

# ECE367: Matrix Algebra and Optimization

Aman Bhargava

September-December 2020

# Contents

0.1	Introduction and Course Information . . . . .	1
<b>1</b>	<b>Vector Space Review</b>	<b>2</b>
1.1	Basic Terminology . . . . .	2
1.2	Norms . . . . .	3
1.3	Inner Products . . . . .	4
1.4	Projection . . . . .	4
1.4.1	Projection onto Subspace . . . . .	4
1.4.2	Affine Projection . . . . .	6
<b>2</b>	<b>Functions and Sets</b>	<b>7</b>
2.1	Basic Terminology . . . . .	7
2.2	Linear and Affine Functions . . . . .	7
2.3	Affine Approximation . . . . .	8
2.4	Chain Rule on Gradients . . . . .	8
<b>3</b>	<b>Matrices</b>	<b>9</b>
3.1	Matrix Introduction . . . . .	9
3.1.1	Matrix Vector Spaces . . . . .	9
3.2	Function Approximations with Matrices . . . . .	10
3.3	Matrices as Linear Maps . . . . .	10
3.3.1	Matrix Inverse . . . . .	10
3.4	Orthogonal Matrices . . . . .	12
3.5	Rank, Range, and Null Space . . . . .	12
<b>4</b>	<b>Eigendecomposition</b>	<b>14</b>
4.1	How to Solve for Eigenthings . . . . .	14
4.2	Diagonalization . . . . .	15
<b>5</b>	<b>Symmetric Matrices</b>	<b>16</b>
5.1	Symmetric Matrices and Quadratic Functions . . . . .	16
5.2	Spectral Theorem . . . . .	16

5.3	Ellipsoids . . . . .	18
5.4	Matrix Square Root and Cholesky Decomposition . . . . .	18
<b>6</b>	<b>Singular Value Decomposition</b>	<b>20</b>
6.1	Motivation . . . . .	20
6.2	Steps to Creating SVD . . . . .	20
6.3	Principal Component Analysis . . . . .	22
6.4	SVD Matrix Properties . . . . .	22
6.5	Moore-Penrose Pseudo-Inverse . . . . .	23
<b>7</b>	<b>Systems of Linear Equations</b>	<b>24</b>
7.1	Mathematical Setup . . . . .	24
7.2	Over, Under, and Uniquely Determined Least Squares . . . . .	25
7.2.1	Overdetermined . . . . .	25
7.2.2	Uniquely Determined . . . . .	26
7.2.3	Underdetermined . . . . .	26
7.3	Regularized Least Squares . . . . .	27
7.3.1	$\ell_2$ Regularization . . . . .	27
7.4	Tikhonov Regularization . . . . .	27
7.5	Theoretical Aspects of Least Squares . . . . .	28
7.5.1	$\ell_1$ Regularization . . . . .	28
7.5.2	SVD Connection . . . . .	28

## 0.1 Introduction and Course Information

This document offers an overview of the ECE367 course. They comprise my condensed course notes for the course. No promises are made relating to the correctness or completeness of the course notes. These notes are meant to highlight difficult concepts and explain them simply, not to comprehensively review the entire course.

### Course Information

- Professor: Stark C. Draper
- Course: Engineering Science, Machine Intelligence Option
- Term: 2020 Fall

# Chapter 1

## Vector Space Review

### 1.1 Basic Terminology

**Subspaces:** Set of vectors closed under addition and scaling.

**Span:**  $\text{span}(S)$  is the set of linear combinations of set  $S$ .

**Linear Independence:** You cannot express one element of a set as a linear combination of the others.

$$\sum \alpha_i v^{(i)} = 0 \text{ iff } \alpha_i = 0 \forall i \in [m] \quad (1.1)$$

$\exists$  unique representation for any vector  $\in \text{span}(V)$

**Basis:** Set  $B$  is a basis iff:

1.  $B$  is linearly independent with  $d$  elements ( $d$  is the number of dimensions of the space).
2.  $\text{Span}(B) = V$ .

**Dimension:**  $\text{Dim}(V)$  is the *cardinality* of any  $\text{Basis}(V)$ .

**Cardinality:**  $\text{cardinality}(S)$  is the **number of elements** in set  $S$ .

**Direct Sum:**

- Let  $W_1, W_2$  be subspaces of  $V$ .
- Let  $U = W_1 + W_2 = \{w_1 + w_2 | w_1 \in W_1, w_2 \in W_2\}$ .

- Then  $\dim(U) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2)$
- **Direct Sum:**  $U = W_1 \oplus W_2$  iff  $W_1 \cap W_2 = \{0\}$  This implies that there is a **unique choice** for a representation of a vector in  $U$ .

## 1.2 Norms

**Norm Definition:** The norm function  $\|\cdot\| : V \rightarrow \mathbb{R}$  with the conditions:

1. Norm is always greater than equal to zero, equalling zero only for the zero vector.
2.  $\|u + v\| \leq \|u\| + \|v\|$  - *Triangle inequality*, very useful in proofs.
3.  $\|\alpha u\| = |\alpha| \cdot \|u\|$

$\ell_p$  **Norm Family:**

$$\|x\|_p = \left( \sum_{k=1}^n |x_k|^p \right)^{\frac{1}{p}} \quad (1.2)$$

For  $1 \leq p \leq \infty$ .

- **Euclidian Norm:**  $\ell_2$  norm is our conventional measure of distance in Euclidian space.
- $\ell_1$ : Is simply the sum of the absolute values in a vector.
- $\ell_\infty$ : Is the absolute value of the maximum magnitude element in the vector.

**Norm Ball:**  $B_p = \{x \in V \mid \|x\|_p \leq 1\}$ .

**Level Sets:**  $\{x \in V \mid \|x\|_p = c, c \in \mathbb{R}\}$

**Cardinality Function:**  $\text{card}(x) = \sum_{k=1}^n \mathbb{1}\{x_k \neq 0\}$  i.e. the number of non-zero elements.

## 1.3 Inner Products

**Inner Product Definition:**  $\langle \cdot, \cdot \rangle : x, y \in \mathbb{C}^n \rightarrow \mathbb{C}$ , as

$$\langle x, y \rangle = x^T \bar{y} = \sum_{k=1}^n x_k \bar{y}_k \quad (1.3)$$

Where  $\bar{y}$  is the complex conjugate of  $y$ .

**Cauchy-Schwartz:**  $|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$  holds for **all** inner product spaces.

**Holder's Inequality:**  $|\langle x, y \rangle| \leq \|x\|_p \|y\|_q$  under the conditions  $p, q \geq 1$  AND  $\frac{1}{p} + \frac{1}{q} = 1$ .

**Inner Product Space:** Space where the inner product **induces** a norm (somewhat hand-wavey definition).

**Induced Norm:** For vector space  $\mathbb{C}^n$ :  $\|x\|_2 = \sqrt{\langle x, x \rangle}$ .

## 1.4 Projection

**Goal:** Find vector  $y^*$  for some vector  $x$  such that

$$y^* = \Pi Y(x) = \operatorname{argmin}_{y \in Y} \|y - x\| \quad (1.4)$$

Where

- $Y$  is the set of vectors from which you choose your  $y^*$ .
- $\Pi$  is the projection operator.
- You are trying to choose optimal  $y^*$ .

### 1.4.1 Projection onto Subspace

**Theorem 1 *Projection onto multidimensional subspaces:*** Let  $x \in V$  and  $S \subseteq V$ . Then there exists unique  $y^* \in S$  such that

$$y^* = \operatorname{argmin}_{y \in S} \|x - y\| \quad (1.5)$$

- We let  $S = \operatorname{span}(\{v^{(1)}, \dots, v^{(n)}\})$ .

- We can now write  $y^* = \sum_{i=1}^d \alpha_i v^{(i)}$ .

$$\begin{bmatrix} \langle v^{(1)}, v^{(1)} \rangle & \dots & \langle v^{(1)}, v^{(d)} \rangle \\ \vdots & \ddots & \vdots \\ \langle v^{(d)}, v^{(1)} \rangle & \dots & \langle v^{(d)}, v^{(d)} \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{bmatrix} = \begin{bmatrix} \langle v^{(1)}, x \rangle \\ \vdots \\ \langle v^{(d)}, x \rangle \end{bmatrix} \quad (1.6)$$

Note that for an orthonormal set of vectors  $V$ , the lefthand matrix becomes the identity matrix. Therefore, one can project onto an orthonormal set by taking the inner products between the set elements and the vector in question.

**Gram-Schmidt Procedure:** How to make an orthonormal basis out of a set (go from  $v^{(i)}$  to orthonormal  $z^{(i)}$ ).

1. Normalize  $v^{(1)} \rightarrow z^{(1)}$ .
2. Let  $u^* = \langle v^{(2)}, z^{(1)} \rangle z^{(1)}$ .
3.  $z^{(2)} = \frac{v^{(2)} - u^*}{\|v^{(2)} - u^*\|}$ .
4. For higher dimension: Repeat steps before where the  $m$ th vector  $z^{(m)}$  is calculated as

$$z^{(m)} = \frac{w^{(m)}}{\|w^{(m)}\|}$$

$$\text{Where } w^{(m)} = v^{(m)} - \sum_{i=1}^{m-1} \langle v^{(m)}, z^{(i)} \rangle z^{(i)}$$

**QR Decomposition:** When you write a matrix as the product of an orthonormal basis and an upper triangular matrix.

$$A = \begin{bmatrix} \vdots & & \vdots \\ v^{(1)} & \dots & v^{(m)} \\ \vdots & & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & & \vdots \\ z^{(1)} & \dots & z^{(m)} \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ 0 & r_{22} & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & r_{mm} \end{bmatrix} \quad (1.7)$$

Note that this is a representation of taking the Gram-Schmidt procedure in reverse (i.e. the inverse of the  $R$  matrix corresponds to the process described above).

### 1.4.2 Affine Projection

**Affine Set Definition:** Affine set is a shift or translation of subspace  $S \subseteq \mathbb{R}^n$  by some vector  $x^{(0)} \in \mathbb{R}$ .

$$A = x^{(0)} + S = \{x^{(0)} + u | u \in S\}$$

1. Given  $x, A, x^{(0)}$ . You want to project  $x$  onto affine set  $A = S + x^{(0)}$
2. Translate  $x, A$  by  $-x^{(0)}$ .
3. Project onto the conventional subspace using the procedure above.
4. Translate the projection back using  $+x^{(0)}$ .



# Chapter 2

## Functions and Sets

### 2.1 Basic Terminology

**Prototypical Function:**

$$f : X \rightarrow Y \quad (2.1)$$

Means that  $f$  is a function that maps members of set  $X$  to members of set  $Y$ .

**Sets Related to Functions:**

- graph  $f = \{(x, f(x)) \in \mathbb{R}^{n+1} | x \in \mathbb{R}^n\}$
- epigraph  $f = \{(x, t) \in \mathbb{R}^{n+1} | x \in \mathbb{R}^n, t \geq f(x)\}$
- level sets:  $c_f(t) = \{x \in \mathbb{R}^n | f(x) = t\}$
- sub level sets:  $l_f(t) = \{x \in \mathbb{R}^n | f(x) \leq t\}$

### 2.2 Linear and Affine Functions

**Linearity Conditions:**

1. **Homogeneity:**  $f(\alpha x) = \alpha f(x)$ .
2. **Additivity:**  $f(x + y) = f(x) + f(y)$ .

**Resultant property:** Super position;  $f(\sum_{i=1}^n \alpha_i x^{(i)}) = \sum_{i=1}^n \alpha_i f(x^{(i)})$ .

**Affine Function:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is affine if  $\tilde{f}(x)$  (defined below) is linear

$$\tilde{f}(x) = f(x) - f(0) \quad (2.2)$$

**Useful property:**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is affine **iff**  $\exists a, b \in \mathbb{R}^n$  such that

$$f(x) = a^T x + b \quad (2.3)$$

Which also implies that all linear functions can be expressed as  $f(x) = a^T x$ .

## 2.3 Affine Approximation

**Affine Approximation Definition:** The affine approximation for function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  about point  $\bar{x}$  is:

$$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \quad (2.4)$$

- **First-order increment:**  $\nabla f(\bar{x})(x - \bar{x})$ .
- First-order increment equal to zero (or any constant value value) results in a **hyperplane**.

## 2.4 Chain Rule on Gradients

**Given**

1.  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$
2.  $f : \mathbb{R}^m \rightarrow \mathbb{R}$
3.  $\phi(x) = f(g(x)), x \in \mathbb{R}^n$

**Goal:** Find  $\nabla \phi(x)$ .

$$\nabla \phi(x) = \begin{bmatrix} \frac{\partial \phi}{\partial x_1} \\ \vdots \\ \frac{\partial \phi}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \frac{\partial g_2(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_1} \\ \vdots & \ddots & & \vdots \\ \frac{\partial g_1(x)}{\partial x_n} & \cdots & & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} \nabla f(g(x)) \quad (2.5)$$

# Chapter 3

## Matrices

### 3.1 Matrix Introduction

*This chapter began with a basic review of matrix algebra that will not be re-covered here.*

**Outer Product View of Matrix Multiplication:**

$$AB = \begin{bmatrix} | & & | \\ a^{(1)} & \dots & a^{(m)} \\ | & & | \end{bmatrix} \begin{bmatrix} - & b^{(1)T} & - \\ & \vdots & \\ - & b^{(m)T} & - \end{bmatrix} = \sum_{k=1}^m a^{(k)} (b^{(k)})^T \quad (3.1)$$

**Important Matrices:**

- **Square**
- **Symmetric:**  $A_{ij} = A_{ji}$ ;  $A = A^T$
- **Diagonal:**  $A_{ij} = 0$  for  $i \neq j$
- **Identity.**
- **Upper Triangular:** Square matrix with condition  $A_{ij} = 0$  if  $i > j$ .

#### 3.1.1 Matrix Vector Spaces

**Inner Product:**  $A, B \in \mathbb{C}^{m \times n}$

$$\langle A, B \rangle = \text{trace}(A^T B) = \text{trace}(B A^T) = \sum_i \sum_j [A]_{ij} [B]_{ij} \quad (3.2)$$

**Trace Operator:** For matrix  $A \in \mathbb{C}^{n \times n}$ :

$$\text{tr}(A) = \sum_{i=1}^n [A]_{ii} = [A]_{11} + \dots + [A]_{nn} \quad (3.3)$$

**Frobenius Norm:**  $\|\cdot\|_F$  where

$$\|A\|_F \equiv \sqrt{\langle A, A \rangle} \quad (3.4)$$

Which is the equivalent of the  $\ell_2$  norm of the vector made by flattening  $A$ .

These definitions and properties make proofs a lot easier when you deeply understand them.

## 3.2 Function Approximations with Matrices

**Second Order (Quadratic) Approximations:** For map  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , we can approximate by

$$f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T \nabla^2 f(\bar{x}) (x - \bar{x}) \quad (3.5)$$

Where  $\nabla^2 f(\bar{x})$  is the **Hessian** of  $f$  about point  $\bar{x}$  and is defined as:

$$\nabla^2 f(\bar{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (3.6)$$

Conveniently, the Hessian matrix is also symmetric.

## 3.3 Matrices as Linear Maps

**Linear Map Definition:** map  $f : V \rightarrow W$  is linear if

$$f(\alpha_1 x^{(1)} + \alpha_2 x^{(2)}) = \alpha_1 f(x^{(1)}) + \alpha_2 f(x^{(2)}) \quad (3.7)$$

**Key Finding:** **ANY LINEAR MAP** can be represented by a matrix  $\in \mathbb{R}^{m \times n}$  for map  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$

### 3.3.1 Matrix Inverse

A matrix is **NOT** invertable if  $m < n$  for map represented by  $A \in \mathbb{R}^{m \times n}$  due to the many-to-one mapping. It is **POTENTIALLY** invertible if  $m \geq n$ .

**Square Matrix Invertability:**  $A \in \mathbb{R}^{n \times n}$  is invertible iff there exists  $B \in \mathbb{R}^{n \times n}$  such that

$$AB = BA = I \quad (3.8)$$

If  $B$  exists, it is unique and is equal to  $A^{-1}$ .

**Invertability conditions:** If any of these hold true, then the matrix is invertible.

1.  $\det(A) \neq 0$
2. None of  $A$ 's eigenvalues are equal to zero.

**Invertible Matrix Properties:** Let  $A, B \in \mathbb{R}^{n \times n}$  both be invertible. Then:

1.  $(AB)^{-1} = B^{-1}A^{-1}$ .
2.  $(A^T)^{-1} = (A^{-1})^T$ .
3.  $\det(A) = \det(A^T) = \frac{1}{\det(A^{-1})}$

**Pseudo Inverse:** For non-square matrix there is no inverse, though the pseudo inverse  $A^{pi}$  satisfies:

$$AA^{pi}A = A \quad (3.9)$$

That is, either  $AA^{pi} = I_m$  or  $A^{pi}A = I_n$ .

**Case I:** Tall and thin matrix  $A$  with  $m > n$

- The “left inverse” exists iff the **columns** of the matrix are **linearly independent**.
- $A^{li}A = I_n$ .
- This only exists if the mapping of  $A$  is unique (i.e. it is a full-rank matrix).

**Case II:** Wide and fat matrix  $A$  with  $m < n$

- The “right inverse” exists iff the **rows** of the matrix are linearly independent.
- $AA^{ri} = I_m$ .
- If  $A^{ri}$  exists then  $A$  is “right invertible”.
- $A^{ri}y$  gives one of *infinitely many*  $x \in \mathbb{R}^n$  that satisfy  $Ax = y$ .

### 3.4 Orthogonal Matrices

**Orthogonality Conditions:**  $A \in \mathbb{R}^{n \times n}$  is orthogonal if:

1. Column vectors are all **mutually orthogonal**.
2. Column vectors are all **normalized**.

If we let  $A$  have columns  $q^{(i)}$  such that  $A = [q^{(1)} \dots q^{(n)}]$ , then

$$\langle q^{(i)}, q^{(j)} \rangle = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \quad (3.10)$$

It follows:

- $A^T A = I$
- $AA^T = I$
- $A^T = A^{-1}$

**Necessary and sufficient condition for orthogonality:**  $A^T = A^{-1}$ .

**Unitary Matrix:** If  $A \in \mathbb{C}^{n \times n}$  and all of the above apply, then:

$$U^H = \text{conjugate}(U^T)$$

**Geometric Implications of Orthogonality:** If  $Ax = y$  and  $A\tilde{x} = \tilde{y}$ :

- $\|x\| = \|y\|$  – That is, there is no **scaling** because all the columns of  $A$  are normalized.
- $\langle y, \tilde{y} \rangle = \langle x, \tilde{x} \rangle$  – that is, the **angles** between vectors are maintained.

Which implies that the transformation of  $A$  can **only** be composed of (1) rotation and (2) reflection.

### 3.5 Rank, Range, and Null Space

Let  $A \in \mathbb{R}^{m \times n}$ . Then  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  and  $A^T : \mathbb{R}^m \rightarrow \mathbb{R}^n$ .

**Domain and Range:**

- **Domain:**  $\text{dom}(\cdot)$  is the vector space the operator acts on.
- **Range:**  $\mathcal{R}\{\cdot\}$  is the **SPAN OF THE COLUMNS** of the matrix.

### How to Find Basis of $\mathcal{R}(A)$ :

1. Manipulate  $A$  into reduced row-echelon form.
2. We define **pivots** as the first non-zero entries of each row in the reduced row-echelon form.
3. Basis( $\mathcal{R}(A^T)$ ) are the **non-zero rows** of  $A_{REF}$ .
4. Basis( $\mathcal{R}(A)$ ) are the columns of  $A$  corresponding to the columns of the **pivots**.

$$\text{Thus } \dim(\mathcal{R}(A^T)) = \dim(\mathcal{R}(A))$$

### Nullspace of a Matrix

$$\mathcal{N}(A) = \{x | Ax = 0\} \quad (3.11)$$

It is the **orthogonal compliment** of  $\mathcal{R}(A^T)$ . To solve for this, you must start with  $Ax = 0$  and find the implied conditions on  $x$ .

### Fundamental Theorem of Linear Algebra

1.  $\mathcal{N}(A) \oplus \mathcal{R}(A^T) = \mathbb{R}^n$
2.  $\dim(\mathcal{N}(A)) + \text{rank}(A) = n$
3.  $\mathcal{R}(A) \oplus \mathcal{N}(A^T) = \mathbb{R}^m$
4.  $\text{rank}(A) + \dim(\mathcal{N}(A^T)) = m$

**Consequence of the Fundamental Theorem:** We can express any  $x \in \mathbb{R}^n$  as

$$x = A^T y + z \quad (3.12)$$

- $z \in \mathcal{N}(A)$
- $y \in \mathcal{R}(A^T)$

# Chapter 4

## Eigendecomposition

### 4.1 How to Solve for Eigenthings

$$Av = \lambda v$$

$$Av - I\lambda v = 0$$

$$\det(A - \lambda I) = 0 \tag{4.1}$$

Once you solve for the **characteristic polynomial roots**, you can plug them back into the initial formula and solve for the corresponding eigenvalue.

**Theorem 2** *All  $A \in \mathbb{R}^{n \times n}$  have  $\geq 1$  eigenvalue.*

- *Full set of eigenvectors has the same number of eigenvectors as eigenvalues.*
- *Partial set (defective matrix) has fewer eigenvectors than eigenvalues.*

**Multiplicity of Eigenvalues:** We have two types of multiplicity.

1. **Algebraic Multiplicity:**  $AM(\lambda_i) = \mu_i$  – number of times  $\lambda_i$  is repeated in the **characteristic polynomial**.
2. **Geometric Multiplicity:**  $GM(\lambda_i) = \nu_i = \dim(\mathcal{N}(A - \lambda_i I))$  – number of **eigenvectors** corresponding to the given eigenvalue.

$$1 \leq \nu_i = GM(\lambda_i) \leq AM(\lambda_i) = \mu_i$$



**Eigenspace:** For each  $\lambda$  there is a space  $E_\lambda$ :

$$E_\lambda = \text{span}\{u | Au = \lambda u\} = \mathcal{N}(A - I\lambda) \quad (4.2)$$

The geometric multiplicity of  $\lambda$  gives the dimensionality of the eigenspace. Each  $E_\lambda$  is invariant to  $A$ .

## 4.2 Diagonalization

**Condition for Diagonalizability:** If  $A$  has a **full set** of eigenvalues (i.e. distinct  $\lambda_i$  for  $i \in [k]$ ,  $k \leq n$ ) such that  $\nu_i = GM(\lambda_i) = AM(\lambda_i) = \mu_i$  for all  $i \in [k]$ , then  $A$  is diagonalizable.

This means that we can create  $n \times n$  matrix  $U$  with columns equal to eigenvectors of  $A$  and matrix  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  such that

$$A = U\Lambda U^{-1} \quad (4.3)$$

**Utility of Diagonalization:**

- We can take arbitrary powers of  $A$  very easily now by cancelling  $U$ ,  $U^{-1}$ .

**Determinants of Diagonal Matrices** If we can write  $A = U\Lambda U^{-1}$ , then

$$|\det A| = \left| \prod_{i=1}^n \lambda_i \right| \quad (4.4)$$

# Chapter 5

## Symmetric Matrices

### 5.1 Symmetric Matrices and Quadratic Functions

**Symmetric Matrix Definition:**  $S^n$  is the set of all symmetric matrices and is defined as

$$S^n = \{A \in \mathbb{R}^{n \times n} | A = A^T\} \quad (5.1)$$

**Quadratic Function Definition:**  $q : \mathbb{R}^n \rightarrow \mathbb{R}$  is a quadratic function if

$$q(x) = x^T A x + c^T x + d \quad (5.2)$$

where  $A \in \mathbb{R}^{n \times n}$ ,  $c \in \mathbb{R}^n$ ,  $d \in \mathbb{R}$ .

**Important Alternate Forms of Quadratics:**

$$q(x) = \frac{1}{2} x^T (A + A^T) x + c^T x + d \quad (5.3)$$

$$q(x) = \frac{1}{2} \begin{bmatrix} x^T & 1 \end{bmatrix} \begin{bmatrix} A + A^T & c \\ c^T & 2d \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \quad (5.4)$$

**Important note:** The 2x2 matrix there is  $\in S^{n+1 \times n+1}$ ! This is the strong connection between quadratics and symmetric matrices.

### 5.2 Spectral Theorem

*Tl;dr: Eigenthings are really nice and simple with symmetric matrices.*

**Theorem 3 *Nice things about symmetrical matrices:*** Let  $A \in S^n$  have  $n$  (potentially non-distinct) eigenvalues. Then

1.  $\lambda_i \in \mathbb{R}$  which means that the eigenvectors are **purely real**.
2.  $GM(\lambda_i) = AM(\lambda_i)$  which means there is a full set of eigenvalues  $\rightarrow A$  is diagonalizable!
3. All eigen spaces are **mutually orthogonal** (not just linearly independent).

**Spectral Decomposition:** If  $A \in S^n$  then there are  $n$  orthonormal eigenvectors. We let each eigenvector  $u^{(i)}$  be a column in  $U \in \mathbb{R}^{n \times n}$ . Then:

$$A = U \Lambda U^T \quad (5.5)$$

**Rayleigh Quotients:**

$$\lambda_{min} \leq \frac{x^T A x}{\|x\|^2} \leq \lambda_{max} \quad (5.6)$$

- The Rayleigh quotient when  $x = u^{(1)}$  (the maximum  $\lambda$  eigenvector) is equal to  $\lambda_{max}(A)$ .
- Same goes for when  $x = u^{(n)}$  with respect to the smallest eigenvalue  $\lambda_n$ .
- **This connects eigenvectors and singular values.**

**Positive Semi-Definite Matrices:** if

$$x^T A x \geq 0 \quad (5.7)$$

For all  $x \in \mathbb{R}^n$ , then  $A \in S_+^n$  ( $A$  is a positive semidefinite matrix).

**Positive Definite Matrices:** if strict inequality holds such that

$$x^T A x > 0 \quad (5.8)$$

for all  $x \in \mathbb{R}^n$ , then  $A$  is a positive definite matrix ( $A \in S_{++}^n$ ).

**Theorem 4 *Convenient things about  $S_+^n$ ,  $S_{++}^n$ :***

- $A \in S_+^n$  iff  $\lambda_i \geq 0$  for all  $i \in [n]$ .
- $A \in S_{++}^n$  iff  $\lambda_i > 0$  for all  $i \in [n]$ .
- Any positive definite matrix is invertible. Positive semi-definite matrices are invertible iff they are positive definite.

## 5.3 Ellipsoids

**Ellipsoid definition:** An ellipsoid is a set

$$\varepsilon = \{x \in \mathbb{R}^n | (x - x^{(0)})^T P^{-1} (x - x^{(0)}) \leq 1\} \quad (5.9)$$

Where  $P \in S_{++}^n$ ,  $x^{(0)} \in \mathbb{R}^n$

- The condition is a quadratic function!
- You can apply spectral decomposition to  $P^{-1}$ .

**Plotting Ellipses:**

$$\begin{aligned} 1 &\geq \bar{x}^T (U \Lambda U^T)^{-1} \bar{x} \\ &1 \geq \tilde{x}^T \lambda^{-1} \tilde{x} \\ &= \sum_{i=1}^n \left( \frac{\tilde{x}_i}{\sqrt{\lambda_i}} \right)^2 \end{aligned} \quad (5.10)$$

Where  $\bar{x} = (x - x^{(0)})$ ,  $\tilde{x} = U^T \bar{x}$ .

To plot:

1. Generate ‘plot’ in  $\tilde{x}$  space (note the key positions).
2. Unrotate by multiplying key position vectors by  $U$ .
3. Un-shift by adding  $x^{(0)}$

**Sample Mean and Covariance:** for  $m$  data vectors  $x^{(i)}$ ,  $i \in [m]$ :

1. **Sample Mean:**  $\hat{x} = \frac{1}{m} \sum_{i=1}^m x^{(i)}$
2. **Sample Covariance:**  $S = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \hat{x})(x^{(i)} - \hat{x})^T \in S^n$

**Data ellipse:**  $\varepsilon_\lambda = \{x \in \mathbb{R}^n | (x - \hat{x})^T S^{-1} (x - \hat{x}) < \gamma\}$  where  $\gamma$  is usually

1. This ellipse can help describe the geometry of the data.

## 5.4 Matrix Square Root and Cholesky Decomposition

**Matrix Square Root Definition:** For any  $A \in S_+^n$ :

$$A^{\frac{1}{2}} = U \Lambda^{\frac{1}{2}} U \quad (5.11)$$

Where  $\lambda = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$ .

- $A^{\frac{1}{2}}$  is a unique PSD matrix.
- $A$  is a PSD matrix **iff**  $A^{\frac{1}{2}}$  **is PSD**.

**Cholesky Decomposition:** Any PSD  $A$  can be factored as

$$A = B^T B \quad (5.12)$$

Where  $B = \Lambda^{\frac{1}{2}} U$  and  $B^T = U \Lambda^{\frac{1}{2}}$ .

$B$  also has a QR (orthogonal times upper triangular) decomposition, so:

$$A = B^T B = (QR)^T (QR) = R^T Q^T Q R = R^T R \quad (5.13)$$

**To create Choleski Decomposition:**

1. Select  $R$  such that diagonals  $R_{ii} > 0$ .
2. If  $A$  is PD, then **there exists unique**  $R$  which forms the Choleski Decomposition.
3. There exist non-unique  $R$  for  $A$  is PSD.

**Solving  $Ax = b$  with Choleski Decomposition:**

1.  $A = R^T R$
2.  $R^T R x = b$
3. Let  $y = R x$
4. **Forward Substitution:**
  - (a)  $y_1 R_{11} = b_1$  – Everything but  $y_1$  is known, so we solve for  $y_1$ !
  - (b)  $y_1 R_{21} + y_2 R_{22} = b_2$  – Everything but  $y_2$  is known, so we can solve for  $y_2$ !
  - (c) We continue this until we know the entire  $y$  vector!
5. **Back Substitution:**
  - (a) We use a similar procedure to solve from the bottom up for  $R x = y$

# Chapter 6

## Singular Value Decomposition

### 6.1 Motivation

**Theorem 5** We can write any linear map  $A \in \mathbb{R}^{m \times n}$  as

$$A = U\Sigma V^T \tag{6.1}$$

Where

- $U$  is  $m \times m$  orthonormal matrix composed of **left singular vectors**.
- $V$  is  $n \times n$  orthonormal matrix composed of **right singular vectors**.
- $\Sigma$  is  $m \times n$   $\text{diag}(\sigma_1, \dots, \sigma_r)$  singular values.

**Utility of SVD:**

1. Understand rank of relevant spaces to  $A$ .
2. Spectral norm (direction of maximum gain).
3. Orthonormal basis for  $\mathcal{N}(A), \mathcal{R}(A)$ .
4. Solve systems of linear equations relating to  $A$ .

### 6.2 Steps to Creating SVD

**Key Starting Ideas:** For  $m \times n$  real matrix  $A$ :

1.  $AA^T, A^T A$  are both PSD matrices.  $x^T A^T A x = \|Ax\|_2^2$ .
2.  $\text{rank}(A^T A) = \text{rank}(AA^T) = \text{rank}(A)$ .
3.  $AA^T$  and  $A^T A$  have the same  $r$  eigenvalues with  $r = \text{rank}(A)$ .

**Singular Value Definition:** The singular values of  $A$  are

$$\sigma_i = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(AA^T)} > 0 \quad (6.2)$$

Where  $\lambda_i(B)$  is the  $i$ th eigenvalue of matrix  $B$ .

**First  $r$  Right Singular Values of  $A$ :** We take the **spectral decomposition** of  $A^T A = V \Lambda V^T$ .

$$V = \left[ \begin{array}{c|ccc|c} \begin{array}{c} | \\ v^{(1)} \\ | \end{array} & \dots & \begin{array}{c} | \\ v^{(r)} \\ | \end{array} & \begin{array}{c} \text{vrule} \\ v^{(r+1)} \\ \text{vrule} \end{array} & \dots & \begin{array}{c} | \\ v^{(n)} \\ | \end{array} \end{array} \right] \quad (6.3)$$

- The first  $r$   $v^{(i)}$  relate to  $\lambda_i \neq 0$ .
  - **THESE ARE OUR RIGHT SINGULAR VECTORS.**
- They are all **mutually orthogonal**.
- The remaining vectors are associated with  $\lambda_i = 0$ .
  - We get to choose our own arbitrary basis for  $\mathcal{N}(A)$ .

**First  $r$  Left Singular Values of  $A$ :** The left singular vectors are denoted by  $u^{(i)}$  and are given by

$$u^{(i)} = A v^{(i)} \quad (6.4)$$

You could also get them by taking the spectral decomposition of  $AA^T$  but you run the risk of inverting the polarity (i.e.  $u^{(i)} \rightarrow -u^{(i)}$ ) which would mess up the entire decomposition. Just use the above formula to be safe and efficient.

Each  $u^{(i)}$  are associated with the same singular values  $\sigma_i = \sqrt{\lambda_i}$ .

**Compact SVD:** We can package our first  $r$  right ( $v$ ) and left ( $u$ ) singular vectors into  $V_r$  and  $U_r$  respectively to decompose  $A$  as follows:

$$A = U_r \Sigma V_r^T \quad (6.5)$$

Where  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ .

**Full SVD:** To get the ‘Full’ SVD, we add vectors to  $U_r, V_r$  such that they form bases for  $\mathbb{R}^m, \mathbb{R}^n$  respectively.

- $u^{(i)} \in \mathcal{N}(A^T)$  for all  $i \in \{r+1, \dots, m\}$
- $v^{(i)} \in \mathcal{N}(A)$  for all  $i \in \{r+1, \dots, n\}$
- Each of the above sets are **orthonormal bases** for their respective null spaces.

Now we have full SVD:

$$A = U\tilde{\Sigma}V^T \quad (6.6)$$

Where  $\tilde{\Sigma}$  is the original sigma padded with zeros such that it is an  $m \times n$  matrix.

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{m \times n} \quad (6.7)$$

## 6.3 Principal Component Analysis

**Goal:** Get minimum error when you project data vectors onto some affine set.

**Approach:** We can optimize the variance of projected data points onto a unit vector by aligning it such that it is an **eigenevalue** of the covariance matrix  $\frac{1}{m}X^T X$

## 6.4 SVD Matrix Properties

**Fundamental Subspaces from SVD:** Assuming that  $A = U\tilde{\Sigma}V^T$ ,  $U = [U_r, U_{nr}]$ ,  $V = [V_r, V_{nr}]$ :

1. Basis of  $\mathcal{R}(A)$ : columns of  $U_r$ .
2. Basis of  $\mathcal{N}(A)$ : columns of  $V_{nr}$ .
3. Basis of  $\mathcal{N}(A^T)$ : columns of  $U_{nr}$ .
4. Basis of  $\mathcal{R}(A^T)$ : columns of  $V_r$ .



## 6.5 Moore-Penrose Pseudo-Inverse

**MP Pseudo-Inverse Definition:** For rank  $r$  matrix  $A \in \mathbb{R}^{m \times n}$  with SVD  $A = U\tilde{\Sigma}V^T$ ,

$$A^\dagger = V\tilde{\Sigma}^\dagger U^T = V_r \Sigma^{-1} U_r^T \in \mathbb{R}^{n \times m} \quad (6.8)$$

Where

$$\tilde{\Sigma}^\dagger = \begin{bmatrix} \Sigma^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} \quad (6.9)$$

And  $\Sigma^{-1} = \text{diag}(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r})$

**Recall:**  $B$  is a pseudo inverse of  $A$  if  $ABA = A$ . Moore-Penrose gives proper inverse (true, left, right) given the dimensions of the matrix  $A$ .

# Chapter 7

## Systems of Linear Equations

### 7.1 Mathematical Setup

**General Form:**

$$Ax = y \tag{7.1}$$

- $A \in \mathbb{R}^{m \times n}$  is a set of **known** matrix coefficients.
- $x \in \mathbb{R}^n$  is the **unknown** vector we want to solve for.
- $y \in \mathbb{R}^m$  is the **known** constraint vector.

It is possible that there are no solutions, one solution, or infinite solutions.

**Existence of Solutions:** Since  $Ax \in \mathcal{R}(A)$ , the solution only exists for  $y \in \mathcal{R}(A)$ .

**Simple Test for Existence:** If

$$\text{Rank}([A \ y]) = \text{Rank}(A) \tag{7.2}$$

Then the solution exists. Otherwise, it doesn't (in this case,  $y$  adds to the column space and is therefore not in  $\mathcal{R}(A)$ ).

**Simple Test for Uniqueness:** If  $\mathcal{N}(A) = \{\emptyset\}$ , then the solution is unique.

**Consistent Solution to Least Squares:** Using the **Moore-Penrose pseudo inverse**  $A^\dagger$ , we have

$$x^* = A^\dagger y \quad (7.3)$$

This will always work, and in the underdetermined case, will produce the solution that minimizes  $\|x^*\|_2$ .

## 7.2 Over, Under, and Uniquely Determined Least Squares

**Theorem 6 Normal Equation:** *The optimal solution  $x^*$  to any (unconstrained) least squares problem satisfies*

$$A^T A x^* = A^T y \quad (7.4)$$

### 7.2.1 Overdetermined

For “tall and thin” matrix  $A \in \mathbb{R}^{m \times n}$  with  $m > n$ , we have the dimension of the column space of  $A$  (or  $\mathcal{R}(A)$ ) limited by  $n$ . Since the output of  $Ax$  is in  $\mathbb{R}^n$ , **most of the possible  $y$  vectors will not be in  $\mathcal{R}(A)$ .**

Therefore, we are forced to search for the **best solution**:

$$x^* = \arg \min_{x \in \mathbb{R}^n} \|Ax - y\| \quad (7.5)$$

Which looks an awful lot like the definition of **projection** of  $y \rightarrow \text{column span}(A)$ . We use the **normal equation** (seen above) to solve for this.

**Solving the Normal Equation** To solve  $A^T A x^* = A^T y$  we have several cases:

1.  $A^T A$  is invertible (i.e.  $A$  is full rank). In this case,

$$x^* = (A^T A)^{-1} A^T y \quad (7.6)$$

and we can project  $y$  onto the columns of  $A$  to get the *approximation of  $y$*  we achieve.

$$\hat{y}^* = A x^* \quad (7.7)$$

2.  $A^T A$  is **degenerate** (i.e. not full rank). In this case,

- One or more columns of  $A$  can be removed without detracting from  $\mathcal{R}(A)$  since columns are not linearly independent due to degeneracy.
- **Procedure:** Remove redundant columns from  $A$  and follow case 1 to generate  $x^*$  and  $\hat{y}^*$ .

**Using SVD to Solve Over Determined Least Squares:**  $A = U_r \Sigma V_r^T$  where  $U_r$  is a basis for the column space of  $A$ . Then

$$\hat{y}^* = \sum_{i=1}^r \langle u^{(i)}, y \rangle u^{(i)} \quad (7.8)$$

### 7.2.2 Uniquely Determined

If the solution is uniquely determined,  $A$  has full rank, is square, and therefore has an inverse.

$$x^* = A^{-1}y \quad (7.9)$$

### 7.2.3 Underdetermined

For  $A \in \mathbb{R}^{m \times n}$ , we have a “short and fat” matrix. In other words,  $m < n$ . Therefore, we have fewer constraints than we have manipulable variables. We have a **choice** between infinite  $x^*$ .

**Convention:** We use the *shortest*  $x^*$  vector (i.e. we minimize  $\|x^*\|_2$ .)

Our infinite set of possible  $x^*$  can be represented by the affine set where  $\bar{x}$  satisfies  $A\bar{x} = y$ .

$$\mathcal{A} = \bar{x} + \mathcal{N}(A) \quad (7.10)$$

We can then project  $y \rightarrow A$ .

**Solution to Underdetermined Case:**

$$x^* = A^T(AA^T)^{-1}y \quad (7.11)$$

With the norm of  $x^*$  given by the following

$$\|x^*\|_2^2 = \sum_{i=1}^r \frac{1}{\sigma_i^2} \langle u^{(i)}, y \rangle^2 \quad (7.12)$$

Where  $A = U_r \Sigma V_r^T$  is the SVD of  $A$ .

## 7.3 Regularized Least Squares

The idea is to introduce another term to the “cost function” for least squares. We *penalize* the solution  $x^*$ , generally based on some measure of its magnitude (e.g. an  $\ell_p$  norm, a weighted sum, etc.)

### 7.3.1 $\ell_2$ Regularization

$$x^* = \arg \min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 + \gamma \|x\|_2^2 \quad (7.13)$$

**Solving  $\ell_2$  Regularized Least Squares:**

1. Let

$$\bar{A} = \begin{bmatrix} A \\ \sqrt{\gamma} I_{n \times n} \end{bmatrix} \quad (7.14)$$

$$\bar{y} \in \mathbb{R}^n = \begin{bmatrix} y \\ 0 \\ 0 \\ \vdots 0 \end{bmatrix} \quad (7.15)$$

2. Solve as before, knowing that

$$\|Ax - y\|_2^2 + \gamma \|x\|_2^2 \equiv \|\bar{A}x - \bar{y}\|_2^2 \quad (7.16)$$

$$x^* = (\bar{A}^T \bar{A})^{-1} \bar{A}^T \bar{y} = (A^T A + \gamma I)^{-1} A^T y \quad (7.17)$$

## 7.4 Tikhonov Regularization

*Also known as ridge regression, weight decay.*

**Governing Formula:** We introduces matrices  $W_1, W_2$  to indicate “priority” of minimizing different error terms (and linear combinations of error terms) and variable  $x^{(0)}$  to represent the “zero effort point” as follows:

$$\min_x \|W_1(Ax - y)\|_2^2 + \|W_2(x - x^{(0)})\|_2^2 \quad (7.18)$$

- $W_1 \in S_{++}^n$ : How much do we weigh residuals (i.e. indices of  $Ax - y$ )?
- $W_2$ : How much do we weigh deviations of  $x^*$  from the **zero effort point**  $x^{(0)}$ ?

**Solution:** We find algebraic equivalence to the following unregularized least squares problem:

$$\min_x \|\bar{A}x - \bar{y}\|_2^2 \quad (7.19)$$

Where

$$\bar{A} = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} \quad (7.20)$$

$$\bar{y} = \begin{bmatrix} W_1 y \\ W_2 x^{(0)} \end{bmatrix} \quad (7.21)$$

## 7.5 Theoretical Aspects of Least Squares

For the solution of regularized least squares,

$$\min_x \|Ax - y\|_2^2 + \gamma \|x\|_2^2 \quad (7.22)$$

Our objective function can be re-written w.r.t.  $x$  as the following sum of constants and quadratics:

$$\|Ax_{LS}^* - y\|_2^2 + (x - x_{LS}^*)^T A^T A (x - x_{LS}^*) + \gamma x^T x \quad (7.23)$$

One can think of it as two “gravity wells” or “potential energy wells” that pull the optimal solution  $x^*$  towards them. One is centered at the solution to the un-regularized least squares problem and the other at the **zero effort point** (in the above equations, that is the origin). Since they are quadratic cost surfaces, they actually do have very similar to shape as gravity wells (other than their eccentricity).

### 7.5.1 $\ell_1$ Regularization

LASSO = “Least Absolute Shrinkage and Selection Operator” is the term for  $\ell_1$  regularization. It generally produces sparsity in approximations.

### 7.5.2 SVD Connection

*After much symbol-shunting of  $\ell_2$  regularized least squares,*

$$y^* = \sum_{i=1}^r \frac{\sigma_i^2}{\sigma_i^2 + \gamma} \langle u^{(i)}, y \rangle u^{(i)} \quad (7.24)$$