Probability Theory Review for Machine Learning

Samuel Ieong

November 6, 2006

1 Basic Concepts

Broadly speaking, probability theory is the mathematical study of uncertainty. It plays a central role in machine learning, as the design of learning algorithms often relies on probabilistic assumption of the data. This set of notes attempts to cover some basic probability theory that serves as a background for the class.

1.1 Probability Space

When we speak about probability, we often refer to the probability of an *event* of uncertain nature taking place. For example, we speak about the probability of rain next Tuesday. Therefore, in order to discuss probability theory formally, we must first clarify what the possible events are to which we would like to attach probability.

Formally, a *probability space* is defined by the triple $(\Omega, \mathcal{F}, P)$, where

- Ω is the *space of possible outcomes* (or *outcome space*),
- $\mathcal{F} \subseteq 2^{\Omega}$ (the power set of Ω) is the *space of (measurable) events* (or *event space*),
- P is the *probability measure* (or *probability distribution*) that maps an event $E \in \mathcal{F}$ to a real value between 0 and 1 (think of P as a function).

Given the outcome space Ω, there is some restrictions as to what subset of 2^{Ω} can be considered an event space $\mathcal{F}$:

- The trivial event Ω and the empty event $\emptyset$ is in $\mathcal{F}$.
- The event space $\mathcal{F}$ is closed under (countable) union, i.e., if $\alpha, \beta \in \mathcal{F}$, then $\alpha \cup \beta \in \mathcal{F}$.
- The even space $\mathcal{F}$ is closed under complement, i.e., if $\alpha \in \mathcal{F}$, then $(\Omega \setminus \alpha) \in \mathcal{F}$.

Example 1. *Suppose we throw a (six-sided) dice. The space of possible outcomes* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *We may decide that the events of interest is whether the dice throw is odd or even. This event space will be given by* $\mathcal{F} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$.

Note that when the outcome space Ω is finite, as in the previous example, we often take the event space $\mathcal{F}$ to be 2^Ω. This treatment is not fully general, but it is often sufficient for practical purposes. However, when the outcome space is infinite, we must be careful to define what the event space is.

Given an event space $\mathcal{F}$, the probability measure P must satisfy certain axioms.

- (non-negativity) For all $\alpha \in \mathcal{F}$, $P(\alpha) \geq 0$.
- (trivial event) $P(\Omega) = 1$.
- (additivity) For all $\alpha, \beta \in \mathcal{F}$ and $\alpha \cap \beta = \emptyset$, $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.

Example 2. *Returning to our dice example, suppose we now take the event space $\mathcal{F}$ to be 2^Ω. Further, we define a probability distribution P over $\mathcal{F}$ such that*

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = 1/6$$

then this distribution P completely specifies the probability of any given event happening (through the additivity axiom). For example, the probability of an even dice throw will be

$$P(\{2, 4, 6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 1/6 + 1/6 + 1/6 = 1/2$$

since each of these events are disjoint.

1.2 Random Variables

Random variables play an important role in probability theory. The most important fact about random variables is that they are **not** variables. They are actually **functions** that map outcomes (in the outcome space) to real values. In terms of notation, we usually denote random variables by a capital letter. Let's see an example.

Example 3. *Again, consider the process of throwing a dice. Let X be a random variable that depends on the outcome of the throw. A natural choice for X would be to map the outcome i to the value i, i.e., mapping the event of throwing an "one" to the value of 1. Note that we could have chosen some strange mappings too. For example, we could have a random variable Y that maps all outcomes to 0, which would be a very boring function, or a random variable Z that maps the outcome i to the value of 2^i if i is odd and the value of $-i$ if i is even, which would be quite strange indeed.*

In a sense, random variables allow us to abstract away from the formal notion of event space, as we can define random variables that capture the appropriate events. For example, consider the event space of odd or even dice throw in Example 1. We could have defined a random variable that takes on value 1 if outcome i is odd and 0 otherwise. These type of binary random variables are very common in practice, and are known as *indicator variables*, taking its name from its use to indicate whether a certain event has happened. So why did we introduce event space? That is because when one studies probability theory (more rigorously) using measure theory, the distinction between outcome space and event space will be very important. This topic is too advanced to be covered in this short review note. In any case, it is good to keep in mind that event space is not always simply the power set of the outcome space.

From here onwards, we will talk mostly about probability with respect to random variables. While some probability concepts can be defined meaningfully without using them, random variables allow us to provide a more uniform treatment of probability theory. For notations, the probability of a random variable X taking on the value of a will be denoted by either

$$P(X = a) \qquad \text{or} \qquad P_X(a)$$

We will also denote the range of a random variable X by $Val(X)$.

1.3 Distributions, Joint Distributions, and Marginal Distributions

We often speak about the *distribution* of a variable. This formally refers to the probability of a random variable taking on certain values. For example,

Example 4. *Let random variable X be defined on the outcome space Ω of a dice throw (again!). If the dice is fair, then the distribution of X would be*

$$P_X(1) = P_X(2) = \cdots = P_X(6) = 1/6$$

Note that while this example resembles that of Example 2, they have different semantic meaning. The probability distribution defined in Example 2 is over **events**, whereas the one here is defined over **random variables**.

For notation, we will use $P(X)$ to denote the distribution of the random variable X.

Sometimes, we speak about the distribution of more than one variables at a time. We call these distributions *joint distributions*, as the probability is determined jointly by all the variables involved. This is best clarified by an example.

Example 5. *Let X be a random variable defined on the outcome space of a dice throw. Let Y be an indicator variable that takes on value 1 if a coin flip turns up head and 0 if tail. Assuming both the dice and the coin are fair, the joint distribution of X and Y is given by*

P	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$
$Y = 0$	1/12	1/12	1/12	1/12	1/12	1/12
$Y = 1$	1/12	1/12	1/12	1/12	1/12	1/12

As before, we will denote the probability of X taking value a and Y taking value b by either the long hand of $P(X = a, Y = b)$, or the short hand of $P_{X,Y}(a, b)$. We refer to their joint distribution by $P(X, Y)$.

Given a joint distribution, say over random variables X and Y, we can talk about the *marginal distribution* of X or that of Y. The marginal distribution refers to the probability distribution of a random variable on its own. To find out the marginal distribution of a random variable, we *sum out* all the other random variables from the distribution. Formally, we mean

$$P(X) = \sum_{b \in Val(Y)} P(X, Y = b) \tag{1}$$

The name of marginal distribution comes from the fact that if we add up all the entries of a row (or a column) of a joint distribution, and write the answer at the end (i.e., margin) of it, this will be the probability of the random variable taking on that value. Of course, thinking in this way only helps when the joint distribution involves two variables.

1.4 Conditional Distributions

Conditional distributions are one of the key tools in probability theory for reasoning about uncertainty. They specify the distribution of a random variable when the value of another random variable is known (or more generally, when some event is known to be true).

Formally, conditional probability of $X = a$ *given* $Y = b$ is defined as

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)} \tag{2}$$

Note that this is not defined when the probability of $Y = b$ is 0.

Example 6. *Suppose we know that a dice throw was odd, and want to know the probability of an "one" has been thrown. Let X be the random variable of the dice throw, and Y be an indicator variable that takes on the value of* 1 *if the dice throw turns up odd, then we write our desired probability as follows:*

$$P(X = 1|Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{1/6}{1/2} = 1/3$$

The idea of conditional probability extends naturally to the case when the distribution of a random variable is conditioned on several variables, namely

$$P(X = a|Y = b, Z = c) = \frac{P(X = a, Y = b, Z = c)}{P(Y = b, Z = c)}$$

As for notations, we write $P(X|Y = b)$ to denote the distribution of random variable X when $Y = b$. We may also write $P(X|Y)$ to denote a set of distributions of X, one for each of the different values that Y can take.

1.5 Independence

In probability theory, *independence* means that the distribution of a random variable does *not* change on learning the value of another random variable. In machine learning, we often make such assumptions about our data. For example, the training samples are assumed to be drawn independently from some underlying space; the label of sample i is assumed to be independent of the features of sample j $(i \neq j)$.

Mathematically, a random variable X is independent of Y when

$$P(X) = P(X|Y)$$

(Note that we have dropped what values X and Y are taking. This means the statement holds true for any values X and Y may take.)

Using Equation (2), it is easy to verify that if X is independent of Y, then Y is also independent of X. As a notation, we write $X \perp Y$ if X and Y are independent.

An equivalent mathematical statement about the independence of random variables X and Y is

$$P(X, Y) = P(X)P(Y)$$

Sometimes we also talk about *conditional independence*, meaning that if we know the value of a random variable (or more generally, a set of random variables), then some other random variables will be independent of each other. Formally, we say "X and Y are *conditionally* independent given Z" if

$$P(X|Z) = P(X|Y, Z)$$

or, equivalently,

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

An example of conditional independence that we will se in class is the *Naïve Bayes* assumption. This assumption is made in the context of a learning algorithm for learning to classify emails as spams or non-spams. It assumes that the probability of a word x appearing in the email is conditionally independent of a word y appearing given whether the email is spam or not. This clearly is not without loss of generality, as some words almost invariably comes in pair. However, as it turns out, making this simplifying assumption does not hurt the performance much, and in any case allow us to learn to classify spams rapidly. Details can be found in Lecture Notes 2.

1.6 Chain Rule and Bayes Rule

We now present two basic yet important rules for manipulating that relates joint distributions and conditional distributions. The first is known as the *Chain Rule*. It can be seen as a generalization of Equation (2) to multiple random variables.

Theorem 1 (Chain Rule)**.**

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1) \cdots P(X_n|X_1, X_2, \ldots, X_{n-1}) \tag{3}$$

The Chain Rule is often used to evaluate the joint probability of some random variables, and is especially useful when there are (conditional) independence across variables. Notice there is a choice in the order we unravel the random variables when applying the Chain Rule; picking the right order can often make evaluating the probability much easier.

The second rule we are going to introduce is the *Bayes Rule*. The Bayes Rule allows us to compute the conditional probability $P(X|Y)$ from $P(Y|X)$, in a sense "inverting" the conditions. It can be derived simply from Equation (2) as well.

Theorem 2 (Bayes Rule)**.**

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{4}$$

And recall that if $P(Y)$ is not given, we can always apply Equation (1) to find it.

$$P(Y) = \sum_{a \in Val(X)} P(X = a, Y) = \sum_{a \in Val(X)} P(Y|X = a)P(X = a)$$

This application of Equation (1) is sometimes referred to as the *law of total probability*.

Extending the Bayes Rule to the case of multiple random variables can sometimes be tricky. Just to be clear, we would give a few examples. When in doubt, one can always refer to how conditional probabilities are defined and work out the details.

Example 7. *Let's consider the following conditional probabilities:* $P(X, Y|Z)$ *and* $(X|Y, Z)$*.*

$$P(X, Y|Z) = \frac{P(Z|X, Y)P(X, Y)}{P(Z)} = \frac{P(Y, Z|X)P(X)}{P(Z)}$$

$$P(X|Y, Z) = \frac{P(Y|X, Z)P(X, Z)}{P(Y, Z)} = \frac{P(Y|X, Z)P(X|Z)P(Z)}{P(Y|Z)P(Z)} = \frac{P(Y|X, Z)P(X|Z)}{P(Y|Z)}$$

2 Defining a Probability Distribution

We have been talking about probability distributions for a while. But how do we define a distribution? In a broad sense, there are two classes of distribution that require seemingly different treatments (these can be unified using measure theory). Namely, *discrete* distributions and *continuous* distributions. We will discuss how distributions are specified next.

Note that this discussion is distinct from how we can efficiently *represent* a distribution. The topic of efficient representation of probability distribution is in fact a very important and active research area that deserves its own course. If you are interested to learn more about how to efficiently represent, reason, and perform learning on distributions, you are advised to take CS228: Probabilistic Models in Artificial Intelligence.

2.1 Discrete Distribution: Probability Mass Function

By a discrete distribution, we mean that the random variable of the underlying distribution can take on only *finitely many* different values (or that the outcome space is finite).

To define a discrete distribution, we can simply enumerate the probability of the random variable taking on each of the possible values. This enumeration is known as the *probability mass function*, as it divides up a unit mass (the total probability) and places them on the different values a random variable can take. This can be extended analogously to joint distributions and conditional distributions.

2.2 Continuous Distribution: Probability Density Function

By a continuous distribution, we mean that the random variable of the underlying distribution can take on *infinitely many* different values (or that the outcome space is infinite).

This is arguably a trickier situation than the discrete case, since if we place a non-zero amount of mass on each of the values, the total mass will add up to infinity, which violates the requirement that the total probaiblity must sum up to one.

To define a continuous distribution, we will make use of *probability density function* (PDF). A probability density function, f, is a *non-negative*, *integrable* function such that

$$\int_{Val(X)} f(x)dx = 1$$

The probability of a random variable X distributed according to a PDF f is computed as follows

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Note that this, in particular, implies that the probability of a continuously distributed random variable taking on any given single value is zero.

Example 8 (Uniform distribution)**.** *Let's consider a random variable X that is uniformly distributed in the range $[0,1]$. The corresponding PDF would be*

$$f(x) = \begin{cases} 1 & \textit{if } 0 \leq x \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

We can verify that $\int_0^1 1\ dx$ is indeed 1, and therefore f is a PDF. To compute the probability of X smaller than a half,

$$P(X \leq 1/2) = \int_0^{1/2} 1\ dx = [x]_0^{1/2} = 1/2$$

More generally, suppose X is distributed uniformly over the range $[a,b]$, then the PDF would be

$$f(x) = \begin{cases} \frac{1}{b-a} & \textit{if } a \leq x \leq b \\ 0 & \textit{otherwise} \end{cases}$$

Sometimes we will also speak about *cumulative distribution function.* It is a function that gives the probability of a random variable being smaller than some value. A cumulative distribution function F is related to the underlying probability density function f as follows:

$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx$$

and hence $F(x) = \int f(x)dx$ (in the sense of indefinite integral).

To extend the definition of continuous distribution to joint distribution, the probability density function is extended to take multiple arguments, namely,

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots, a_n \leq X_n \leq n_1) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

To extend the definition of conditional distribution to continuous random variables, we ran into the problem that the probability of a continuous random variable taking on a single value is 0, so Equation (2) is not well defined, since the denominator equals 0. To define the conditional distribution of a continuous variable, let $f(x, y)$ be the joint distribution of X and Y. Through application of analysis, we can show that the PDF, $f(y|x)$, underlying the distribution $P(Y|X)$ is given by

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

For example,

$$P(a \leq Y \leq b | X = c) = \int_{a}^{b} f(y|c)dy = \int_{a}^{b} \frac{f(c, y)}{f(c)} dy$$

3 Expectations and Variance

3.1 Expectations

One of the most common operations we perform on a random variable is to compute its *expectation*, also known as its *mean*, *expected value*, or *first moment*. The expectation of a random variable, denoted by $E(X)$, is given by

$$E(X) = \sum_{a \in Val(X)} aP(X = a) \qquad \text{or} \qquad E(X) = \int_{a \in Val(X)} xf(x)\, dx \tag{5}$$

Example 9. *Let X be the outcome of rolling a fair dice. The expectation of X is*

$$E(X) = (1)\frac{1}{6} + (2)\frac{1}{6} + \cdots + 6\frac{1}{6} = 3\frac{1}{2}$$

We may sometimes be interested in computing the expected value of some function f of a random variable X. Recall, however, that a random variable is also a function itself, so the easiest way to think about this is that we define a new random variable $Y = f(X)$, and compute the expected value of Y instead.

When working with indicator variables, a useful identify is the following:

$$E(X) = P(X = 1) \qquad \text{for indicator variable } X$$

When working with the sums of random variables, one of the most important rule is the *linearity of expectations.*

Theorem 3 (Linearity of Expectations)**.** *Let X_1, X_2, ..., X_n be (possibly dependent) random variables,*

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) \tag{6}$$

The linearity of expectations is very powerful because there are no restrictions on whether the random variables are independent or not. When we work on products of random variables, however, there is very little we can say in general. However, when the random variables are independent, then

Theorem 4. *Let X and Y be independent random variables,*

$$E(XY) = E(X)E(Y)$$

3.2 Variance

The *variance* of a distribution is a measure of the "spread" of a distribution. Sometimes it is also referred to as the *second moment*. It is defined as follows:

$$Var(X) = E\left((X - E(X))^2\right) \tag{7}$$

The variance of a random variable is often denoted by σ^2. The reason that this is squared is because we often want to find out σ, known as the *standard deviation*. The variance and the standard deviation is related (obviously) by $\sigma = \sqrt{Var(X)}$.

To find out the variance of a random variable X, it's often easier to compute the following instead

$$Var(X) = E(X^2) - (E(X))^2$$

Note that unlike expectation, variance is not a linear function of a random variable X. In fact, we can verify that the variance of $(aX + b)$ is

$$Var(aX + b) = a^2 Var(X)$$

If random variables X and Y are independent, then

$$Var(X + Y) = Var(X)Var(Y) \qquad \text{if } X \perp Y$$

Sometimes we also talk about the *covariance* of two random variables. This is a measure of how "closely related" two random variables are. Its definition is as follows.

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

4 Some Important Distributions

In this section, we will review some of the probability distributions that we will see in this class. This is by no means a comprehensive list of distribution that one should know. In particular, distributions such as the geoemtric, hypergeometric, and binomial distributions, which are very useful in their own right and studied in introductory probability theory, are not reviewed here.

4.1 Bernoulli

The *Bernoulli distribution* is one of the most basic distribution. A random variable distributed according to the Bernoulli distribution can take on two possible values, $\{0, 1\}$. It can be specified by a single parameter p, and by convention we take p to be $P(X = 1)$. It is often used to indicate whether a trail is successful or not.

Sometimes it is useful to write the probability distribution of a Bernoulli random variable X as follows

$$P(X) = p^x(1-p)^{1-x}$$

An example of the Bernoulli distribution in action is the classification task in Lecture Notes 1. To develop the logistic regression algorithm for the task, we assume that the labels are distributed according to the Bernoulli distribution given the features.

4.2 Poisson

The *Poisson distribution* is a very useful distribution that deals with the arrival of events. It measures probaiblity of the number of events happening over a fixed period of time, given a fixed average rate of occurrence, and that the events take place independently of the time since the last event. It is parametrized by the average arrival rate λ. The probability mass function is given by:

$$P(X = k) = \frac{\exp(-\lambda)\lambda^k}{k!}$$

The mean value of a Poisson random variable is λ, and its variance is also λ.

We will get to work on a learning algorithm that deals with Poisson random variables in Homework 1, Problem 3.

4.3 Gaussian

The *Gaussian distribution*, also known as the *normal distribution*, is one of the most "versatile" distributions in probability theory, and appears in a wide variety of contexts. For example, it can be used to approximate the binomial distribution when the number of experiments is large, or the Poisson distribution when the average arrival rate is high. It is also related to the Law of Large Numbers. For many problems, we will also often assume that when noise in the system is Gaussian distributed. The list of applications is endless.

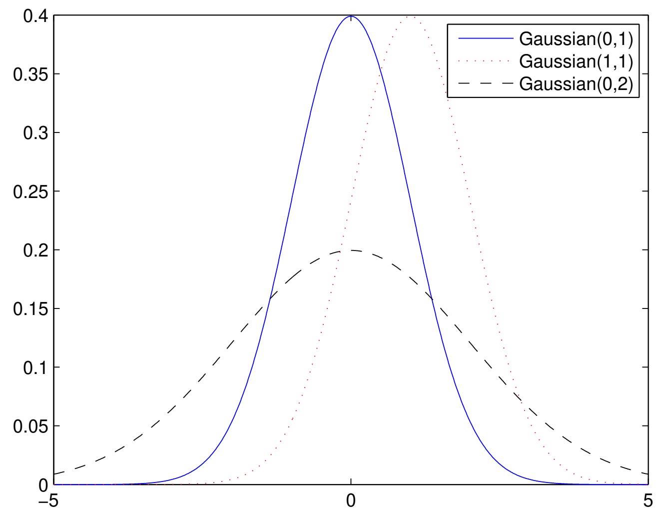

Figure 1: Gaussian distributions under different mean and variance

The Gaussian distribution is determined by two parameters: the mean μ and the variance σ^2. The probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{8}$$

To get a better sense of how the distribution changes with respect to the mean and the variance, we have plotted three different Gaussian distributions in Figure 1.

In our class, we will sometimes work with multi-variate Gaussian distributions. A k-dimensional multi-variate Gaussian distribution is parametrized by (μ, Σ), where μ is now a *vector* of means in $\mathbb{R}^k$, and Σ is the *covariance matrix* in $\mathbb{R}^{k\times k}$, in other words, $\Sigma_{ii} = Var(X_i)$ and $\Sigma_{ij} = Cov(X_i, X_j)$. The probability density function is now defined over vectors of input, given by

$$f(\mathbf{x}) = \frac{1}{\sqrt{2\pi^k|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{9}$$

(Recall that we denote the determinant of a matrix A by $|A|$, and its inverse by A^{-1})

To get a better sense of how a multi-variate Gaussian distribution depends on the covariance matrix, we can look at the figures in Lecture Notes 2, Pages 3—4.

Working with a multi-variate Gaussian distribution can be tricky and daunting at times. One way to make our lives easier, at least as a way to get intuition on a problem, is to assume that the covariances are zero when we first attempt a problem. When the covariances are zero, the determinant $|\Sigma|$ will simply be the product of the variances, and the inverse Σ^{-1} can be found by taking the inverse of the diagonal entries of Σ.

5 Working with Probabilities

As we will be working with probabilities and distributions a lot in this class, listed below are a few tips about efficient manipulation of distributions.

5.1 The log trick

In machine learning, we generally assume the independence of different samples. Therefore, we often have to deal with the product of a (large) number of distributions. When our goal is to optimize functions of such products, it is often easier if we first work with the logarithm of such functions. As the logarithmic function is a strictly increasing function, it will not distort where the maximum is located (although, most certainly, the maximum value of the function before and after taking logarithm will be different).

As an example, consider the likelihood function in Lecture Notes 1, Page 17.

$$L(\theta) = \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

I dare say this is a pretty mean-looking function. But by taking the logarithm of it, termed log-likelihood function, we have instead

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

Not the world's prettiest function, but at least it's more manageable. We can now work on one term (i.e., one training sample) at a time, because they are summed together rather than multiplied together.

5.2 Delayed Normalization

Because probability has to sum up to one, we often have to deal with normalization, especially with continuous distribution. For example, for Gaussian distributions, the term outside of the exponent is to ensure that the integral of the PDF evaluates to one. When we are sure that the end product of some algebra will be a probability distribution, or when we are finding the optimum of some distributions, it's often easier to simply denote the normalization constant to be Z, and not worry about computing the normalization constant all the time.

5.3 Jenson's Inequality

Sometimes when we are evaluating the expectation of a function of a random variable, we may only need a bound rather than its exact value. In these situations, if the function is convex or concave, Jenson's inequality allows us to derive a bound by evaluating the value of the function at the expectation of the random variable itself.

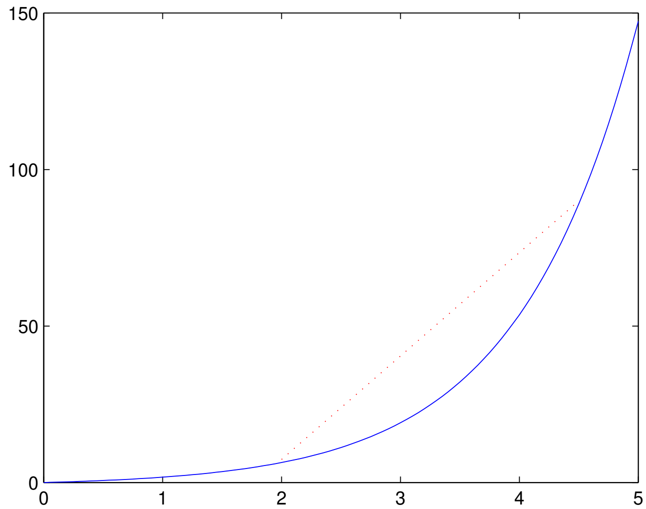

Figure 2: Illustration of Jenson's Inequality

Theorem 5 (Jenson's Inequality)**.** *Let X be a random variable, and f be a convex function. Then*

$$f(E(X)) \leq E(f(X))$$

If f is a concave function, then

$$f(E(X)) \geq E(f(X))$$

While we can show Jenson's inequality by algebra, it's easiest to understand it through a picture. The function in Figure 2 is a convex function. We can see that a straight line between any two points on the function always lie above the function. This shows that if a random variable can take on only two values, then Jenson's inequality holds. It is relatively straight forward to extend this to general random variables.