1 Background on Linear Algebra

We briefly overview a few important topics from linear algebra that will be used throughout this course.

1.1 Eigenvalues and singular values

Definition 1 Given a square matrix $A \in \mathbb{C}^{n \times n}$, any $(\lambda, q) \in \mathbb{C}^1 \times \mathbb{C}^n$ is called a pair of eigenvalue and eigenvector of $A$ if

\[ Aq = q\lambda. \]

Moreover, the largest norm of the eigenvalues of $A$ is called the spectral radius of $A$.

Remark 1 If $A = A^T$, i.e., $A$ is symmetric and also $A \in \mathbb{R}^{n \times n}$, then it can be shown that there exist $n$ pairs of eigenvalue-eigenvectors $(\lambda_i, q_i) \in \mathbb{R}^1 \times \mathbb{R}^n$ such that $q_i^T q_j = 0$, $\forall i \neq j$ and $\|q_i\| = 1$, $\forall i$ and

\[ A = \sum_{i=1}^{n} \lambda_i q_i q_i^T, \quad (1) \]

where $(\lambda_i, q_i)$ are pairs of eigenvalue-eigenvectors of $A$ and $(\lambda_i, q_i) \in \mathbb{R}^1 \times \mathbb{R}^n$. Equivalently, we can write

\[ A = Q\Lambda Q^T, \quad (2) \]

1
where \( Q := [q_1 | \ldots | q_n] \) and \( \Lambda := \text{Diag}(\lambda_1, \ldots, \lambda_n) \). Equation (2) is called an eigenvalue decomposition of \( A \).

**Definition 2** A symmetric matrix \( A \in \mathbb{R}^{n \times n} \) is called positive semidefinite (PSD) if

\[
x^T A x \geq 0, \quad \forall x \in \mathbb{R}^n.
\]  

Moreover, if the inequality in (3) is strict for all \( x \in \mathbb{R} \setminus \{0\} \), then we call \( A \) a positive definite (PD) matrix.

Let \( A \) be a real symmetric matrix with the eigenvalue decomposition

\[
A = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T.
\]

One can write

\[
x^T A x = x^T \left( \sum_{i=1}^{n} \lambda_i q_i q_i^T \right) x
\]

\[
= \sum_{i=1}^{n} \lambda_i (q_i^T x)^2.
\]

Therefore, \( A \) is PSD if and only if \( \lambda_i \geq 0, \forall i = 1, \ldots, n \) and similarly \( A \) is PD if and only if \( \lambda_i > 0, \forall i = 1, \ldots, n \). (Note that since \( q_i \) are orthogonal, if \( x = q_i \), then \( x^T A x = \lambda_i (q_i^T x)^2 \).

Eigenvalue decomposition may not exist if for example \( A \) is non-symmetric or rectangular. However, for any matrix \( A \), \( AA^T \) and \( A^T A \) are both symmetric. This allows us to obtain the so called Singular Value Decomposition (SVD) of \( A \in \mathbb{R}^{n \times m} \) (\( m \leq n \)) as

\[
A = U \Sigma V^T,
\]

where \( U \) and \( A \) are unitary matrices of appropriate dimension i.e. \( U^T U = I_n \) and \( V^T V = I_m \). Moreover, \( \Sigma = \text{Diag}(\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^{n \times m} \), where \( \sigma_i^2 \in \text{eig}(AA^T) \) and \( \sigma_1 \geq \ldots \geq \sigma_n \geq 0 \).

Note that there is no connection between the singular values of a matrix and the matrix being PSD.
1.2 Matrix norms

Definition 3 Given a matrix $A \in \mathbb{R}^{n \times m}$ we define

1. **Frobenius norm:** $||A||_F := \sqrt{\sum_{i,j} A_{ij}^2}$

2. **Induced norm:** $||A||_2 := \sup_x \frac{||Ax||}{||x||}$

3. **Nuclear norm:** $||A||_* = \sum \sigma_i$

The Frobenius norm is analogous to the 2-norm of a vector as $||A||_F = ||\text{vec}(A)||_2$, where $\text{vec}(A) \in \mathbb{R}^{mn}$ is formed by putting the columns of $A$ in a single vector format. Moreover, akin to the 2-norm that is defined using the inner product $\langle u, v \rangle = \sum u_i v_i$, $\forall u, v \in \mathbb{R}^n$, the Frobenius norm can be obtained using the matrix inner product

$$\langle A, B \rangle := \text{trace}(A^T B) = \text{trace}(B^T A) = \sum A_{ij} B_{ij},$$

as

$$||A||_F = \sqrt{\langle A, A \rangle}.$$ 

Furthermore, using the SVD decomposition of $A$, one can write

$$||A||_F^2 = \langle A, A \rangle = \text{trace}(AA^T) = \text{trace}(U \Sigma V^T V \Sigma^T U^T) = \text{trace}(U \Sigma \Sigma^T U^T) = \sum_i \sigma_i^2 \text{trace}(u_i u_i^T) = \sum_i \sigma_i^2. \quad (10)$$

We can also show that the 2-norm of a matrix $A$ is equal to the largest singular value of $A$ i.e. $\sup_x \frac{||Ax||}{||x||} = \max_i \sigma_i$. This is because

$$\sup_x \frac{||Ax||}{||x||} = \sup_x \frac{||U \Sigma V^T x||}{||x||} = \sup_x \frac{||\Sigma x||}{||x||}. \quad (11)$$

Note that $(\sum \sigma_i)^2 \geq \sum \sigma_i^2 \geq \max(\sigma_i)^2$, therefore

$$||A||_* \geq ||A||_F \geq ||A||_2.$$
Note: The matrix inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{A}^T \mathbf{B})$ satisfies the Cauchy-Schwarz inequality

$$\text{trace}(\mathbf{A} \mathbf{A}^T) \text{trace}(\mathbf{B} \mathbf{B}^T) \geq (\text{trace}(\mathbf{A} \mathbf{B}^T))^2. \quad (12)$$

2 Limiting Behavior of Functions

We start this section with an illustrative example. Consider the two functions $f_1(x) = x$ and $f_2(x) = x^2$. It is intuitive that $f_2$ grows faster than $f_1$ as $x \to \infty$. In an attempt to make this notion rigorous, we adopt the following notation.

**Definition 4** Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be real valued functions. As a comparison between the asymptotic behavior of $f$ and $g$ as $x \to \infty$, we call the function $f$

1. $O(g)$ as $x \to \infty$ if $\exists \alpha, \ x_0 > 0 \ s.t. \ |f(x)| \leq \alpha |g(x)|, \ \forall x > x_0$,
2. $\Omega(g)$ as $x \to \infty$ if $\exists \alpha, \ x_0 > 0 \ s.t. \ |f(x)| \geq \alpha |g(x)|, \ \forall x > x_0$,
3. $o(g)$ as $x \to \infty$ if $\forall \alpha, \ \exists x_0 > 0 \ s.t. \ |f(x)| \leq \alpha |g(x)|, \ \forall x > x_0$,
4. $\omega(g)$ as $x \to \infty$ if $\forall \alpha, \ \exists x_0 > 0 \ s.t. \ |f(x)| \geq \alpha |g(x)|, \ \forall x > x_0$.

Similarly, for any $a \in \mathbb{R}$ we call $f$

1. $O(g)$ as $x \to a$ if $\exists \alpha, \ x_0 > 0 \ s.t. \ |f(x)| \leq \alpha |g(x)|, \ \forall x \ s.t. \ |x-a| \leq x_0$,
2. $\Omega(g)$ as $x \to a$ if $\exists \alpha, \ x_0 > 0 \ s.t. \ |f(x)| \geq \alpha |g(x)|, \ \forall x \ s.t. \ |x-a| \leq x_0$,
3. $o(g)$ as $x \to a$ if $\forall \alpha, \ \exists x_0 > 0 \ s.t. \ |f(x)| \leq \alpha |g(x)|, \ \forall x \ s.t. \ |x-a| \leq x_0$,
4. $\omega(g)$ as $x \to a$ if $\forall \alpha, \ \exists x_0 > 0 \ s.t. \ |f(x)| \geq \alpha |g(x)|, \ \forall x \ s.t. \ |x-a| \leq x_0$.

The big Oh notation e.g. means that if $h(x) := |\frac{f(x)}{g(x)}|$, then there exist $\alpha > 0$ and $x_0$ such that $h(x) < \alpha$ for all $x \geq x_0$. This means that the plot of $h$ lies below the horizontal line $y = \alpha$ for any $x \geq x_0$. A function $f$ being $O(g)$ basically means that $f$ is below $g$ “asymptotically”. Similarly, $f = \Omega(g)$ can be inferred as $f$ lies above $g$ “asymptotically”. Alternatively, $f = o(g)$ means that $\lim_{x \to \infty} |h(x)| = 0$ and $f = \omega(g)$ means that $\lim_{x \to \infty} |h(x)| = \infty$.

**Example:** It is straightforward to verify that
1. $10 \sin(x) = O(1)$, however, $10 \sin(x) \neq \Omega(1)$ as $x \to \infty$.

2. $4x^4 + 3x^2 = O(x^2)$ as $x \to 0$.

3 A Few Basic Differentiation Rules

Let $e_1, e_2, \ldots, e_n$ be the standard basis of $\mathbb{R}^n$. Given a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the partial derivatives of $f$ ($e_i$-directional derivates of $f$) are defined as

$$\frac{\partial f(x)}{\partial x_i} := \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}. \quad (13)$$

If we concatenate all $\frac{\partial f(x)}{\partial x_i}$ in a column, we obtain the gradient of $f$ as

$$\nabla f(x) := \left( \frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T. \quad (14)$$

The second derivative of $f$ is called the Hessian matrix of $f$ and is defined as

$$\nabla^2 f(x) = \left[ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right]. \quad (15)$$

Note that the Hessian is always symmetric.

**Example** Let $f(x) = x_1^2 + x_2^2 + 2x_1x_2 + x_1$. The gradient and Hessian of $f$ are as follows.

$$\nabla f(x) = \begin{bmatrix} 2x_1 + 2x_2 + 1 \\ 2x_1 + 2x_2 \end{bmatrix}, \quad (16)$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}. \quad (17)$$
3.1 Taylor Expansion

Using the first and second derivatives of a function $f$ evaluated at $x$ we can approximate the value of $f$ at a neighborhood of $x$ as

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x) + o(||y - x||^2),$$

(18)

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x).$$

(19)

The right hand side of (19) is called the Taylor Expansion of $f$ around $x$ which is the second order approximation of $f$ around $x$. Specifically, in Equation (18), $o(||y - x||^2)$ means that the error in (19) divided by $||y - x||^2$ goes to zero as $y$ goes to $x$.

Note that the first and higher order approximations of $f$ can also be obtained similarly.

3.2 Mean value theorem

Let $f$ be a differentiable function and $x, y \in \mathbb{R}^n$. The mean value theorem states that there exist $\zeta, \eta$ in the line segment between $x$ and $y$ such that

$$f(y) = f(x) + \nabla f(\zeta)^T (y - x),$$

(20)

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(\eta) (y - x).$$

(21)

In other words, Equation (20) means that $\frac{f(y) - f(x)}{y - x} = f'(\zeta)$, for some $\zeta$ in the line segment between $x$ and $y$.

Note that the difference between Equation (20) and the first order Taylor approximation of $f$ at $x$ is that in Equation (20), the derivative is evaluated at $\zeta$ instead of $x$ and this makes the approximation of $f$ at $y$ precise.

3.3 Chain rule

So far, we have only discussed the derivatives of real valued functions. Now, consider a multivalued function $f: \mathbb{R}^n \to \mathbb{R}^m$. The Jacobian matrix of $f$ is
defined as
\[ \nabla f(x) = [\nabla f_1(x)] \ldots [\nabla f_m(x)] \] (22)
where \( f_i \) is the \( i \)th component of \( f \). Note that Jacobian is clearly an \( n \times m \) matrix.

Example: Let \( f(x_1, x_2) = [x_1^2, x_2, x_1]^T \). The Jacobian of \( f \) is
\[ \nabla f = \begin{bmatrix} 2x_1 & 0 & 1 \\ 0 & 2x_2 & 0 \end{bmatrix}. \] (23)

Now, let \( f : \mathbb{R}^k \to \mathbb{R}^m \) and \( g : \mathbb{R}^m \to \mathbb{R}^n \) be two differentiable mappings. In order to evaluate the derivative of \( h := g \circ f \), one can use the Chain Rule
\[ \nabla h(x) = \nabla f(x) \nabla g(f(x)). \] (24)

Example: Let \( f(x) = Ax \) be a linear function and let \( g(y) \) be differentiable. Using chain rule, one can write
1. \( \nabla g(Ax) = \nabla g(f) = A^T \nabla g(Ax) \).
2. \( \nabla^2 g(Ax) = \nabla^2 g(f) = A^T \nabla^2 g(Ax) A \).

The second example is obtained by two times applying the chain rule to \( g \circ f \).

4 Lipschitz Continuity and Contraction Mapping

A function is continuous if for any convergent sequence \( (a_i) \in \mathbb{N} \) with \( \lim_{i \to \infty} a_i = a \), it satisfies \( \lim_{i \to \infty} f(a_i) = f(a) \). Intuitively, the graph of continuous functions have no jump.

Definition 5 A function \( f \) is called Lipschitz continuous if there exist \( \lambda > 0 \) such that
\[ \frac{|f(x) - f(y)|}{|y - x|} \leq \gamma, \quad \forall x, y. \] (25)

Moreover, any \( \lambda > 0 \) that satisfies (25) is called a Lipschitz continuity factor of \( f \).
Note that any Lipschitz continuous function is continuous, however the converse is not true. In (25), \(| . |\) is the 2-norm, however, Lipschitz continuity can be defined with respect to different norms as well.

**Example:** The functions \(g(x) = 1\) and \(h(x) = x\) are Lipschitz continuous however \(f(x) = x^2\) is not. In order to show that \(f\) is not Lipschitz continuous, we can write

\[
|f(x) - f(y)| \leq \lambda |(x - y)|, \quad \forall x, y \implies (26)
\]

\[
|(x - y)(x + y)| \leq \lambda |(x - y)|, \quad \forall x, y \implies (27)
\]

\[
(x + y)^2 \leq \lambda^2, \quad \forall x, y, (28)
\]

which is a contraction.

Note that if we restrict the domain of \(f\) to an interval, then \(f\) becomes Lipschitz continuous.

**Proposition 1** If a function \(f\) has bounded derivative, then \(f\) is Lipschitz continuous.

**Definition 6** Let function \(f\) be Lipschitz continuous with factor \(\lambda \in \mathbb{R}\). If \(\lambda \leq 1\), then \(f\) is called non-expansive. Furthermore, if \(\lambda < 1\), then \(f\) is called contraction.

A mapping being contraction implies that \(|f(x) - f(y)| < |x - y|, \forall x, y\).

**Theorem 1 (Fixed point theorem)** Any contraction mapping \(f : \mathbb{R}^n \to \mathbb{R}^m\) has a unique fixed point \(x^*\), i.e., \(f(x^*) = x^*\). Moreover, \(\lim_{n \to \infty} f^n(x) = x^*, \forall x\).

The fixed point theorem holds for any closed set. Also note that it is straightforward to show that the fixed point is unique. Let \(f(x) = x\) and \(f(y) = y\). We have \(|y - x| = |f(y) - f(x)| \leq \lambda |y - x| < |y - x|, \) therefore \(|y - x| = 0\).

**Example:** The function \(f(x) := \frac{x}{2}\) is contraction because any \(\lambda \geq \frac{1}{2}\) is a Lipschitz continuity factor for \(f\). Moreover, \(0\) is the unique minimizer of \(f\).

Note that the fixed point theorem does not hold for non-expansive functions in general. For example, any rotation is a non-expansive mapping and although has a unique fixed point, any number of combinations of the rotation with itself does not converge to the fixed point.
5 Background on Probability

The reader is assumed to be familiar with the following concepts.

- Probability,
- Conditional probability,
- Random Variable (RV),
- Independence.

Let $X$ and $Y$ be RVs. We denote

- Expected value: $E(X) = \mu_x$,
- Variance: $\text{Var}(X) := E((X - \mu_x)^2) = E(X^2) - E(X)^2$,
- Covariance: $\text{Cov}(X, Y) = E((X - \mu_x)(Y - \mu_y)^T)$.

It can be shown that if $X$ and $Y$ are independent, then

$$E(XY^T) = 0 \quad \text{and} \quad \text{Cov}(X, Y) = 0. \quad (29)$$

We denote $X \sim N(\mu, \sigma^2)$ if $X$ has a Gaussian(Normal) distribution $f$, i.e.,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (30)$$

In the case that $X$ is multivariate, the notation $X \sim N(\mu, \Sigma)$ means that $X$ is jointly Normal, i.e.,

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \text{det}(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right). \quad (31)$$

Let $X_1, X_2, \ldots$ be Independent Identically Distributed (i.i.d.) RVs with $E(X_i) = \mu$ and $\text{Cov}(X_i) = \Sigma$. Let $S_n := \frac{1}{n} \sum_{i=1}^{n} X_i$.

**Law of large numbers:**

$$S_n \xrightarrow{a.s.} \mu$$

**Central limit theorem:**

$$\sqrt{n}(S_n - \mu) \xrightarrow{\text{dist.}} N(0, \Sigma)$$
Example: Consider a random experiment in which we flip a coin $n$ times and let $S_n$ denote the average number of wins. In fact,

$$S_n := \frac{1}{n} \sum_{i=1}^{n} X_i,$$

where $X_i$ is a Bernoulli distributed RV with probability $p \in [0,1]$. The law of large numbers states that $S_n$ converges to $p$ almost surely. The central limit theorem also states that $\sqrt{n}(S_n - p) \sim N(0, p - p^2)$.

Lemma 1 (Markov’s inequality) Let $X$ be a nonnegative RV.

$$P(X \geq a) \leq \frac{E(x)}{a}.$$  \hspace{1cm} (32)

Lemma 2 (Chebyshev’s inequality) Let $X$ be an RV with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$  \hspace{1cm} (33)

Note that if we think of $|X - \mu|$ as a nonnegative random variable, then we can use the Markov’s inequality to prove the Chebyshev’s inequality.

Lemma 3 (Cauchy-Schwarz inequality) Let $X$ and $Y$ be RVs.

$$|E(XY)|^2 \leq E(X^2)E(Y^2).$$
Contraction Mapping

Lecturer: Meisam Razaviyayn

Aug 24, 2017

**Contraction Mapping Theorem:** Suppose that $f : \mathcal{X} \rightarrow \mathcal{X}$ is a contraction mapping, i.e., $\|f(x) - f(y)\| \leq \gamma \|x - y\|$, $\forall x, y \in \mathcal{X}$ with $\gamma < 1$. Assume $\mathcal{X}$ is a closed subset of $\mathbb{R}^n$. Then the iterates defined by

$$x^{r+1} = f(x^r), \quad r = 0, 1, \ldots$$

is convergent to $x^* \in \mathcal{X}$ where $x^*$ is the unique fixed point of the mapping $f$, i.e., $f(x^*) = x^*$.

**Proof:** First notice that for any indices $k, \ell$, we have

$$\|x^{k+\ell} - x^k\| \leq \sum_{i=1}^{\ell} \|x^{k+i} - x^{k+i-1}\|$$

$$\leq \sum_{i=1}^{\ell} \gamma^{i-1} \|x^{k+1} - x^k\|$$

$$\leq \frac{1}{1-\gamma} \|x^{k+1} - x^k\|$$

$$\leq \frac{\gamma^k}{1-\gamma} \|x^1 - x^0\|.$$

Furthermore, for any given $\epsilon > 0$, we can choose $k$ large enough so that $\frac{\gamma^k}{1-\gamma} \|x^1 - x^0\| \leq \epsilon$ and consequently, $\|x^{m+\ell} - x^m\| \leq \epsilon$, $\forall m \geq k$, $\forall \ell$. Hence, $\{x^r\}$ is a Cauchy sequence and it converges to a limit point $x^*$. Moreover, $x^*$ is a fixed point of the mapping $f(\cdot)$ since

$$\|f(x^r) - x^r\| = \|x^{r+1} - x^r\| \leq \gamma^r \|x^1 - x^0\|.$$
Letting $r \to \infty$ implies $\|f(x^*) - x^*\| = 0$, i.e., $f(x^*) = x^*$. Moreover, it can be shown that the fixed point $x^*$ is unique since if $y^*$ is another fixed point with $f(y^*) = y^*$, we have

$$\|x^* - y^*\| = \|f(x^*) - f(y^*)\| \leq \gamma \|x^* - y^*\|.$$  

Since $\gamma < 1$, we must have $x^* = y^*$. 