Part VI

Mathematics for Machine Learning

18
Probability and Statistics

18.1 Probability and Event

18.1.1 Sample Space and Event

Probability is related to the uncertainty and randomness of the world. It measures the likelihood of some outcome or event happening. To formalize this concept, we introduce the following definition:

Definition 18.1.1 (Sample Space and Event). *A set S of all possible outcomes of a random phenomenon in the world is called a **sample space**. A subset $A \subset S$ is called an **event**.*

Example 18.1.2. *The sample space of "the outcome of tossing two dice" is the set $S = \{(1,1),(1,2),\ldots,(6,6)\}$ of 36 elements. The event "the sum of the numbers on the two dice is 5" is the subset $A = \{(1,4),(2,3),(3,2),(4,1)\}$ of 4 elements.*

18.1.2 Probability

Given a sample space S, we define a probability for each event A of that space. This probability measures the likelihood that the outcome of the random phenomenon belongs A.

Definition 18.1.3 (Probability). *A **probability** $\Pr : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is a mapping from each **event** $A \subset S$ to a non-negative real number $\Pr[A] \geq 0$ such that the following properties are satisfied:*

1. $0 \leq \Pr[A] \leq 1$ *for any* $A \subset S$
2. $\Pr[S] = 1$
3. *For any countable collection $\{A_1, A_2, \ldots\}$ of events that are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset$ for any $i \neq j$),*

$$\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \Pr[A_i]$$

When the sample space is finite or countably infinite, [1] *the properties above can be simplified into the following condition:*

$$\sum_{x \in S} \Pr[\{x\}] = 1$$

Example 18.1.4. *Consider the sample space of "the outcome of tossing two dice" again. Assuming the two dice are fair, the probability of each outcome can be defined as* $1/36$*. Then the probability of the event "the sum of the numbers on the two dice is* 5*" is* $4/36$*.*

Example 18.1.5. *We are picking a point uniformly at random from the sample space* $[0,2] \times [0,2]$ *in the Cartesian coordinate system. The probability of the event that the point is drawn from the bottom left quarter* $[0,1] \times [0,1]$ *is* $1/4$*.*

18.1.3 Joint and Conditional Probability

In many cases, we are interested in not just one event, but multiple events, possibly happening in a sequence.

Definition 18.1.6 (Joint Probability)**.** *For any set of events* $\mathcal{A} = \{A_1, \ldots, A_n\}$ *of a sample space* S*, the* ***joint probability*** *of* $\mathcal{A}$ *is the probability* $\Pr[A_1 \cap \ldots \cap A_n]$ *of the intersection of all of the events. The probability* $\Pr[A_i]$ *of each of the events is also known as the* ***marginal probability****.*

Example 18.1.7. *Consider the sample space of "the outcome of tossing two dice" again. Let* A_1 *be the event "the number on the first die is* 1*" and let* A_2 *be the event "the number on the second die is* 4*." The joint probability of* A_1, A_2 *is* $1/36$*. The marginal probability of each of the events is* $1/6$*.*

It is also useful to define the probability of an event A, based on the knowledge that other events $A_1, \ldots, A_n$ have occurred.

Definition 18.1.8 (Conditional Probability)**.** *For any event* A *and any set of events* $\mathcal{A} = \{A_1, \ldots, A_n\}$ *of a sample space* S*, where* $\Pr[A_1 \cap \ldots \cap A_n] > 0$*, the* ***conditional probability of*** A ***given*** $\mathcal{A}$ *is*

$$\Pr[A \mid A_1, \ldots, A_n] = \frac{\Pr[A \cap A_1 \cap \ldots \cap A_n]}{\Pr[A_1 \cap \ldots \cap A_n]}$$

Example 18.1.9. *Consider the sample space of "the outcome of tossing two dice" again. Let* A_1 *be the event "the number on the first die is* 1*" and let* A_2 *be the event "the sum of the numbers on the two dice is* 5*." The conditional probability of* A_1 *given* A_2 *is* $1/4$*. The conditional probability of* A_2 *given* A_1 *is* $1/6$*.*

Using the definition of a conditional probability, we can define a formula to find the joint probability of a set $\mathcal{A}$ of events of a sample space.

[1] A countably infinite set refers to a set whose elements can be numbered with indices. The set $\mathbb{N}$ of natural numbers or the set $\mathbb{Q}$ of rational numbers are examples of countably infinite sets.

Proposition 18.1.10 (Chain Rule for Conditional Probability). *Given a set $\mathcal{A} = \{A_1, \ldots, A_n\}$ of events of a sample space S, where all appropriate conditional probabilities are defined, we have the following*

$$\begin{aligned}\Pr[A_1 \cap \ldots \cap A_n] &= \Pr[A_1 \mid A_2 \cap \ldots \cap A_n] \cdot \Pr[A_2 \cap \ldots \cap A_n] \\ &= \Pr[A_1 \mid A_2 \cap \ldots \cap A_n] \cdot \Pr[A_2 \mid A_3 \cap \ldots \cap A_n] \cdot \Pr[A_3 \cap \ldots \cap A_n] \\ &\vdots \\ &= \Pr[A_1 \mid A_2 \cap \ldots \cap A_n] \cdot \Pr[A_2 \mid A_3 \cap \ldots \cap A_n] \cdots \Pr[A_n]\end{aligned}$$

Finally, from the definition of a conditional probability, we see that

$$\Pr[B \mid A] \Pr[A] = \Pr[A \cap B] = \Pr[A \mid B] \Pr[B]$$

This shows that

$$\Pr[B \mid A] = \frac{\Pr[A \mid B] \Pr[B]}{\Pr[A]}$$

This is known as the *Bayes's Rule*.

18.1.4 Independent Events

Definition 18.1.11 (Independent Events). *Two events A, B are **independent** if $\Pr[A], \Pr[B] > 0$ and*

$$\Pr[A] = \Pr[A \mid B]$$

or equivalently

$$\Pr[B] = \Pr[B \mid A]$$

or equivalently

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

Example 18.1.12. *Consider the sample space of "the outcome of tossing two dice" again. Let A_1 be the event "the number on the first die is 1" and let A_2 be the event "the number on the second die is 4." A_1 and A_2 are independent.*

Example 18.1.13. *Consider the sample space of "the outcome of tossing two dice" again. Let A_1 be the event "the number on the first die is 1" and let A_2 be the event "the sum of the numbers on the two dice is 5." A_1 and A_2 are not independent.*

18.2 Random Variable

In the previous section, we only learned how to assign a probability to an event, a subset of the sample space. But in general, we can assign a probability to a broader concept called a *random variable*, associated to the sample space.

Definition 18.2.1 (Random Variable)**.** *Given a sample space S, a mapping $X : S \to \mathbb{R}$ that maps each outcome $x \in S$ to a value $i \in \mathbb{R}$ is called a **random variable**.*

Example 18.2.2. *Consider the sample space of "the outcome of tossing two dice" again. Then the random variable $X =$ "sum of the numbers on the two dice" maps the outcome $(1,4)$ to the value 5.*

Definition 18.2.3 (Sum and Product of Random Variables)**.** *If $X, X_1, \ldots, X_n$ are random variables defined on the same sample space S such that $X(x) = X_1(x) + \ldots + X_n(x)$ for every outcome $x \in S$, then we say that X is the **sum** of the random variables $X_1, \ldots, X_n$ and denote*

$$X = X_1 + \ldots + X_n$$

*If $X(x) = X_1(x) \times \ldots \times X_n(x)$ for every outcome $x \in S$, then we say that X is the **product** of the random variables $X_1, \ldots, X_n$ and denote*

$$X = X_1 \cdots X_n$$

Example 18.2.4. *Consider the sample space of "the outcome of tossing two dice" again. Then the random variable $X =$ "sum of the numbers on the two dice" is the sum of the two random variables $X_1 =$ "the number on the first die" and $X_2 =$ "the number on the second die."*

18.2.1 Probability of Random Variable

There is a natural relationship between the definition of an event and a random variable. Given a sample space S and random variable $X : S \to \mathbb{R}$, the "event that X takes a value in B" is denoted $\Pr[X \in B]$. It is the total probability of all outcomes $x \in S$ such that $X(x) \in B$. In particular, the event that X takes a particular value $i \in \mathbb{R}$ is denoted as $X = i$ and the event that X takes a value in the interval $[a, b]$ is denoted as $a \leq X \leq b$ and so on.

Example 18.2.5. *Consider the sample space of "the outcome of tossing two dice" and the random variable $X =$ "sum of the numbers on the two dice" again. Then*

$$\Pr[X = 5] = \Pr[\{(1,4),(2,3),(3,2),(4,1)\}] = 4/36$$

Often we are interested in the probability of the events of the form $X \leq x$. Plotting the values of $\Pr[X \leq x]$ with respect to x completely identifies the *distribution* of the values of X.

Definition 18.2.6 (Cumulative Distribution Function)**.** *Given a random variable X, there is an associated **cumulative distribution function (cdf)** $F_X : \mathbb{R} \to [0,1]$ defined as*

$$F_X(x) = \Pr[X \leq x]$$

Proposition 18.2.7. *The following properties hold for a cumulative distribution function F_X:*

1. *F_X is increasing*

2. $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$

18.2.2 Discrete Random Variable

If the set of possible values of a random variable X is finite or countably infinite, we call it a *discrete random variable*. For a discrete random variable, the probability $\Pr[X = i]$ for each value i that the random variable can take completely identifies the *distribution* of X. In view of this fact, we denote the *probability mass function (pmf)* by

$$p_X(i) = \Pr[X = i]$$

Proposition 18.2.8. *The following properties hold for a probability mass function p_X:*

1. $\sum_i p_X(i) = 1$

2. $F_X(x) = \sum_{i \leq x} p_X(i)$

18.2.3 Continuous Random Variable

We now consider the case where the set of all possible values of a random variable X is an interval or a disjoint union of intervals in $\mathbb{R}$. We call such X a *continuous random variable*. In this case, the probability of the event $X = i$ is zero for any $i \in \mathbb{R}$. Instead, we care about the probability of the events of the form $a \leq X \leq b$.

Definition 18.2.9 (Probability Density Function)**.** *Given a continuous random variable X, there is an associated **probability density function (pdf)** $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that*

$$\Pr[a \leq X \leq b] = \int_a^b f(x)dx$$

for any $a, b \in \mathbb{R}$.

Proposition 18.2.10. *The following properties hold for a probability density function f_X:*

1. $\int_{-\infty}^{\infty} f_X(x)dx = 1$

2. $F_X(x) = \int_{-\infty}^{x} f_X(y)dy$

18.2.4 Expectation and Variance

Definition 18.2.11 (Expectation). *The **expectation** or the **expected value** of a discrete random variable X is defined as*

$$\mathbb{E}[X] = \sum_i i \cdot p_X(i) = \sum_i i \cdot \Pr[X = i]$$

where p_X is its associated probability mass function. Similarly, the expectation for a continuous random variable X is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

where f_X is the associated probability density function. In either case, it is customary to denote the expected value of X as μ_X or just μ if there is no source of confusion.

Example 18.2.12. *Consider the sample space of "the outcome of tossing one die." Then the expected value of the random variable X = "the number on the first die" can be computed as*

$$\mathbb{E}[X] = 1 \cdot \frac{6}{36} + 2 \cdot \frac{6}{36} + 3 \cdot \frac{6}{36} + 4 \cdot \frac{6}{36} + 5 \cdot \frac{6}{36} + 6 \cdot \frac{6}{36} = 3.5$$

Proposition 18.2.13 (Linearity of Expectation). *If X is the sum of the random variables $X_1, \ldots, X_n$, then the following holds:*

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]$$

Also, if $a, b \in \mathbb{R}$ and X is a random variable, then

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$$

Example 18.2.14. *Consider the sample space of "the outcome of tossing two dice." Then the expected value of the random variable X = "the sum of the numbers of the two dice" can be computed as*

$$\mathbb{E}[X] = 3.5 + 3.5 = 7$$

since the expected value of the number on each die is 3.5*.*

Definition 18.2.15 (Variance). *The **variance** of a random variable X, whose expected value is μ, is defined as*

$$Var[X] = \mathbb{E}[(X - \mu)^2]$$

*Its **standard deviation** is defined as*

$$\sigma_X = \sqrt{Var[X]}$$

It is customary to denote the variance of X as σ_X^2.

Proposition 18.2.16. *If $a \in \mathbb{R}$ and X is a random variable, then*

$$Var[aX] = a^2 Var[X] \quad \sigma_{aX} = |a| \sigma_X$$

Problem 18.2.17. *Prove **Chebyshev's inequality**:*

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

for any $k > 0$. (Hint: Suppose the probability was greater than $1/k^2$. What could you conclude about $\mathbb{E}[(X-\mu)^2]$?)

18.2.5 Joint and Conditional Distribution of Random Variables

Just as in events, we are interested in multiple random variables defined on the sample space.

Definition 18.2.18 (Joint Distribution)**.** *If X, Y are discrete random variables defined on the same sample space S, the **joint probability mass function** $p_{X,Y}$ is defined as*

$$p_{X,Y}(i,j) = \Pr[X = i, Y = j]$$

where the event $X = i, Y = j$ refers to the intersection $(X = i) \cap (Y = j)$.

*If X, Y are continuous random variables defined on S, there is an associated **joint probability density function** $f_{X,Y} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that*

$$\Pr[a \leq X \leq b, c \leq Y \leq d] = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

*The joint probability mass/density function defines the **joint distribution** of the two random variables.*

Definition 18.2.19 (Marginal Distribution)**.** *Given a joint distribution $p_{X,Y}$ or $f_{X,Y}$ of two random variables X, Y, the **marginal distribution** of X can be found as*

$$p_X(i) = \sum_j p_{X,Y}(i,j)$$

if X, Y are discrete and

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

if continuous. We can equivalently define the marginal distribution of Y.

Definition 18.2.20 (Conditional Distribution)**.** *Given a joint distribution $p_{X,Y}$ or $f_{X,Y}$ of two random variables X, Y, we define the **conditional distribution of** X **given** Y as*

$$p_{X \mid Y}(i \mid j) = \frac{p_{X,Y}(i,j)}{p_Y(j)}$$

if X, Y are discrete and

$$f_{X\mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

if continuous. We can equivalently define the marginal distribution of Y given X.

18.2.6 Bayes' Rule for Random Variables

Sometimes it is easy to calculate the conditional distribution of X given Y, but not the other way around. In this case, we can apply the *Bayes' Rule* to compute the conditional distribution of Y given X. Here, we assume that X, Y are discrete random variables. By a simple application of Bayes' Rule, we have

$$\Pr[Y = j \mid X = i] = \frac{\Pr[X = i \mid Y = j]\Pr[Y = j]}{\Pr[X = i]}$$

Now by the definition of a marginal distribution, we have

$$\Pr[X = i] = \sum_{j'} \Pr[X = i, Y = j] = \sum_{j'} \Pr[X = i \mid Y = j']\Pr[Y = j']$$

for all possible values j' that Y can take. If we plug this into the denominator above,

$$\Pr[Y = j \mid X = i] = \frac{\Pr[X = i \mid Y = j]\Pr[Y = j]}{\sum\limits_{j'} \Pr[X = i \mid Y = j']\Pr[Y = j']}$$

Example 18.2.21. *There is a coin, where the probability of **Heads** is unknown and is denoted as θ. You are told that there is a 50% chance that it is a fair coin (i.e., $\theta = 0.5$) and 50% chance that it is biased to be $\theta = 0.7$. To find out if the coin is biased, you decide to flip the coin. Let D be the result of a coin flip. Then it is easy to calculate the conditional distribution of D given θ. For example,*

$$\Pr[D = H \mid \theta = 0.5] = 0.5$$

But we are more interested in the probability that the coin is fair/biased based on the observation of the coin flip. Therefore, we can apply the Bayes' Rule.

$$\Pr[\theta = 0.7 \mid D = H] = \frac{\Pr[D = H \mid \theta = 0.7]\Pr[\theta = 0.7]}{\Pr[D = H]}$$

which can be calculated as

$$\begin{aligned}&\frac{\Pr[D = H \mid \theta = 0.7]\Pr[\theta = 0.7]}{\Pr[D = H \mid \theta = 0.7]\Pr[\theta = 0.7] + \Pr[D = H \mid \theta = 0.5]\Pr[\theta = 0.5]}\\ &= \frac{0.7 \cdot 0.5}{0.7 \cdot 0.5 + 0.5 \cdot 0.5} \simeq 0.58\end{aligned}$$

*So if we observe one **Heads**, there is a 58% chance that the coin was biased and a 42% chance that it was fair.*

Problem 18.2.22. *Consider Example 18.2.21 again. This time, we decide to throw the coin 10 times in a row. Let N be the number of observed **Heads**. What is the probability that the coin is biased if $N = 7$?*

18.2.7 Independent Random Variables

Analogous to events, we can define the independence of two random variables.

Definition 18.2.23 (Independent Random Variables)**.** *Two discrete random variables X, Y are **independent** if for every i, j, we have*

$$p_X(i) = p_{X \mid Y}(i \mid j)$$

or equivalently,

$$p_Y(j) = p_{Y \mid X}(j \mid i)$$

or equivalently

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

*Two continuous random variables X, Y are **independent** if the analogous conditions hold for the probability density functions.*

Definition 18.2.24 (Mutually Independent Random Variables)**.** *If any pair of n random variables $X_1, X_2, \ldots, X_n$ are independent of each other, then the random variables are **mutually independent**.*

Proposition 18.2.25. *If $X_1, \ldots, X_n$ are mutually independent random variables, the following properties are satisfied:*

1. $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$
2. $Var[X_1 + \ldots + X_n] = Var(X_1) + \ldots + Var(X_n)$

We are particularly interested in independent random variables that have the same probability distribution. This is because if we repeat the same random process multiple times and define a random variable for each iteration, the random variables will be *independent and identically distributed*.

Definition 18.2.26. *If $X_1, \ldots, X_n$ are mutually independent random variables that have the same probability distribution, we call them **independent, identically distributed** random variables, which is more commonly denoted as **iid** or **i.i.d.** random variables.*

18.3 Central Limit Theorem and Confidence Intervals

Now we turn our attention to two very important topics in statistics: *Central Limit Theorem* and *confidence intervals*.

You may have seen *confidence intervals* or *margin of error* in the context of election polls. The pollster usually attaches a caveat to the prediction, saying that there is some probability that the true opinion of the public is $\pm\epsilon$ of the pollster's estimate, where ϵ is typically a few percent. This section is about the most basic form of confidence intervals, calculated using the famous Gaussian distribution. It also explains why the Gaussian pops up unexpectedly in so many settings.

A running example in this chapter is estimating the bias of a coin we have been given. Specifically, $\Pr[\textbf{Heads}] = p$ where p is unknown and may not be $1/2$. We wish to estimate p by repeatedly tossing the coin. If we toss the coin n times, we expect to see around np Heads. Confidence intervals ask the converse question: after having seen the number of heads in n tosses, how "confidently" can we estimate p?

18.3.1 Coin Tossing

Suppose we toss the same coin n times. For each $i = 1, 2, \ldots, n$, define the random variable X_i as an *indicator random variable* such that

$$X_i = \begin{cases} 1 & i\text{-th toss was } \textbf{Heads} \\ 0 & \textit{otherwise} \end{cases}$$

It is easily checked that $X_1, \ldots, X_n$ are iid random variables, each with $\mathbb{E}[X_i] = p$ and $Var[X_i] = p(1-p)$. Also if we have another random variable $X =$ "number of heads," notice that X is the sum of $X_1, \ldots, X_n$. Therefore, $\mathbb{E}[X] = np$ and $Var[X] = np(1-p)$.

Problem 18.3.1. *Show that if* $\Pr[\textbf{\textit{Heads}}] = p$ *then* $\mathbb{E}[X] = np$ *and* $Var[X] = np(1-p)$*. (Hint: use linearity of expectation and the fact that* X_i*'s are mutually independent.)*

Suppose $p = 0.8$. What is the distribution of X? Figure 18.1 gives the distribution of X for different n's.

Let's make some observations about Figure 18.1.

Expected value may not happen too often. For $n = 10$, the expected number of Heads is 8, but that is seen only with probability 0.3. In other words, with probability 0.7, the number of Heads is different from the expectation. [2]

The highly likely values fall in a smaller and smaller band around the expected value, as n increases.

For $n = 10$, there is a good chance that the number of Heads is

[2] In such cases, *expectation* can be a misleading term. It may in fact be *never* seen. For instance, the expected number of eyes in an individual drawn from the human population is somewhere between 1 and 2 but no individual has a non-integral number of eyes. Thus *mean value* is a more intuitive term.

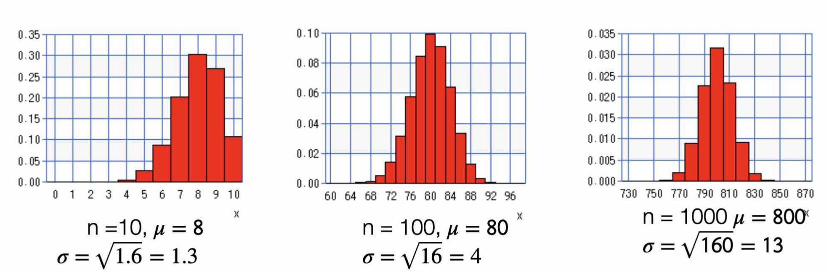


Figure 18.1: Distribution of X when we toss a coin n times, and $p = 0.8$. The plots were generated using a calculator.

quite far from the the expectation. For $n = 100$, the number of Heads lies in $[68, 90]$ with quite high probability. For $n = 1000$ it lies in $[770, 830]$ with high probability.

The probability curve becomes more symmetrical around the mean. Contrast between the case where $n = 10$ and the case where $n = 100$.

Probability curve starts resembling the famous Gaussian distribution . Also called *Normal Distribution* and in popular math, the *Bell curve*, due to its bell-like shape.

18.3.2 Gaussian Distribution

We say that a real-valued random variable X is distributed according to $\mathcal{N}(\mu, \sigma^2)$, the Gaussian distribution with mean μ and variance σ^2, if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{18.1}$$

It is hard to make an intuitive sense of this expression. The following figure gives us a better handle.

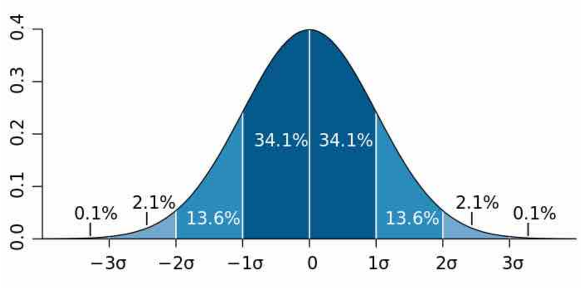


Figure 18.2: Cheatsheet for the Gaussian distribution with mean μ and variance σ^2. It is tightly concentrated in the interval $[\mu - k\sigma, \mu + k\sigma]$ for even $k = 1$ and certainly for $k = 2, 3$. *Source:* `https://en.wikipedia.org/wiki/Normal_distribution`

Figure 18.2 shows that X concentrates very strongly around the mean μ. The probability that X lies in various intervals around μ of the type $[\mu - k\sigma, \mu + k\sigma]$ are as follows: (i) For $k = 1$ it is 68.2%; (ii) For $k = 2$ it is 95.4%; (iii) For $k = 3$ it is 99.6%.

18.3.3 Central Limit Theorem (CLT)

This fundamental result explains our observations in Subsection 18.3.1.

Theorem 18.3.2 (Central Limit Theorem, informal statement)**.** *Suppose $X_1, X_2, \ldots,$ is a sequence of random variables that are mutually independent and each of whose variance is upper bounded by some constant C. Then as $n \to \infty$, the sum $X_1 + X_2 + \ldots + X_n$ tends to $\mathcal{N}(\mu, \sigma^2)$ where $\mu = \sum_i \mathbb{E}[X_i]$ and $\sigma^2 = \sum_i Var(X_i)$.*

We won't prove this theorem. We will use it primarily via the "cheatsheet" of Figure 18.2.

18.3.4 Confidence Intervals

We return to the problem of estimating the bias of a coin, namely $p = \Pr[\mathbf{Heads}]$. Suppose we toss it n times and observe X heads. Then $X = \sum_i X_i$ where X_i is the indicator random variable that signifies if i-th toss is Heads.

Since the X_i's are mutually independent, we can apply the CLT and conclude that X will approximately follow a Gaussian distribution as n grows. This is clear from Figure 18.1, where the probability histogram (which is a discrete approximation to the probability density) looks quite Gaussian-like for $n = 1000$. In this course we will assume for simplicity that CLT applies exactly. Using the mean and variance calculations from Problem 18.3.1, X is distributed like $\mathcal{N}(\mu, \sigma^2)$ where $\mu = np$, $\sigma^2 = np(1-p)$. Using the cheatsheet of Figure 18.2, we can conclude that

$$\Pr[X \notin [np - 2\sigma, np + 2\sigma]] \leq 4.6\%$$

Since $X \in [np - 2\sigma, np + 2\sigma]$ if and only if $np \in [X - 2\sigma, X + 2\sigma]$, some students have the following misconception:

> *Given the observation of X heads in n coin tosses, the probability that $np \notin [X - 2\sigma, X + 2\sigma]$ is at most* 4.6%.

But there is no *a priori distribution* on p. It is simply some (unknown) *constant* of nature that we're trying to estimate. So the correct inference should be:

> *If $np \notin [X - 2\sigma, X + 2\sigma]$, then the probability (over the n coin tosses) that we would have seen X heads is at most* 4.6%.

The above is an example of confidence bounds. Of course, you may note that σ also depends on p, so the above conclusion doesn't give us a clean confidence interval. In this course we use a simplifying assumption: to do the calculation we estimate σ^2 as $np'(1-p')$ where $p' = X/n$. (The intuitive justification is that we expect p to be close to X/n.)

Example 18.3.3. *Suppose $X = 0.8n$. Using our simplified calculation, $\sigma^2 \approx n(0.8)(0.2)$, implying $\sigma = 0.4\sqrt{n}$. Thus we conclude that if $p \notin [0.8 - 0.4/\sqrt{n}, 0.8 + 0.4/\sqrt{n}]$, then the probability of observing this many Heads in n tosses would have been less than $100 - 68.2\%$, that is, less than 31.8%.*

The concept of confidence intervals is also relevant to ML models.

Example 18.3.4. *A deep neural network model was trained to predict cancer patients' chances of staying in remission a year after chemotherapy, and we are interested in finding out its accuracy p. When the model is tested on $n = 1000$ held-out data points, this problem is equivalent to the coin flipping problem. For each of the held-out data point, the probability that the model makes the correct prediction is p. By observing the number of correct predictions on the held-out data, we can construct a confidence interval for p. Say the test accuracy was $p' = 70\%$. Then the 68% confidence interval can be written as*

$$np \in [np' - \sigma, np' + \sigma]$$

Substituting $p' = 0.7, \sigma \approx \sqrt{np'(1-p')}, n = 1000$, we get

$$1000p \in [685.5, 714.5]$$

or equivalently,

$$p \in [0.6855, 0.7145]$$

18.3.5 Confidence Intervals for Vectors

In the above settings, sampling was being used to estimate a real number, namely, $\Pr[\textbf{Heads}]$ for a coin. How about estimating a vector? For instance, in an opinion poll, respondents are being asked for opinions on multiple questions. Similarly, in *stochastic gradient descent* (Chapter 3), the gradient vector is being estimated by sampling a small number of data points. How can we develop confidence bounds for estimating a vector in $\mathbb{R}^k$ from n samples?

The confidence intervals for the coin toss setting can be easily extended to this case using the so called *Union Bound*:

$$\Pr[A_1 \cup A_2 \cup \cdots \cup A_k] \leq Pr[A_1] + \Pr[A_2] + \cdots + \Pr[A_k] \qquad (18.2)$$

This leads to the simplest confidence bound for estimating a vector in $\mathbb{R}^k$. Suppose the probability of the estimate being off by δ_i in the i-th coordinate is at most q_i. Then

$$\Pr[\text{estimate is off by } \vec{\delta}] \leq q_1 + q_2 + \cdots + q_k$$

where $\vec{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)$

18.4 Final Remarks

The CLT applies to many settings, but it doesn't apply everywhere. It is useful to clear up a couple of frequent misconceptions that students have:

1. Not every distribution involving a large number of samples is Gaussian. For example, scores on the final exam are usually not distributed like a Gaussian. Similarly, human heights are not really distributed like Gaussians.

2. Not everything that looks Gaussian-like is a result of the Central Limit Theorem. For instance, we saw that the distribution of weights in the sentiment model in Chapter 1 looked vaguely Gaussian-like, but they are not the sum of independent random variables as far as we can tell.

19
Calculus

19.1 Calculus in One Variable

In this section, we briefly review calculus in one variable.

19.1.1 Exponential and Logarithmic Functions

When we multiply the same number a by n times, we denote it as a^n. The *exponential function* is a natural extension of this concept.

Definition 19.1.1 (Exponential Function)**.** *There is a unique function $f : \mathbb{R} \to \mathbb{R}$ such that $f(n) = e^n$ for any $n \in \mathbb{N}$ and $f(x+y) = f(x)f(y)$ for any $x, y \in \mathbb{R}$. This function is called the **exponential function** and is denoted as e^x or $\exp(x)$.*

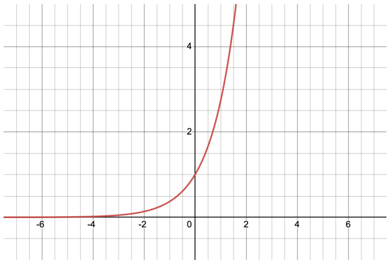


Figure 19.1: The graph of the exponential function.

Proposition 19.1.2. *The following properties hold for the exponential function:*

1. $\exp(x) > 0$ *for any* $x \in \mathbb{R}$
2. $\exp(x)$ *is increasing*
3. $\lim_{x \to -\infty} \exp(x) = 0$
4. $\lim_{x \to \infty} \exp(x) = \infty$
5. $\exp(-x) = \frac{1}{\exp(x)}$

We are also interested in the inverse function of the exponential function.

Definition 19.1.3 (Logarithmic Function)**.** *The **logarithmic function** $\log : (0, \infty) \to \mathbb{R}$ is defined as the inverse function of the exponential function. That is, $\log(x) = y$ where $x = e^y$.*

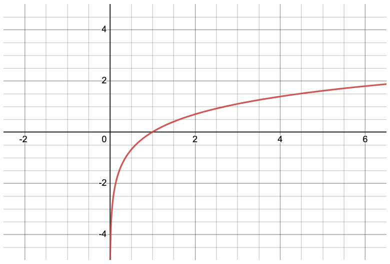


Figure 19.2: The graph of the logarithmic function.

Proposition 19.1.4. *The following properties hold for the logarithmic function:*

1. $\log(x)$ *is increasing*
2. $\lim\limits_{x \to 0+} \log(x) = -\infty$
3. $\lim\limits_{x \to \infty} \log(x) = \infty$
4. $\log(xy) = \log(x) + \log(y)$

19.1.2 Sigmoid Function

In Machine Learning, a slight variant of the exponential function, known as the *sigmoid function* is widely used.

Definition 19.1.5 (Sigmoid Function)**.** *The* ***sigmoid function*** *denoted as* $\sigma : \mathbb{R} \to \mathbb{R}$ *is defined as*

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

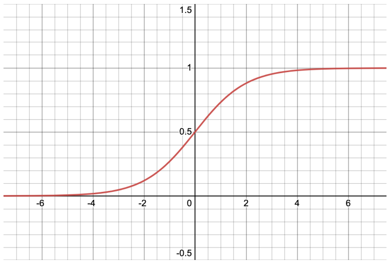


Figure 19.3: The graph of the sigmoid function.

Proposition 19.1.6. *The following properties hold for the sigmoid function:*

1. $0 < \sigma(x) < 1$ *for any* $x \in \mathbb{R}$
2. $\sigma(x)$ *is increasing*
3. $\lim\limits_{x \to -\infty} \sigma(x) = 0$
4. $\lim\limits_{x \to \infty} \sigma(x) = 1$
5. *The graph of* σ *is symmetrical to the point* $\left(0, \frac{1}{2}\right)$*. In particular,*

$$\sigma(x) + \sigma(-x) = 1$$

Because of the last property in Proposition 19.1.6, the sigmoid function is well suited for binary classification (*e.g.*, in logistic regression in Chapter 1). Given some output value x of a classification model, we interpret it as the measure of confidence that the input is of label 1, where we implicitly assume that the measure of confidence that the input is of label 2 is $-x$. Then we apply the sigmoid function to translate this into a probability distribution over the two labels.

19.1.3 Differentiation

Definition 19.1.7 (Derivative)**.** *Given a function* $f : \mathbb{R} \to \mathbb{R}$*, its* ***derivative*** f' *is defined as*

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

We alternatively denote $f'(x)$ *as* $\frac{d}{dx} f(x)$.

Example 19.1.8. *The derivative of the exponential function is itself:*

$$\exp'(x) = \exp(x)$$

and the derivative of the logarithmic function is:

$$\log'(x) = \frac{1}{x}$$

In general, there are more than two variables, that are related to each other through a composite function. The *chain rule* helps us find the derivative of the composite function.

Definition 19.1.9 (Chain Rule). *If there are functions $f, g : \mathbb{R} \to \mathbb{R}$ such that $y = f(x)$ and $z = g(y)$, then*

$$(g \circ f)'(x) = g'(f(x))f'(x) = \frac{d}{dy}g(f(x)) \cdot \frac{d}{dx}f(x)$$

or equivalently

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

19.2 Multivariable Calculus

In this section, we introduce the basics of multivariable calculus, which is widely used in Machine Learning. Since this is a generalization of the calculus in one variable, it will be useful to pay close attention to the similarity with the results from the previous section.

19.2.1 Mappings of Several Variables

So far, we only considered functions of the form $f : \mathbb{R} \to \mathbb{R}$ that map a real value x to a real value y. But now we are interested in mappings $f : \mathbb{R}^n \to \mathbb{R}^m$ that map a vector $\vec{\mathbf{x}} = (x_1, \ldots, x_n)$ with n coordinates to a vector $\vec{\mathbf{y}} = (y_1, \ldots, y_m)$ with m coordinates. In general, a *function* is a special case of a *mapping* where the range is $\mathbb{R}$. If the mappings are of the form $f : \mathbb{R}^n \to \mathbb{R}$ (*i.e.*, $m = 1$), it can still be called a *function* of several variables.

First consider an example where $m = 1$.

Example 19.2.1. *Let $f(x_1, x_2) = x_1^2 + x_2^2$ be a function in two variables. This can be understood as mapping a point $\vec{\mathbf{x}} = (x_1, x_2)$ in the Cartesian coordinate system to its squared distance from the origin. For example, $f(3,4) = 25$ shows that the point the squared distance between $(3,4)$ and the origin $(0,0)$ is 25.*

When $m > 1$, we notice that each coordinate $y_1, \ldots, y_m$ is a function of $x_1, \ldots, x_n$. Therefore, we can decompose f into m functions $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(\vec{\mathbf{x}}) = (f_1(\vec{\mathbf{x}}), \ldots, f_m(\vec{\mathbf{x}}))$$

Example 19.2.2. *Let $f(x_1, x_2) = (x_1^2 x_2, x_1 x_2^2)$ be a mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$. Then we can decompose f into two functions f_1, f_2 in two variables where*

$$f_1(x_1, x_2) = x_1^2 x_2$$
$$f_2(x_1, x_2) = x_1 x_2^2$$

19.2.2 Softmax Function

The *softmax function* is a multivariable function widely used in Machine Learning, especially for multi-class classification (see Chapter 4, Chapter 10). It takes in a vector of k values, each corresponding to a particular class, and outputs a probability distribution over the k classes — that is, a vector of k non-negative values that sum up to 1. The resulting probability is *exponentially proportional* to the input value of that class. We formally write this as:

Definition 19.2.3 (Softmax Function)**.** *Given a vector $\vec{\mathbf{z}} = (z_1, z_2, \ldots, z_k) \in \mathbb{R}^k$, we define the* ***softmax function*** *as a probability function $softmax : \mathbb{R}^k \to [0,1]^k$ where the "probability of predicting class i" is:*

$$softmax(\vec{\mathbf{z}})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{19.1}$$

Problem 19.2.4. *Show that for $k = 2$, the definition of the softmax function is equivalent to the sigmoid function (after slight rearrangement/renaming of terms).*

The sigmoid function is used for binary classification, where it takes in a single real value and converts it to a probability of one class (and the probability of the other class can be inferred as its complement). The softmax function is used for multi-class classification, where it takes in k real values and converts them to k probabilities, one for each class.

19.2.3 Differentiation

Just like with functions in one variable, we can define differentiation for mappings in several variables. The key point is that now we will define a *partial derivative* for each pair (x_i, y_j) of coordinate x_i of the domain and coordinate y_j of the range.

Definition 19.2.5 (Partial Derivative)**.** *Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the* ***partial derivative of*** y_j ***with respect to*** x_i ***at the point*** $\vec{\mathbf{x}}$ *is defined as*

$$\left.\frac{\partial y_j}{\partial x_i}\right|_{\vec{\mathbf{x}}} = \lim_{h \to 0} \frac{f_j(x_1, \ldots, x_j + h, \ldots, x_n) - f_j(x_1, \ldots, x_j, \ldots, x_n)}{h}$$

Definition 19.2.6 (Gradient). *If $f : \mathbb{R}^n \to \mathbb{R}$ is a function of several variables, the* gradient of f *is defined as a mapping $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ that maps each input vector to the vector of partial derivatives at that point:*

$$\nabla f(\vec{\mathbf{x}}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)\Bigg|_{\vec{\mathbf{x}}}$$

Similarly to the chain rule in one variable, we can define a chain rule for multivariable settings. The key point is that there are multiple ways that a coordinate x_j can affect the value of z_i. Definition 19.2.7 can be thought as applying the chain rule for one variable in each of the paths, and adding up the results.

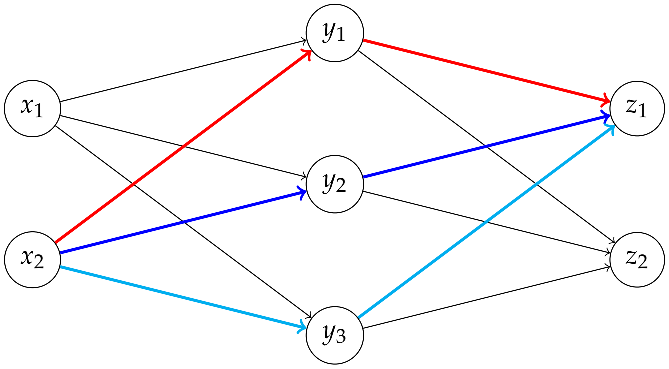


Figure 19.4: A visualization of the chain rule in multivariable settings. Notice that x_2 can affect the value of z_1 in three different paths. The amount of effect from each path will respectively be calculated as $(\partial z_1/\partial y_1)(\partial y_1/\partial x_2)$ (red), $(\partial z_1/\partial y_2)(\partial y_2/\partial x_2)$ (blue), and $(\partial z_1/\partial y_3)(\partial y_3/\partial x_2)$ (cyan).

Definition 19.2.7 (Chain Rule). *If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^\ell$ are mappings of several variables, where $\vec{\mathbf{y}} = f(\vec{\mathbf{x}})$ and $\vec{\mathbf{z}} = g(\vec{\mathbf{y}})$, the following **chain rule** holds for each $1 \le i \le \ell$ and $1 \le j \le n$:*

$$\frac{\partial z_i}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial z_i}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_j}$$

Example 19.2.8. *Suppose we define the functions $h = s + t^2$, $s = 3x$, and $t = x - 2$. Then, we can find the partial derivative $\frac{\partial h}{\partial x}$ using the chain rule:*

$$\begin{aligned}\frac{\partial h}{\partial x} &= \frac{\partial s}{\partial x} + \frac{\partial (t^2)}{\partial x}\\ &= \frac{\partial s}{\partial x} + \frac{\partial (t^2)}{\partial t} \cdot \frac{\partial t}{\partial x}\\ &= 3 + 2t \cdot 1\\ &= 2x - 1\end{aligned}$$

Problem 19.2.9. *Suppose we define the functions $h = s + t^2$, $s = xy$, and $t = x - 2y$. Compute the partial derivative $\partial h/\partial x$.*

20

Linear Algebra

20.1 Vectors

Vectors are a collection of entries (here, we focus only on real numbers). For example, the pair $(1, 2)$ is a real vector of size 2, and the 3-tuple $(1, 0, 2)$ is a real vector of size 3. We primarily categorize vectors by their size. For example, the set of all real vectors of size n is denoted as $\mathbb{R}^n$. Any element of $\mathbb{R}^n$ can be thought of as representing a point (or equivalently, the direction from the origin to the point) in the n-dimensional Cartesian space. A real number in $\mathbb{R}$ is also known as a *scalar*, as opposed to *vectors* in $\mathbb{R}^n$ where $n > 1$.

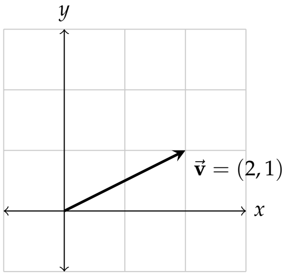
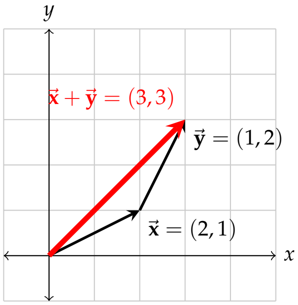


Figure 20.1: A visualization of a vector $\vec{\mathbf{v}} = (2, 1)$ in $\mathbb{R}^2$.

20.1.1 Vector Space

We are interested in two operations defined on vectors — vector addition and scalar multiplication. Given vectors $\vec{\mathbf{x}} = (x_1, x_2, \ldots, x_n)$ and $\vec{\mathbf{y}} = (y_1, y_2, \ldots, y_n)$ and a scalar $c \in \mathbb{R}$, the *vector addition* is defined as

$$\vec{\mathbf{x}} + \vec{\mathbf{y}} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) \in \mathbb{R}^n$$



y
$\vec{\mathbf{x}} + \vec{\mathbf{y}} = (3, 3)$
$\vec{\mathbf{y}} = (1, 2)$
$\vec{\mathbf{x}} = (2, 1)$
x

Figure 20.2: A visualization of $\vec{\mathbf{x}} + \vec{\mathbf{y}}$ where $\vec{\mathbf{x}} = (2, 1)$ and $\vec{\mathbf{y}} = (1, 2)$.

where we add each of the coordinates element-wise. As shown in Figure 20.2, vector addition is the process of finding the diagonal of the parallelogram made by the two vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$. The *scalar multiplication* is similarly defined as

$$c\vec{\mathbf{x}} = (cx_1, cx_2, \ldots, cx_n) \in \mathbb{R}^n$$

As shown in Figure 20.3, scalar multiplication is the process of scaling one vector up or down.

$\mathbb{R}^n$ is closed under these two operations — *i.e.*, the resulting vector of either operation is still in $\mathbb{R}^n$. Any subset S of $\mathbb{R}^n$ that is closed under vector addition and scalar multiplication is known as a *subspace* of R^n.

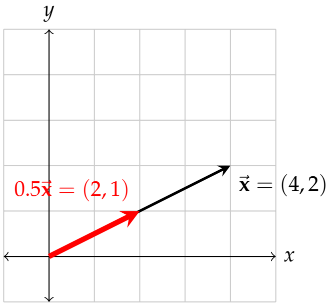


Figure 20.3: A visualization of $0.5\vec{\mathbf{x}}$ where $\vec{\mathbf{x}} = (4, 2)$.

20.1.2 Inner Product

The *inner product* is defined as

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}$$

Closely related to the inner product is the *norm* of a vector, which measures the *length* of it. It is defined as $\|\vec{\mathbf{x}}\| = \sqrt{\vec{\mathbf{x}} \cdot \vec{\mathbf{x}}}$. [1]

Proposition 20.1.1. *The inner product satisfies the following properties:*

- *Symmetry:* $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = \vec{\mathbf{y}} \cdot \vec{\mathbf{x}}$
- *Linearity:* $(a_1\vec{\mathbf{x}}_1 + a_2\vec{\mathbf{x}}_2) \cdot \vec{\mathbf{y}} = a_1(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{y}}) + a_2(\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{y}})$

and the norm satisfies the following property:

- *Absolute Homogeneity:* $\|a\vec{\mathbf{x}}\| = |a| \, \|\vec{\mathbf{x}}\|$

20.1.3 Linear Independence

Any vector of the form

$$a_1\vec{\mathbf{x}}_1 + a_2\vec{\mathbf{x}}_2 + \ldots + a_k\vec{\mathbf{x}}_k$$

where a_i's are scalars and $\vec{\mathbf{x}}_i$'s are vectors is called a *linear combination* of the vectors $\vec{\mathbf{x}}_i$'s. Notice that the zero vector $\vec{\mathbf{0}}$ (*i.e.*, the vector with all zero entries) can always be represented as a linear combination of an arbitrary collection of vectors, if all a_i's are chosen as zero. This is known as a *trivial linear combination*, and any other choice of a_i's is known as a *non-trivial linear combination*.

Definition 20.1.2. *k vectors $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_k \in \mathbb{R}^n$ are called* ***linearly dependent*** *if $\vec{\mathbf{0}}$ can be represented as a non-trivial linear combination of the vectors $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k$; or equivalently, if one of the vectors can be represented as a linear combination of the remaining $k-1$ vectors. The vectors that are not linearly dependent with each other are called* ***linearly independent****.*

Consider the following analogy. Imagine trying to have a family style dinner at a fast food restaurant, where the first person orders a burger, the second person orders a chilli cheese fries, and the third person orders a set menu with a burger and a chili cheese fries. The third person's order did not contribute to the diversity of the food on the dinner table. Similarly, if some set of vectors are linearly dependent, it means that at least one of the vectors is redundant.

Example 20.1.3. *The set $\{(-1, 2), (3, 0), (1, 4)\}$ of three vectors is linearly dependent because*

$$(1, 4) = 2 \cdot (-1, 2) + (3, 0)$$

can be represented as the linear combination of the remaining two vectors.

[1] There are many other definitions of a norm. This particular one is called an ℓ_2 norm.

Example 20.1.4. *The set* $\{(-1,2,1),(3,0,0),(1,4,1)\}$ *of three vectors is linearly independent because there is no way to write one vector as a linear combination of the remaining two vectors.*

20.1.4 Span

Definition 20.1.5. *The* ***span*** *of a set of vectors* $\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k$ *is the set of all vectors that can be represented as a linear combination of* $\vec{\mathbf{x}}_i$*'s.*

Example 20.1.6. $(1,4)$ *is in the span of* $\{(-1,2),(3,0)\}$ *because*

$$(1,4) = 2\cdot(-1,2)+(3,0)$$

Example 20.1.7. $(1,4,1)$ *is not in the span of* $\{(-1,2,1),(3,0,0)\}$ *because there is no way to choose* $a_1,a_2\in\mathbb{R}$ *such that*

$$(1,4,1) = a_1(-1,2,1)+a_2(3,0,0)$$

The span is also known as the *subspace generated by the vectors* $\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k$. This is because if you add any two vectors in the span, or multiply one by a scalar, it is still in the span (*i.e.*, the span is closed under vector addition and scalar multiplication).

Example 20.1.8. *In the* $\mathbb{R}^3$*, the two vectors* $(1,0,0)$ *and* $(0,1,0)$ *span the* 2*-dimensional XY-plane. Similarly, the vectors* $(1,0,1)$ *and* $(0,2,1)$ *span the* 2*-dimensional plane* $2x+y-2z=0$. [2]

In Example 20.1.8, we see examples where 2 vectors span a 2-dimensional subspace. In general, the dimension of the subspace spanned by k vectors can go up to k, but it can also be strictly smaller than k. This is related to the *linear independence* of the vectors.

Proposition 20.1.9. *Given k vectors,* $\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k\in\mathbb{R}^n$*, there is a maximum number* $d\geq 1$ *such that there is some subcollection* $\vec{\mathbf{x}}_{i_1},\ldots,\vec{\mathbf{x}}_{i_d}$ *of these vectors that are linearly independent. Then*

$$span(\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_k) = span(\vec{\mathbf{x}}_{i_1},\ldots,\vec{\mathbf{x}}_{i_d}) \tag{20.1}$$

is a d-dimensional subspace of $\mathbb{R}^n$.

Conversely, if we know that the span of the k vectors is a d-dimensional subspace, then the maximum number of vectors that are linearly independent with each other is d, and any subcollection of linearly independent d vectors satisfies (20.1).

Proposition 20.1.9 states that the span of some set of k vectors is equivalent to the maximum number d of linearly independent vectors. It also states that the span of the k vectors is equal to the span of the linearly independent d vectors, meaning all of the information

[2] The term *dimension* will be formally defined soon. Here, we rely on your intuition.

is captured by the d vectors; the remaining $k-d$ vectors are just redundancies. But trying to directly compute the maximum number of linearly independent vectors is inefficient — it may require checking the linear independence of an exponential number of subsets of the vectors. In the next section, we discuss a concept called *matrix rank* that is very closely related to this topic.

20.1.5 Orthogonal Vectors

Definition 20.1.10. *If vectors $\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k \in \mathbb{R}^n$ satisfy $\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j = 0$ for any $i \neq j$, then they are called* ***orthogonal*** *vectors. In particular, if they also satisfy the condition that $\|\vec{\mathbf{x}}_i\| = 1$ for each i, then they are also* ***orthonormal****.*

In $\mathbb{R}^n$, orthogonal vectors form a 90 degrees angle with each other.

Example 20.1.11. *The two vectors $(1,0),(0,2)$ are orthogonal. So are the vectors $(1,2),(-2,1)$.*

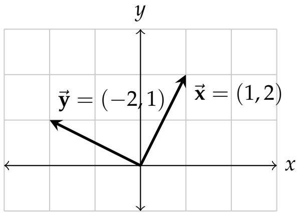


Figure 20.4: A visualization of orthogonal vectors $\vec{\mathbf{x}} = (1,2)$ and $\vec{\mathbf{y}} = (-2,1)$.

Given any set of orthogonal vectors, it is possible to transform it into a set of orthonormal vectors, by normalizing each vector (*i.e.*, scale it such that the norm is 1).

20.1.6 Basis

Definition 20.1.12. *A collection $\{\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_k\}$ of linearly independent vectors in $\mathbb{R}^n$ that span a set S is known as a* ***basis*** *of S. In particular, if the vectors of the basis are orthogonal/orthonormal, the basis is called an* ***orthogonal/orthonormal basis*** *of S.*

The set S in Definition 20.1.12 can be the entire vector space $\mathbb{R}^n$, but it can also be some subspace of $\mathbb{R}^n$ with a lower dimension.

Example 20.1.13. *The set $\{(1,0,0),(0,1,0),(0,0,1)\}$ of three vectors is a basis for $\mathbb{R}^3$. When we exclude the last vector $(0,0,1)$, the set $\{(1,0,0),(0,1,0)\}$ is a basis of the 2-dimensional XY-plane in $\mathbb{R}^3$.*

Given some subspace S, the basis of S is not unique. However, every basis of S must have the same size — this size is called the *dimension* of S. For a finite dimensional space S, it is known that there exists an *orthogonal* basis of S. There is a well-known algorithm — Gram-Schmidt process — that can transform an arbitrary basis into an orthogonal basis (and eventually an orthonormal basis via normalization).

20.1.7 Projection

Vector projection is the key concept used in the Gram-Schmidt process that computes an orthogonal basis. Given a fixed vector $\vec{\mathbf{a}}$, it decomposes any given vector $\vec{\mathbf{x}}$ into a sum of two components — one that is orthogonal to $\vec{\mathbf{a}}$ ("distinct information") and the other that is parallel to $\vec{\mathbf{a}}$ ("redundant information").

Definition 20.1.14 (Vector Projection)**.** *Fix a vector $\vec{\mathbf{a}} \in \mathbb{R}^n$. Given another vector $\vec{\mathbf{x}}$, the **projection of $\vec{\mathbf{x}}$ on $\vec{\mathbf{a}}$** is defined as*

$$proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}}) = \frac{\vec{\mathbf{x}} \cdot \vec{\mathbf{a}}}{\vec{\mathbf{a}} \cdot \vec{\mathbf{a}}}\vec{\mathbf{a}}$$

and is parallel to the fixed vector $\vec{\mathbf{a}}$. The remaining component

$$\vec{\mathbf{x}} - proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}})$$

*is called the **rejection of $\vec{\mathbf{x}}$ from $\vec{\mathbf{a}}$** and is orthogonal to $\vec{\mathbf{a}}$.*

Proposition 20.1.15 (Pythagorean Theorem)**.** *If $\vec{\mathbf{x}}, \vec{\mathbf{y}}$ are orthogonal, then*

$$\|\vec{\mathbf{x}} + \vec{\mathbf{y}}\|^2 = \|\vec{\mathbf{x}}\|^2 + \|\vec{\mathbf{y}}\|^2$$

In particular, given two vectors $\vec{\mathbf{a}}, \vec{\mathbf{x}}$, we have

$$\|\vec{\mathbf{x}} - proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}})\|^2 = \|\vec{\mathbf{x}}\|^2 - \|proj_{\vec{\mathbf{a}}}(\vec{\mathbf{x}})\|^2$$

Now assume we are given a space S and a subspace $T \subset S$. Then a vector $\vec{\mathbf{x}} \in S$ in the larger space does not necessarily belong in T. Instead, we can find a vector $\vec{\mathbf{x}}' \in T$ that is "closest" to $\vec{\mathbf{x}}$ using vector projection. [3]

Definition 20.1.16 (Vector Projection on Subspace)**.** *Given a space S, its subspace T with an orthogonal basis $\{\vec{\mathbf{t}}_1, \ldots, \vec{\mathbf{t}}_k\}$, and a vector $\vec{\mathbf{x}} \in S$, the **projection of $\vec{\mathbf{x}}$ on** T is defined as*

$$proj_T(\vec{\mathbf{x}}) = \sum_{i=1}^{k} proj_{\vec{\mathbf{t}}_i}(\vec{\mathbf{x}}) = \sum_{i=1}^{k} \frac{\vec{\mathbf{x}} \cdot \vec{\mathbf{t}}_i}{\vec{\mathbf{t}}_i \cdot \vec{\mathbf{t}}_i}\vec{\mathbf{t}}_i$$

the sum of projection of $\vec{\mathbf{x}}$ on each of the basis vectors of T.

20.2 Matrices

Matrices are a generalization of *vectors* in 2-dimension — a $m \times n$ matrix is a collection of numbers assembled in a rectangular shape of m rows and n columns. The set of all real matrices of size $m \times n$ is denoted as $\mathbb{R}^{m \times n}$. A vector of size n is customarily understood as a column vector — that is, a $n \times 1$ matrix. Also, if $m = n$, then the matrix is known as a *square matrix*.

[3] We ask you to prove this in Problem 7.1.3.

20.2.1 Matrix Operation

Similarly to vector operations, we are interested in four matrix operations — matrix addition, scalar multiplication, matrix multiplication, and transpose. Given a scalar $c \in \mathbb{R}$ and matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix} \quad \textit{and} \quad \mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,n} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m,1} & y_{m,2} & \cdots & y_{m,n} \end{bmatrix}$$

the matrix addition is defined as

$$\mathbf{X} + \mathbf{Y} = \begin{bmatrix} x_{1,1} + y_{1,1} & x_{1,2} + y_{1,2} & \cdots & x_{1,n} + y_{1,n} \\ x_{2,1} + y_{2,1} & x_{2,2} + y_{2,2} & \cdots & x_{2,n} + y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} + y_{m,1} & x_{m,2} + y_{m,2} & \cdots & x_{m,n} + y_{m,n} \end{bmatrix}$$

where we add each of the coordinates element-wise. The *scalar multiplication* is similarly defined as

$$c\mathbf{X} = \begin{bmatrix} cx_{1,1} & cx_{1,2} & \cdots & cx_{1,n} \\ cx_{2,1} & cx_{2,2} & \cdots & cx_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ cx_{m,1} & cx_{m,2} & \cdots & cx_{m,n} \end{bmatrix}$$

The *matrix multiplication* $\mathbf{XY}$ is defined for a matrix $\mathbf{X} \in \mathbb{R}^{\ell \times m}$ and a matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$; that is, when the number of columns of the first matrix is equal to the number of rows of the second matrix. The output $\mathbf{XY}$ of the matrix multiplication will be a $\ell \times n$ matrix. The (i, j) entry of the matrix $\mathbf{XY}$ is defined as

$$(\mathbf{XY})_{i,j} = \sum_{k=1}^{m} x_{i,k} y_{k,j}$$

That is, it is defined as the inner product of the i-th row of $\mathbf{X}$ and the j-th column of $\mathbf{Y}$.

Proposition 20.2.1. *The above matrix operations satisfy the following properties:*

- $c(\mathbf{XY}) = (c\mathbf{X})\mathbf{Y} = \mathbf{X}(c\mathbf{Y})$
- $(\mathbf{X}_1 + \mathbf{X}_2)\mathbf{Y} = \mathbf{X}_1\mathbf{Y} + \mathbf{X}_2\mathbf{Y}$
- $\mathbf{X}(\mathbf{Y}_1 + \mathbf{Y}_2) = \mathbf{XY}_1 + \mathbf{XY}_2$

Finally, the *transpose* $\mathbf{X}^\intercal \in \mathbb{R}^{n \times m}$ of a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is the resulting matrix when the entries of $\mathbf{X}$ are reflected down the diagonal. That is,

$$(\mathbf{X}^\intercal)_{i,j} = \mathbf{X}_{j,i}$$

Proposition 20.2.2. *The transpose of a matrix satisfies the following properties:*

- $(\mathbf{X} + \mathbf{Y})^\intercal = \mathbf{X}^\intercal + \mathbf{Y}^\intercal$
- $(c\mathbf{X})^\intercal = c(\mathbf{X}^\intercal)$
- $(\mathbf{X}\mathbf{Y})^\intercal = \mathbf{Y}^\intercal \mathbf{X}^\intercal$

20.2.2 Matrix and Linear Transformation

Recall that a vector of size n is often considered a $n \times 1$ matrix. Therefore, given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\vec{\mathbf{x}} \in \mathbb{R}^n$, we can define the following operation

$$\vec{\mathbf{y}} = \mathbf{A}\vec{\mathbf{x}} \in \mathbb{R}^m$$

through matrix multiplication. This shows that $\mathbf{A}$ can be understood as a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. We see that $a_{i,j}$ (the (i, j) entry of the matrix $\mathbf{A}$) is the coefficient of x_j (the j-th coordinate of the input vector) when computing y_i (the i-th coordinate of the output vector). Since each y_i is linear in terms of each x_j, we say that $\mathbf{A}$ is a *linear transformation*.

20.2.3 Matrix Rank

Matrix rank is one of the most important concepts in basic linear algebra.

Definition 20.2.3. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of m rows and n columns, the number of linearly independent rows is known to be always equal to the number of linearly independent columns. This common number is known as the* **rank** *of $\mathbf{A}$ and is denoted as rank$(\mathbf{A})$.*

The following property of rank is implied in the definition, but we state it explicitly as follows.

Proposition 20.2.4. *The rank of a matrix is invariant to reordering rows/-columns.*

Example 20.2.5. *Consider the matrix $M = \begin{bmatrix} 1 & 1 & -2 & 0 \\ -1 & -1 & 2 & 0 \end{bmatrix}$, we notice that the second row is simply the first row negated, and thus the* rank *of M is* 1.

Example 20.2.6. *Consider the matrix* $M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$*, the rank of* M *is* 3 *because all the row (or column) vectors are linearly independent (they form basis vectors of* $\mathbb{R}^3$*).*

Example 20.2.7. *Consider the matrix* $M = \begin{bmatrix} 1 & 0 & 1 \\ -2 & -3 & 1 \\ 3 & 3 & 0 \end{bmatrix}$*, the rank of* M *is* 2 *because the third row can be expressed as the first row subtracted by the second row.*

When we interpret a matrix as a linear transformation, the rank measures the dimension of the output space.

Proposition 20.2.8. $\mathbf{A} \in \mathbb{R}^{m \times n}$ *has rank* k *if and only if the image of the linear transformation; i.e., the subspace*

$$\{\mathbf{A}\vec{\mathbf{x}} \mid \vec{\mathbf{x}} \in \mathbb{R}^n\}$$

of $\mathbb{R}^m$ *has dimension* k*.*

There are many known algorithms to compute the rank of a matrix. Examples include Gaussian elimination or certain decompositions (expressing a matrix as the product of other matrices with certain properties). Given m vectors in $\mathbb{R}^n$, we can find the maximum number of linearly independent vectors by constructing a matrix with each row equal to each vector [4] and finding the rank of that matrix.

20.2.4 Eigenvalues and Eigenvectors

Say we have a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. This means that the linear transformation expressed by $\mathbf{A}$ is a mapping from $\mathbb{R}^n$ to itself. Most vectors $\vec{\mathbf{x}} \in \mathbb{R}^n$ is mapped to a very "different" vector $\mathbf{A}\vec{\mathbf{x}}$ under this mapping. However, some vectors are "special" and they are mapped to another vector with the same direction.

Definition 20.2.9 (Eigenvalue/Eigenvector)**.** *Given a square matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$*, if a vector* $\vec{\mathbf{v}} \in \mathbb{R}^n$ *satisfies*

$$\mathbf{A}\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$$

for some scalar $\lambda \in \mathbb{R}$*, then* $\vec{\mathbf{v}}$ *is known as an* ***eigenvector*** *of* $\mathbf{A}$*, and* λ *is its corresponding* ***eigenvalue****.*

Each eigenvector can only be associated with one eigenvalue, but each eigenvalue may be associated with multiple eigenvectors.

Proposition 20.2.10. *If* $\vec{\mathbf{x}}, \vec{\mathbf{y}}$ *are both eigenvectors of* $\mathbf{A}$ *for the same eigenvalue* λ*, then any linear combination of them is also an eigenvector for* $\mathbf{A}$ *with the same eigenvalue* λ*.*

[4] By Proposition 20.2.4, the order of the rows can be arbitrary.

Proposition 20.2.10 shows that the set of eigenvectors for a particular eigenvalue forms a subspace, known as the *eigenspace* of that eigenvalue. The dimension of this subspace is known as the *geometric multiplicity* of the eigenvalue. The following result ties together some of the concepts we discussed so far.

Proposition 20.2.11 (Rank-Nullity Theorem). *Given a square matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$*, the eigenspace of* 0 *is the set of all vectors that get mapped to zero vector* $\vec{\mathbf{0}}$ *under the linear transformation* $\mathbf{A}$*. This subspace is known as the* ***null space*** *of* $\mathbf{A}$ *and its dimension (i.e., the geometric multiplicity of* 0*) is known as the* ***nullity*** *of* $\mathbf{A}$ *and is denoted as* $nullity(\mathbf{A})$*. Then*

$$rank(\mathbf{A}) + nullity(\mathbf{A}) = n$$

20.3 Advanced: SVD/PCA Procedures

Now we briefly introduce a procedure called *Principal Component Analysis (PCA)*, which is commonly used in low-dimensional representation as in Chapter 7.

We are given vectors $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots, \vec{\mathbf{v}}_N \in \mathbb{R}^d$ and a positive integer k and wish to obtain the low-dimensional representation in the sense of Definition 7.1.1 that minimizes ϵ. This is what we mean by "best" representation.

Theorem 20.3.1. *The best low-dimensional representation consists of* k *eigenvectors corresponding to the top* k *eigenvalues (largest numerical values) of the matrix* $\mathbf{A}\mathbf{A}^\intercal$ *where the columns of* $\mathbf{A}$ *are* $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \ldots, \vec{\mathbf{v}}_N$*.*

Theorem 20.3.1 shows what the best low-dimensional representation is, but it does not show *how* to compute it. It turns out something called the *Singular Value Decomposition (SVD)* of the matrix $\mathbf{A}$ is useful. It is known that any matrix $\mathbf{A}$ can be decomposed into the following product

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$$

where $\mathbf{\Sigma}$ is a diagonal matrix with entries equal to the square root of the nonzero eigenvalues of $\mathbf{A}\mathbf{A}^\intercal$ and the columns of $\mathbf{U}$ are the orthonormal eigenvectors of $\mathbf{A}\mathbf{A}^\intercal$, where the i-th column is the eigenvector that corresponds to the eigenvalue at the i-th diagonal entry of $\mathbf{\Sigma}$. There are known computationally efficient algorithms that will perform the SVD of a matrix.

In this section, we will prove Theorem 20.3.1 for the case where $k = 1$. To do this, we need to introduce some preliminary results.

Theorem 20.3.2. *If a square matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *is symmetric (i.e.,* $\mathbf{A} = \mathbf{A}^\intercal$*), then there is an orthonormal basis of* $\mathbb{R}^n$ *consisting of* n *eigenvectors of* $\mathbf{A}$*.* [5]

[5] This is known as the Spectral Theorem.

Proof. A real symmetric matrix is known to be *diagonalizable*, and diagonalizable matrices are known to have n eigenvectors that form a basis for $\mathbb{R}^n$. In particular, the eigenvectors are linearly independent, meaning the eigenvectors corresponding to a particular eigenvalue λ will form a basis for the corresponding eigenspace. Through the Gram-Schmidt process, we can replace some of these eigenvectors such that the eigenvectors for λ are orthogonal to each other. That is, if $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ are eigenvectors for the same eigenvalue λ, then $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$. Now assume $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ are two eigenvectors with distinct eigenvalues λ, μ respectively. Then

$$\begin{aligned}\lambda\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = (\lambda\vec{\mathbf{u}}) \cdot \vec{\mathbf{v}} = (\mathbf{A}\vec{\mathbf{u}}) \cdot \vec{\mathbf{v}} &= \sum_{i,j=1}^{n} a_{i,j}u_j v_i \\ &= \vec{\mathbf{u}} \cdot (\mathbf{A}^\intercal\vec{\mathbf{v}}) = \vec{\mathbf{u}} \cdot (\mathbf{A}\vec{\mathbf{v}}) = \vec{\mathbf{u}} \cdot (\mu\vec{\mathbf{v}}) = \mu\vec{\mathbf{u}} \cdot \vec{\mathbf{v}}\end{aligned}$$

where the third and the fourth equality can be verified by direct computation. Since $\lambda \neq \mu$, we conclude $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$. We have now showed that $\vec{\mathbf{u}} \cdot \vec{\mathbf{v}} = 0$ for any pair of eigenvectors $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ — this means that the basis of eigenvectors is also orthogonal. After normalization, the basis can be made orthonormal. □

The following result is not necessarily needed for the proof of Theorem 20.3.1, but the proofs are similar.

Theorem 20.3.3. *If* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *is symmetric, then the unit vector* $\vec{\mathbf{x}}$ *that maximizes* $\|\mathbf{A}\vec{\mathbf{x}}\|$ *is an eigenvector of* $\mathbf{A}$ *with an eigenvalue, whose absolute values is the largest out of all eigenvalues.*

Proof. By Theorem 20.3.2, there is an orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$. Then any vector $\vec{\mathbf{x}}$ is in the span of the eigenvectors and can be represented as the linear combination

$$\vec{\mathbf{x}} = \alpha_1\vec{\mathbf{u}}_1 + \alpha_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\vec{\mathbf{u}}_n$$

for some scalars α_i's. Then

$$\begin{aligned}\|\vec{\mathbf{x}}\|^2 &= \vec{\mathbf{x}} \cdot \vec{\mathbf{x}} \\ &= (\alpha_1\vec{\mathbf{u}}_1 + \alpha_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\vec{\mathbf{u}}_n) \cdot (\alpha_1\vec{\mathbf{u}}_1 + \alpha_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\vec{\mathbf{u}}_n) \\ &= \sum_{i,j=1}^{n} \alpha_i\alpha_j(\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j) \\ &= \sum_{i=1}^{n} \alpha_i^2\end{aligned}$$

where for the last equality, we use the fact that $\vec{\mathbf{u}}_i$'s are orthonormal — that is, $\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j = 0$ if $i \neq j$ and $\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_i = 1$. Since $\vec{\mathbf{x}}$ has norm 1, we see that $\sum\limits_{i=1}^{n} \alpha_i^2 = 1$. Now notice that

$$\begin{aligned}\mathbf{A}\vec{\mathbf{x}} &= \mathbf{A}(\alpha_1\vec{\mathbf{u}}_1 + \alpha_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\vec{\mathbf{u}}_n) \\ &= \alpha_1\mathbf{A}\vec{\mathbf{u}}_1 + \alpha_2\mathbf{A}\vec{\mathbf{u}}_2 + \ldots + \alpha_n\mathbf{A}\vec{\mathbf{u}}_n \\ &= \alpha_1\lambda_1\vec{\mathbf{u}}_1 + \alpha_2\lambda_2\vec{\mathbf{u}}_2 + \ldots + \alpha_n\lambda_n\vec{\mathbf{u}}_n\end{aligned}$$

where λ_i is the eigenvalue for the eigenvector $\vec{\mathbf{u}}_i$. Following a similar computation as above,

$$\|\mathbf{A}\vec{\mathbf{x}}\|^2 = \sum_{i=1}^{n} \alpha_i^2\lambda_i^2$$

The allocation of weights α_i that will maximize $\sum\limits_{i=1}^{n} \alpha_i^2\lambda_i^2$ while maintaining $\sum\limits_{i=1}^{n} \alpha_i^2 = 1$ is assigning $\alpha_i = \pm 1$ to the eigenvalue λ_i that has the highest value of λ_i^2. This shows that the unit vector $\vec{\mathbf{x}} = \pm\vec{\mathbf{u}}_i$ is an eigenvector with the eigenvalue λ_i. □

We now prove one last preliminary result.

Theorem 20.3.4. *For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the matrix $\mathbf{A}\mathbf{A}^\intercal$ is symmetric and its eigenvalues are non-negative.*

Proof. The first part can be verified easily by observing that

$$(\mathbf{A}\mathbf{A}^\intercal)^\intercal = (\mathbf{A}^\intercal)^\intercal\mathbf{A}^\intercal = \mathbf{A}\mathbf{A}^\intercal$$

Now assume $\vec{\mathbf{x}}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ. Then

$$\mathbf{A}\mathbf{A}^\intercal\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$$

We multiply $\vec{\mathbf{x}}^\intercal$ on the left on both sides of the equation.

$$\vec{\mathbf{x}}^\intercal\mathbf{A}\mathbf{A}^\intercal\vec{\mathbf{x}} = \vec{\mathbf{x}}^\intercal(\lambda\vec{\mathbf{x}}) = \lambda\,\|\vec{\mathbf{x}}\|^2$$

At the same time, notice that

$$\vec{\mathbf{x}}^\intercal\mathbf{A}\mathbf{A}^\intercal\vec{\mathbf{x}} = (\mathbf{A}^\intercal\vec{\mathbf{x}})^\intercal(\mathbf{A}^\intercal\vec{\mathbf{x}}) = \|\mathbf{A}^\intercal\vec{\mathbf{x}}\|^2$$

which shows that

$$\lambda\,\|\vec{\mathbf{x}}\|^2 = \|\mathbf{A}^\intercal\vec{\mathbf{x}}\|^2$$

Since $\|\vec{\mathbf{x}}\|^2, \|\mathbf{A}^\intercal\vec{\mathbf{x}}\|^2$ are both non-negative, λ is also non-negative. □

We are now ready to (partially) prove the main result of this section.

Proof of Theorem 20.3.1. We prove the case where $k = 1$. Recall that we want to find a vector $\vec{\mathbf{u}}$ that minimizes the error of the low-dimensional representation:

$$\sum_{i=1}^{N} \left\| \vec{\mathbf{v}}_i - \widehat{\vec{\mathbf{v}}}_i \right\|^2$$

where $\widehat{\vec{\mathbf{v}}}_i$ is the low-dimensional representation of $\vec{\mathbf{v}}_i$ that can be computed as

$$\widehat{\vec{\mathbf{v}}}_i = (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}$$

by the result of Problem 7.1.3. Now by Proposition 20.1.15, we see that

$$\begin{aligned}\sum_{i=1}^{N} \|\vec{\mathbf{v}}_i - (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}\|^2 &= \sum_{i=1}^{N} \left(\|\vec{\mathbf{v}}_i\|^2 - \|(\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})\vec{\mathbf{u}}\|^2 \right) \\ &= \sum_{i=1}^{N} \left(\|\vec{\mathbf{v}}_i\|^2 - (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})^2 \right)\end{aligned}$$

Since we are already given a fixed set of vectors $\vec{\mathbf{v}}_i$, we cannot change the values of $\|\vec{\mathbf{v}}_i\|^2$. Therefore, minimizing the last term of the equation above amounts to maximizing $\sum_{i=1}^{N} (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})^2$. Notice that

$$\sum_{i=1}^{N} (\vec{\mathbf{v}}_i \cdot \vec{\mathbf{u}})^2 = \|\mathbf{A}^\intercal \vec{\mathbf{u}}\|^2 = \vec{\mathbf{u}}^\intercal \mathbf{A}\mathbf{A}^\intercal \vec{\mathbf{u}}$$

By Theorem 20.3.2 and by Theorem 20.3.4, there is an orthonormal basis $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_n\}$ of $\mathbb{R}^n$ that consist of the eigenvectors of the matrix $\mathbf{A}\mathbf{A}^\intercal$. Let λ_i be the eigenvalue corresponding to the eigenvector $\vec{\mathbf{u}}_i$. Then similarly to the proof of Theorem 20.3.3, we can represent any vector $\vec{\mathbf{u}}$ as a linear combination of the eigenvectors as

$$\vec{\mathbf{u}} = \alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n$$

Then we have $\sum_{i=1}^{n} \alpha_i^2 = 1$ and

$$\begin{aligned}\vec{\mathbf{u}}^\intercal \mathbf{A}\mathbf{A}^\intercal \vec{\mathbf{u}} &= (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n)^\intercal \mathbf{A}\mathbf{A}^\intercal (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n) \\ &= (\alpha_1 \vec{\mathbf{u}}_1 + \alpha_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \vec{\mathbf{u}}_n)^\intercal (\alpha_1 \lambda_1 \vec{\mathbf{u}}_1 + \alpha_2 \lambda_2 \vec{\mathbf{u}}_2 + \ldots + \alpha_n \lambda_n \vec{\mathbf{u}}_n) \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \lambda_j (\vec{\mathbf{u}}_i \cdot \vec{\mathbf{u}}_j) \\ &= \sum_{i=1}^{n} \alpha_i^2 \lambda_i\end{aligned}$$

Again, the allocation of α_i's that maximize $\sum_{i=1}^{n} \alpha_i^2 \lambda_i$ while maintaining $\sum_{i=1}^{n} \alpha_i^2 = 1$ is assigning $\alpha_i = \pm 1$ to the eigenvector corresponding to the highest value of λ_i. □