A vector space on which an inner product has been defined is called an **inner product space**. Denote the inner product space as $(V, (\cdot, \cdot))$*

remark. You can verify that an inner product also satisfies the following:

$$p(u, \beta_1 v_1 + \beta_2 v_2) = \bar{\beta}_1 p(u, v_1) + \bar{\beta}_2 p(u, v_2)$$

where $u, v_1, v_2 \in V, \beta_1, \beta_2 \in \mathbb{F}$

Definition. *Let $(V, (\cdot, \cdot))$ be an inner product space. Pick $u, v \in V$, we claim that u and v are **orthogonal** if*

$$(u, v) = 0$$

One important property for the inner product is that it satisfies the Cauthy-Schwarz Inequality.

Theorem (Cauthy-Schwarz Inequality)**.** *Let $(V, (\cdot, \cdot))$ be an inner product space. If $u, v \in V$, then:*

$$|(u, v)| \leqslant \sqrt{(u, u)(v, v)}$$

Proof. Suppose $\mathbb{F} = \mathbb{C}$, pick $\alpha = \frac{\overline{(v,u)}}{(v,v)} \in \mathbb{C}$, then:

$$\begin{aligned}
0 &\leqslant (u - \alpha v, u - \alpha v) \\
&= (u, u) - \alpha(v, u) - \bar{\alpha}(u, v) + \alpha\bar{\alpha}(v, v) \\
&= (u, u) - \frac{\overline{(v, u)}}{(v, v)}(v, u) - \frac{\overline{(u, v)}}{(v, v)}(u, v) + \frac{\overline{(v, u)(u, v)}}{(v, v)^2}(v, v) \\
&= \frac{1}{(v, v)} \left[(u, u)(v, v) - 2|(u, v)|^2 + |(u, v)|^2\right]
\end{aligned}$$

Therefore $|(u, v)|^2 \leqslant (u, u)(v, v)$. □

Next, we want to connect the inner product space with normed space.

Theorem. *Every inner product space is a normed space with norm :*

$$\sqrt{(u, u)} = \|u\|$$

Proof. Recall the definition of the norm, all you need is to verify that

- $\|u\| \geqslant 0$ and $\|u\| = 0 \iff u = 0$
- $\|u + v\| \leqslant \|u\| + \|v\|$
- $\forall \alpha \in \mathbb{F}, \|\alpha u\| = |\alpha|\|u\|$

□

remark. It is understood that the inner product space V is induced with the topology induced by the norm $(u, u)^{\frac{1}{2}}$
Now, we introduce an important type of space:

Definition. *An inner product space is a **Hilbert Space** if and only if it is complete (with respect to the norm induced by the inner product)*

A typical example of a Hilbert Space will be the Euclidean Space $\mathbb{R}^d$ with an inner product defined as:

$$(x, y) = \sum_{i=1}^{d} x_i y_i$$

Theorem. *Suppose an inner product space $(V, (\cdot,\cdot))$ has two convergence sequence in norm:*

$$v_m \to v \text{ and } u_m \to u$$

Then

$$(v_m, u_m) \to (v, u)$$

Proof. In fact, we have:

$$\begin{aligned} |(v_m, u_m) - (v, u)| &= |(v_m, u_m) + (v_m, u) - (v_m, u) - (v, u)| \\ &= |(v_m, u_m - u) + (v_m - v, u)| \\ &\leqslant \|v_m\|\|u_m - u\| + \|v_m - v\|\|u\| \\ &\to 0 \text{ as } m \to \infty \end{aligned} \tag{7}$$

□

remark. Similar as Euclidean Space, inner product shares some geometric properties in general vector space:

- $\cos\theta \overset{def}{=} \frac{(u,v)}{\|u\|\|v\|} \quad (\mathbb{F} = \mathbb{R})$
- Pythagoras: $(u, v) = 0 \Rightarrow \|u + v\|^2 = \|u\|^2 + \|v\|^2$
- Sphere: $(u - u_0, u - u_0) = a^2$
- Hyperplane: $(u - a, n) = 0$
- Parallelogram Law: $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$

3 l_p Function norm and approximation

We consider the choice of norm in the approximations of functions. First we define what is the l_p norm of a function:

Definition. *The l_p norm of the function f given at some finite data points set $X = \{x_i \ : \ i = 1, \ldots, m\}$, is defined by*

$$l_p(f) = \|f\|_p := \left(\sum_{i=1}^{m} |f(x_i)|^p\right)^{1/p}, \quad 1 \le p < \infty.$$

For $p = \infty$, the l_∞ norm is defined by $\|f\|_\infty = \max_{i\in\{1,\ldots,m\}} |f(x_i)|$.

Next, we define what is the best approximation under l_p norm:

Definition. *The function $F(a^*, x)$, is said to be the best approximation of the function $f : \mathbb{R}^n \to \mathbb{R}$ in the norm $\|\cdot\|_p$, if*

$$\|F(a^*) - f\|_p \le \|F(a) - f\|_p, \quad \forall a \in \mathbb{R}^n$$

In that way, the approximation problem is reduced to the problem of minimization of the functional $\|F(a) - f\|_p)$. In general, the best approximation in the lp norm is different from the best approximation in the l_q norm ($p \neq q$). The lp norms can be generalized by introducing the weights ($w(x_i), \ i = 1, \ldots, m$).

If the approximating function $F(a, x)$ is linear in parameters a_j , $j = 1, \ldots, n$, i.e. if $F(a, x) = x^T a$,then

$$1 \le p < q \le \infty \Longrightarrow \min_{a\in\mathbb{R}^n} \|\mathbf{X}a - f\|_q \le \min_{a\in\mathbb{R}^n} \|\mathbf{X}a - f\|_p$$

(where $\mathbf{X}$ is a corresponding data matrix, and f is a vector of values of the dependent variable).

References

[BHK] Avrim Blum, John Hopcroft and Ravindran Kannan. *Foundations of Data Science.*

[EL] Edo Liberty. *Online Lectures by Edo Liberty (source)*

[Fl] Hermann Flaschka. *Principles of Analysis*, 1995

[S] Philip Sabes. *Online Lectures by Philip Sabes, UCSF*

[SE] Stefan Evert. *Online Lectures by Stefan Evert, Osnabruck, Germany*

[TM] Tomislav Marosevic. *Online lecture presented at the Mathematical Colloquium in Osijek organized by Croatian Mathematical Society*