Preliminaries

Xiangyu Yang

March 15, 2020

1 Preface

This note can only be used in **SI 151: Optimization and Machine Learning**, which aims at introducing some preliminaries for students who participate in this class. In particular, most of the contents (e.g., words, figures) come from [1, 2] and the book Introduction to Probability, 2nd Edition. We are very grateful to these authors. On the other hand, I must declare that this note is prohibited from distributing. In other words, one can not casually share the copy of this note with others due to te additional restrictions.

We

- Introduce some necessary preliminaries of Linear algebra used in class. In particular, the singular value decomposition is emphasized.
- Introduce some necessary preliminaries of probability and statistics used in class. In particular, the MLE and MAP rule are emphasized.

2 Some Common Concepts In Linear Algebra

2.1 Vector Norm

Definition 1 (General Vector Norms)**.** *A norm for a real or complex vector space $\mathcal{V}$ is a function $\|\cdot\|$ mapping $\mathcal{V}$ into $\mathcal{R}$ that satisfies the following conditions.*

$$\begin{aligned} &\|\boldsymbol{x}\| \geq 0 \text{ and } \|\boldsymbol{x}\| = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{0}, && (\textit{Nonnegativity and Definiteness}) \\ &\|\alpha\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\| \text{ for all scalars } \alpha, && (\textit{Nonnegative homogeneity}) \\ &\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|. && (\textit{Triangle inequality}) \end{aligned} \tag{1}$$

Typically, the p-norm ($p \geq 1$) of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. In particular

- **The Euclidean norm** is the case that $p = 2$: $\|\boldsymbol{x}\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2} = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$.
- **The grid norm** is the case that $p = 1$: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$.
- **The max norm** is the case that $p = \infty$: $\|\boldsymbol{x}\|_\infty = \lim_{p\to\infty} \|\boldsymbol{x}\|_p = \lim_{p\to\infty} (\sum_{i=1}^n |x_i|^p)^{1/p} = \max_i |x_i|$.

Moreover, for $p \in [1, \infty]$, $\|\boldsymbol{x}\|_p$ is a norm. While for $p \in (0, 1)$, $\|\cdot\|_p$ defines a quasi-norm since it only satisfies the triangle inequality. Further, we call $\|\boldsymbol{x}\|_0$ as 0-norm, which counts the number of nonzero entries in $\boldsymbol{x}$. Obviously, $\|\boldsymbol{x}\|_0$ does not satisfy the definition of vector norm.

2.2 Matrix Norm

Likewise, we have encountered some matrix norm in class. For example,

Definition 2 (Frobenius Matrix Norm)**.** *The Frobenius norm of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is defined by*

$$\|\boldsymbol{A}\|_F^2 = \sum_{i,j} |a_{ij}|^2 = trace(\boldsymbol{A}^T\boldsymbol{A}), \tag{2}$$

where $trace(\boldsymbol{A}^T\boldsymbol{B}) = \langle \boldsymbol{A}, \boldsymbol{B}\rangle$.

Moreover, here are some induced matrix norms listed below:

- The matrix norm induced by the Euclidean vector norm is
$$\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{Ax}\|_2 = \sqrt{\lambda_{\max}}, \tag{3}$$
where $\lambda_{\max}$ is the largest number λ such that $\boldsymbol{A}^T\boldsymbol{A} - \lambda\boldsymbol{I}$ is singular.

- The matrix norms induced by the vector 1-norm is
$$\|\boldsymbol{A}\|_1 = \max_{\|\boldsymbol{x}\|_1=1} \|\boldsymbol{Ax}\|_1 = \max_j \sum_i |a_{ij}| = \text{ the largest absolute column sum.} \tag{4}$$

- The matrix norms induced by the vector ∞-norm is
$$\|\boldsymbol{A}\|_\infty = \max_{\|\boldsymbol{x}\|_\infty=1} \|\boldsymbol{Ax}\|_\infty = \max_i \sum_j |a_{ij}| = \text{ the largest absolute row sum.} \tag{5}$$

2.3 Eigenvalues and Eigenvectors

Definition 3. *Let $\boldsymbol{A}$ be an $n \times n$ matrix.*

1. *An **eigenvector** of $\boldsymbol{A}$ is a nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{Ax} = \lambda\boldsymbol{x}$, for some scalar λ.*

2. *An **eigenvalue** of $\boldsymbol{A}$ is a scalar such that the equation $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ has a nontrivial solution.*

If $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ for $\boldsymbol{x} \neq \boldsymbol{0}$, we say that λ is the eigenvalue for $\boldsymbol{x}$, and that $\boldsymbol{x}$ is an eigenvector for λ.

Geometrically, $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ says that under **transformation** (e.g., **projection**, **reflection**, **rotation**, et al) by $\boldsymbol{A}$, eigenvectors experience only changes in magnitude or sign—the orientation of $\boldsymbol{Ax}$ in $\mathbb{R}^n$ is the same as that of $\boldsymbol{x}$. The eigenvalue λ is simply the amount of "stretch" or "shrink" to which the eigenvector $\boldsymbol{x}$ is subjected when transformed by $\boldsymbol{A}$ (see Fig. 1). For example,

- The projection matrix $P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 0$.

- The reflection matrix $R = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1$.

- The rotation matrix $\bar{R} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ for 90° counter-clockwise rotation has eigenvalues $\lambda_1 = i$ and $\lambda_2 = -i$.

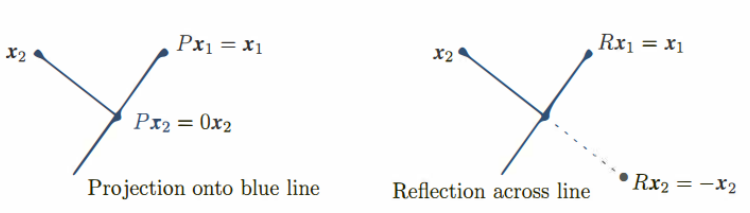

$\boldsymbol{x}_2$
$P\boldsymbol{x}_1 = \boldsymbol{x}_1$
$P\boldsymbol{x}_2 = 0\boldsymbol{x}_2$
Projection onto blue line
$\boldsymbol{x}_2$
$R\boldsymbol{x}_1 = \boldsymbol{x}_1$
Reflection across line
$R\boldsymbol{x}_2 = -\boldsymbol{x}_2$

Figure 1: Projections P have eigenvalues 1 and 0. Reflections R have $\lambda = 1$ and -1. A typical $\boldsymbol{x}$ changes direction, but an eigenvector stays along the same line.

2.4 Positive Definite Matrices

Definition 4. *For real-symmetric matrices $\boldsymbol{A}$, we define the positive definite matrix as follows*

- $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ *for every nonzero* $\boldsymbol{x} \in \mathbb{R}^n$*.*
- *All eigenvalues of* $\boldsymbol{A}$ *are positive.*
- $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ *for some nonsingular* $\boldsymbol{B}$*. (*$\boldsymbol{B}$ *is not unique!)*

2.5 Diagonalizing

Diagonalizing a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is closely related to the problem of finding the eigenvalues and eigenvectors of a matrix.

Theorem 1 (The Spectral Theorem)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$*. Then* $\boldsymbol{A}$ *is symmetric if and only if* $\boldsymbol{A}$ *is orthogonally diagonalizable.*

The above theorem states that there is an orthogonal matrix $\boldsymbol{V}$ and a diagonal $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$ whenever $\boldsymbol{A}$ is symmetric real square matrix. Here the columns of $\boldsymbol{V}$ are eigenvectors for $\boldsymbol{A}$ and form an orthonormal basis for $\mathbb{R}^n$, and the diagonal entries of $\boldsymbol{D}$ are the eigenvalues of $\boldsymbol{A}$.

Remark 1. *Since* $\boldsymbol{V}^T = \boldsymbol{V}^{-1}$ *for orthogonal* $\boldsymbol{V}$*, the equality* $\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D}$ *is the same as* $\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D}$*, so* $\boldsymbol{A} \sim \boldsymbol{D}$*, so this a special case of diagonalization: the diagonal entries of* $\boldsymbol{D}$ *are eigenvalues of* $\boldsymbol{A}$*, and the columns of* $\boldsymbol{V}$ *are corresponding eigenvectors. The only difference is the additional requirement that* $\boldsymbol{V}$ *be orthogonal, which is equivalent to the fact that those eigenvectors (columns of* $\boldsymbol{V}$*) form an orthonormal basis of* $\mathbb{R}^n$*.*

3 Singularly Value Decomposition

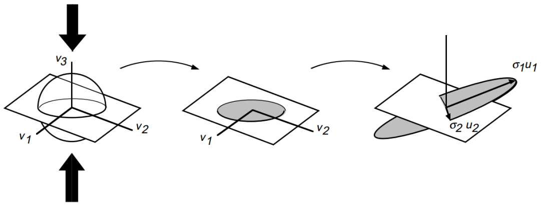
The SVD is intimately related to the familiar theory of diagonalizing a symmetric matrix. To emphasize the connection of with the SVD, we here refer to $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$ as the eigenvalue decomposition, or EVD, for $\boldsymbol{A}$.

Definition 5 (SVD, refer to https://en.wikipedia.org/wiki/Singular_value_decomposition)**.** *Suppose* $\boldsymbol{A}$ *is a* $m \times n$ *matrix whose entries come from the* $\mathbb{R}$*, which is either the field of real numbers or the field of complex numbers. Then there exists a factorization, called a `singular value decomposition´of* $\boldsymbol{A}$*, of the form*

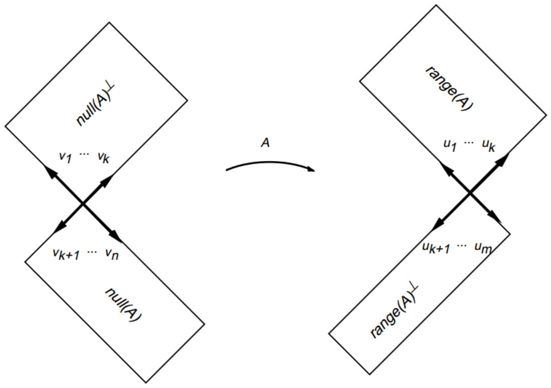
$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

where

- $\boldsymbol{U}$ *is an* $m \times m$ *orthogonal matrix;*
- $\boldsymbol{\Sigma}$ *is a diagonal* $m \times n$ *matrix with non-negative real numbers on the diagonal;*
- $\boldsymbol{V}$ *is an* $n \times n$ *orthogonal matrix.*



(a) How $\boldsymbol{A}$ deforms $\mathbb{R}^n$



(b) Strang's Diagram

The analogy between the EVD for a symmetric matrix and SVD for an arbitrary matrix can be extended a little by thinking of matrices as linear transformations. In particular, most of our notes come from article [1].

- For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the transformation takes $\mathbb{R}^n$ to itself, and the columns of $\boldsymbol{V}$ define an especially nice basis. What does this mean?

  When vectors are expressed relative to this basis, we see that the transformation simply dilates some components and contracts others, according to the magnitudes of the eigenvalues. Moreover, the basis is orthonormal, which is the best kind of basis to have.

- For an arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the transformation takes $\mathbb{R}^n$ to a diferent space, $\mathbb{R}^m$. Is there any a natural basis for each of domain and range? Yes, the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$ provide these bases.

  It simply dilates some components and contracts others, according to the magnitudes of the singular values, and possibly discards components or appends zeros as needed to account for a change in dimension.

Select an orthonormal basis $\{v_1, v_2, \cdots, v_n\}$ for $\mathbb{R}^n$ so that the first k elements span the row space of $\boldsymbol{A}$ and the remaining $n-k$ elements span the null space of $\boldsymbol{A}$, where k is the rank of $\boldsymbol{A}$. Then for $1 \le i \le k$ define u_i to be a unit vector parallel to $\boldsymbol{A}v_i$, and extend this to a basis for $\mathbb{R}^m$. Relative to these bases, $\boldsymbol{A}$ will have a diagonal representation. Here is a question: **But in general, although the v's are orthogonal, is there any reason to expect the u's to be?**

We first show that the EVD of the $n \times n$ symmetric matrix $\boldsymbol{A}^T\boldsymbol{A}$ provides just such a basis, namely, the eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$.

Proof. Let $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$, with the diagonal entries λ_i of $\boldsymbol{D}$ arranged in nonincreasing order, and let the columns of $\boldsymbol{V}$ (which are eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$) be the orthogonal basis $\{v_1, v_2, \cdots, v_n\}$. Then

$$\boldsymbol{A}v_i \cdot \boldsymbol{A}v_j = (\boldsymbol{A}v_i)^T(\boldsymbol{A}v_j) = v_i^T\boldsymbol{A}^T\boldsymbol{A}v_j = v_i^T(\lambda_j v_j) = \lambda_j v_i \cdot v_j, \tag{6}$$

so the image set $\{\boldsymbol{A}v_1, \cdots, \boldsymbol{A}v_n\}$ is orthogonal, and the nonzero vectors in this set form a basis for the range of $\boldsymbol{A}$. □

Two points to emphasize

- The orthogonality of the v-basis is preserved under $\boldsymbol{A}$.
- The eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$ and their images under $\boldsymbol{A}$ provide orthogonal bases allowing $\boldsymbol{A}$ to be expressed in a diagonal form.

Next, we will do the following procedures to complete the construction.

1. Normalize the vectors $\boldsymbol{A}v_i$. Taking $i = j$ in (6) gives $\|\boldsymbol{A}v_i\|^2 = \lambda_i$, which means $\lambda_i \geq 0$. Notice that the assumptions that these eigenvalues were arranged in nonincreasing order, we conclude that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and, $\lambda_i = 0, \forall i > k$. The orthonormal basis for the range is therefore defined by

$$u_i = \frac{\boldsymbol{A}v_i}{\|\boldsymbol{A}v_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}v_i, \quad 1 \leq i \leq k. \tag{7}$$

   If $k < m$, we extend this to an orthonormal basis for $\mathbb{R}^m$. This completes the construction of the desired orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^m$.

2. Setting $\sigma_i = \sqrt{\lambda_i}$, we have $\boldsymbol{A}v_i = \sigma_i u_i$ for all $i \leq k$. Assembling the v_i as the columns of a matrix $\boldsymbol{V}$ and the u_i to form $\boldsymbol{U}$, this shows that $\boldsymbol{AV} = \boldsymbol{U\Sigma}$. Hence, $\boldsymbol{A} = \boldsymbol{U\Sigma V}^T$, which is the singular value decomposition of $\boldsymbol{A}$.

In summary,

- The matrix $\boldsymbol{V}$ is obtained from the diagonal factorization $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{VDV}^T$, in which the diagonal entries of $\boldsymbol{D}$ appear in non-increasing order;

- The columns of $\boldsymbol{U}$ come from normalizing the nonvanishing images under $\boldsymbol{A}$ of the columns of $\boldsymbol{V}$, and extending (if necessary) to an orthonormal basis for $\mathbb{R}^m$;

- The nonzero entries of $\boldsymbol{\Sigma}$ are the respective square roots of corresponding diagonal entries of $\boldsymbol{D}$.

Example 1. *Consider the SVD of the following* 4×5 *matrix*

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

Solution 1. *First of all, rank*$(M) = 3$.

1. *Compute* V.

$$M^TM = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 4 \end{bmatrix},$$

   we obtain $\lambda_1 = 9, \lambda_2 = 5, \lambda_3 = 4, \lambda_4 = 0, \lambda_5 = 0$, *and the corresponding eigenvectors are* $v_1 = [0\ 0\ 1\ 0\ 0]^T$, $v_2 = [0.4472\ 0\ 0\ 0\ 0.8944]^T$, $v_3 = [0\ 1\ 0\ 0\ 0]^T$, $v_4 = [-0.8944\ 0\ 0\ 0\ -0.4472]^T$, $v_5 = [0\ 0\ 0\ 1\ 0]^T$. *Thus,*

$$V^T = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0.4472 & 0 & 0 & 0 & 0.8944 \\ 0 & 1 & 0 & 0 & 0 \\ -0.8944 & 0 & 0 & 0 & 0.4472 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

2. *According to* (7), *we have* $u_1 = [0\ 1\ 0\ 0]^T$, $u_2 = [1\ 0\ 0\ 0]^T$, $u_3 = [0\ 0\ 0\ 1]^T$, $u_4 = [0\ 0\ 1\ 0]^T$. *Thus*

$$U = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

3.

$$\Sigma = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{5} & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

4 Main Concepts on Probability and Statistic

4.1 Random Variables

Let Ω denote a sample space. Starting with a probability model of an experiment:

- A random variable is a real-valued function of the outcome of the experiment.
- A function of a random variable defines another random variable.

4.2 Conditional Probability

Conditional probability provides us with a way to reason about the outcome of an experiment, based on partial information. Specifically,

Definition 6. *The condition probability of an event A, given an event B with $P(B) > 0$, is defined by*

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

and specifies a new (conditional) probability law on the same sample space Ω.

4.3 Total Probability Theorem and Bayes' Rule

Definition 7 (Total Probability Theorem)**.** *Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, \ldots, A_n$) and assume that $P(A_i) > 0$ for all i. Then, for any event B, we have*

$$\begin{aligned} P(B) &= P(A_1 \cap B) + \cdots + P(A_n \cap B) \\ &= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n). \end{aligned} \tag{9}$$

Remark 2. *We have the following comments:*

- *Partition the sample space in to a number of events A_i.*
- *Calculate a weighted average of condition probability under each event.*

The total probability theorem is often used in conjunction with the Bayes' rule.

Definition 8 (Bayes' Rule)**.** *Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $P(A_i) > 0$ for all i. Then, for any event B such that $P(B) > 0$, we have*

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i)P(B|A_i)}{P(B)} \\ &= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}. \end{aligned} \tag{10}$$

Bayes' rule is often used for **inference**. Specifically, we observe the effect, and we wish to infer the cause. For example, given that the effect B has been observed, we wish to evaluate the probability $P(B|A_i)$ that the cause A_i is present. In particular, we refer to

- $P(A_i)$ as the **prior probability**, which would express one's beliefs about this quantity before some evidence is taken into account.
- $P(B|A_i)$ as the **posterior probability** of event A_i given the information.

5 Bayesian Statistical Inference and Classical Statistical Inference

Within the field of statistics there are two prominent schools of thought, with opposing views: the **Bayesian** and the **classical** (also called **frequentist**). **Their fundamental difference relates to the nature of the unknown models or variables**.

- In the Bayesian view. the unknown variables are treated as random variables with known distributions. In particular,
  - when trying to infer the nature of an unknown model, it views the model as chosen randomly from a given model class. This is done by introducing a random variable Θ that characterizes the model, and by postulating a prior probability distribution $P_\Theta(\theta)$.
  - Given the observed data $\boldsymbol{x}$, one can, in principle, use Bayes' rule to derive a posterior probability distribution $P_{\Theta|X}(\theta|x)$. This captures all the information that x can provide about θ.
- In the classical view, the variables are treated as deterministic quantities that happen to be unknown. In particular,
  - we are not dealing with a single probabilistic model, but rather with multiple candidate probabilistic models, one for each possible value of θ.

5.1 Bayesian Inference and the Posterior Distribution

In Bayesian inference, the unknown quantity of interest, which is denoted by Θ, is modeled as a single random variable. **Our goal** is to extract information about Θ based on observing a collection $X = (X_1, \cdots, X_n)$ of observations measurements. We assume that we know

- A prior distribution P_Θ or f_Θ.
- A conditional distribution $P_{X|\Theta}$ or $f_{X|\Theta}$.

Once a particular value x of X has been observed, a complete answer to the Bayesian inference problem is provided by the posterior distribution $P_{\Theta|X}(\theta|x)$ or $f_{\Theta|X}(\theta|x)$ of Θ (see Fig. 2),

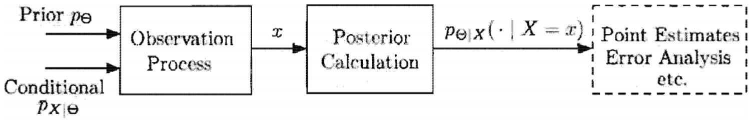


Figure 2: A summary of a Bayesian inference model.

5.2 Maximum A Posterior probability (MAP) rule

Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $P_{\Theta|X}(\theta|x)$ (if Θ is discrete) or $f_{\Theta|X}(\theta|x)$ (if Θ is continuous) (see Fig. 3):

$$\begin{aligned} \hat{\theta} &= \text{argmax}_\theta \; P_{\Theta|X}(\theta|x) \quad (\Theta \text{ discrete}), \\ \hat{\theta} &= \text{argmax}_\theta \; f_{\Theta|X}(\theta|x) \quad (\Theta \text{ continuous}) \end{aligned} \tag{11}$$

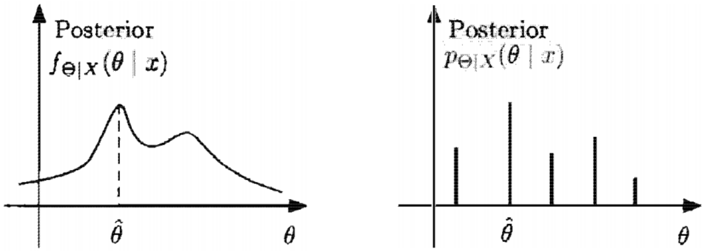


Figure 3: Illustration of the MAP rule for inference of a continuous parameter and a discrete parameter.

5.3 Classical Parameter Estimation

In contrast to the Bayesian view, the classical view the unknown parameter θ as a deterministic (not random) but unknown quantity. Instead of working within a single probabilistic model, we will be dealing simultaneously with multiple candidate models. one model for each possible value of θ (see Fig. 4). In particular,

- The observation X is random.
- A "good" hypothesis testing or estimation procedure will be one that possess certain desirable properties under every candidate model, that is, for every possible value of θ.

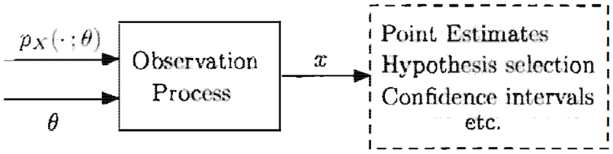


Figure 4: A summary of a classical inference model.

5.4 Maximum Likelihood Estimation

Let the vector of observations $X = (X_1, \cdots, X_n)$ be described by a joint PMF $P_X(x;\theta)$ or PDF $f_X(\boldsymbol{x};\theta)$.

- The maximum likelihood estimate (MLE) is a value of e that maximizes the likelihood function, $P_X(\boldsymbol{x};\theta)$ or PDF $f_X(\boldsymbol{x};\theta)$, overall all θ (see Fig. 5):

$$\begin{aligned} \hat{\theta} &= \text{argmax}_\theta \; P_X(x_1, \cdots, x_n;\theta), \\ \hat{\theta} &= \text{argmax}_\theta \; f_X(x_1, \cdots, x_n;\theta). \end{aligned} \tag{12}$$

We refer to $P_X(x;\theta)$ (or $f_X(x;\theta)$ if X is continuous) as the likelihood function.

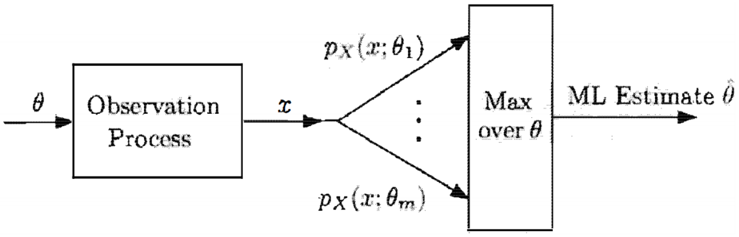


Figure 5: Illustration of MLE, assuming X is discrete and θ takes one of the m values $\theta_1, \cdots, \theta_m$.

- If the observations X_i are assumed to be **independent**, in which case, the likelihood function is of the form

$$P_X(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} P_{X_i}(x_i; \theta)$$
$$f_X(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta)$$

  for discrete X_i and continuous X_i, respectively. For computational convenience, we solve the following **log-likelihood function** in terms of the discrete and continuous X correspondingly,

$$\log P_X(x_1, \cdots, x_n; \theta) = \log \prod_{i=1}^{n} P_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log P_{X_i}(x_i; \theta),$$
$$\log f_X(x_1, \cdots, x_n; \theta) = \log \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i; \theta)$$

  over θ.

Remark 3. *The term "likelihood" needs to be interpreted properly:*

- *Having observed the value x of X, $P_X(x; \theta)$ is the probability that the observed value x can arise when the parameter is equal to θ.*
- *Recall that in Bayesian MAP estimation, the estimate is chosen to maximize the expression $P_\Theta(\theta)P_{X|\Theta}(x|\theta)$ over all θ, where P_Θ is the prior PMF of an unknown discrete parameter θ. Thus, if we view $P_X(x; \theta)$ as a conditional PMF, we may interpret MLE as MAP estimation with a **flat prior**, i.e., a prior which is the same for all θ, indicating the absence of any useful prior knowledge.*

References

[1] Dan Kalman. A singularly valuable decomposition: the SVD of a matrix. *The college mathematics journal*, 27(1):2–23, 1996.

[2] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. SIAM, 2000.