02 Probability Theory

June 3, 2021

Information: *A brief review of basic concepts of Probability, Statistics and Information related theory which is often used in Machine Learning and Deep Learning*

Written by: *Zihao Xu*

Last update date: *06.03.2021*

1 Basic concepts

1.1 Introduction

1.1.1 Definition

- Probability Theory is a mathematical frame work for representing uncertain statements
  - Provides a means of quantifying uncertainty as well as axioms for deriving new uncertainty statements
  - Allows making uncertain statements and reasoning in the presence of uncertainty

1.1.2 Motivation

- Machine learning must always deal with uncertain quantities and sometimes stochastic (nondeterministic) quantities resulting from:
  - Inherent stochasticity in the system being modeled
  - Incomplete observability
  - Incomplete modeling
- It's usually more practical to use a *simple but uncertain rule* rather than a complex but certain one
  - "Most birds fly" versus "Birds fly, except for …"
- Uncertainty can be modeled by probability

1.1.3 Interpretation of probability

- **Frequentist probability**
  - Probability theory was originally developed to analyze the *frequencies of events*, which are often *repeatable*
  - When we say that an outcome has a probability p of occurring, it means that if we repeated the experiment infinitely times, then a proportion p of the repetitions would result in that outcome
  - Not seemingly applicable to propositions that are not repeatable
- **Bayesian probability**

  - Use probability to represent a *degree of belief* for an outcome to occur, with 1 indicating absolute certainty that the outcome occurs and 0 indicating absolute certainty that the outcome doesn't occur
  - For properties that we expect common sense reasoning about uncertainty to have, we need to treat Bayesian probabilities as behaving exactly the same as frequentist probabilities.
- **The extension logic to deal with uncertainty**
  - Probability can be seen as the extension of logic to deal with uncertainty. Probability theory provides a set of formal rules for determining the likelihood of a proposition being true given the likelihood of other propositions

1.2 Probability Space

- A *probability space* is defined by the triple $(\Omega, \mathcal{F}, P)$, where:
  - Ω is the *space of possible outcomes* (or outcome space)
  - $\mathcal{F} \subseteq 2^{\Omega}$ 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 an 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 $\nvdash$ is in $\mathcal{F}$
  - The event space $\mathcal{F}$ is closed under (countable) union
    - if $\alpha, \beta \in \mathcal{F}$, then $\alpha \cup \beta \in \mathcal{F}$
  - The event space $\mathcal{F}$ is closed under complement
    - if $\alpha \in \mathcal{F}$, then $(\Omega \setminus \alpha) \in \mathcal{F}$
- Notice: Event space is not always simply the power set of the outcome space

1.3 Random Variables

1.3.1 Definition

- A *random variable* is a function maps outcomes in the outcome space Ω to real values
  - **Not random nor a variable**
- May be discrete or continuous
  - Discrete: When the image of a random variable is countable, the random variable is called a *discrete random variable*
  - Continuous: When the image of random variable is uncountably infinite (usually an interval), then the random variable is called a *continuous random variable*
  - Image:In mathematics, the image of a function is the set of all output values it may produce.

1.3.2 Notation

- Upper case letters to represent random variables
  - $X, Y, Z, ...$
- Lower case letters to represent the values of random variables
  - $x, y, z, ...$

1.4 Probability Distributions

- A **probability distribution** is a description of how likely a random variable or set of random variables is to take on each of its possible states. The way to describe probability distribution depends on whether the variables are discrete or continuous.

1.4.1 Discrete Variables and Probability Mass Functions

- **Probability Mass Function (PMF)**: A function that gives the probability a discrete random variable is exactly equal to some value
  - Called probability mass function as it divides up a unit mass (the total probability) and places them on different values a random variable can take
  - Sometimes called *discrete density function*
  - The primary means of defining a discrete probability distribution
- **Notations for PMF**:
  - Let X be a discrete random variable, the probability mass function of X is:

$$P(X = x) = \text{Probability that the random variable } X \text{ takes the value } x$$

    * Can be written in $p_X(x)$ or even directly $p(x)$ when there is no ambiguity
  - Sometimes we define a variable first, then use $\sim$ notation to specify which distribution it follows later: $X \sim P(X)$
- **Properties of PMF**: Given that X is a *discrete random variable*, then:
  - The domain of p must be the set of all possible states of X
  - $\forall x \in X, 0 \leq p(x) \leq 1$. An impossible event has probability 0, and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring
  - $\sum_x p(x) = 1$.Usually refer to this property as being *normalized*.
  - Probability of X taking either the value x_1 or the value x_2 (assuming $x_1 \neq x_2$) is:

$$P(X = x_1 \text{ or } X = x_2) \equiv P(X \in \{x_1, x_2\}) = P(X = x_1) + P(X = x_2) = p(x_1) + p(x_2)$$

    * More generally, the probability that the random variable X takes any value in a set A is given by $P(X \in A) = \sum_{x \in A} p(x)$
  - Let $g(x)$ be a function and $Y = g(X)$ be a new random variable, then the PMF would be:

$$P(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} p(x)$$

1.4.2 Continuous Variables and Probability Density Functions

- **Cumulative Distribution Function(CDF)**:A function that gives the probability that a random variable will take a value less than some value
  - The random variable can be either discrete or continuous
  - Sometimes called *distribution function*
- **Notations for CDF**
  - Let X be a random variable, then its cumulative distribution function is denoted by:

$$F(x) = P(X \leq x)$$

- **Properties of CDF**: Given that X is a random variable, then:

  - The domain of F must be the set of all possible states of X
  - $F(x)$ is an increasing function for x
  - $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$
  - $F(+\infty) = \lim_{x \to -\infty} F(x) = 1$
  - $P(a \leq X \leq b) = F(b) - F(a)$
- **Probability Density Function (PDF)**: A function whose value at any given sample in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a *relative likelihood* that the value of the random variable would equal that sample
  - Definition only for continuous random variables
  - While the *absolute likelihood* for a continuous random variable to take on any particular value is 0 (since there is an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer how much more likely it is that the random variable would equal one sample compared to the other sample.
- **Notations for PDF**
  - Let X be a continuous random variable, then its probability density function is denoted by:
$$p(x) \simeq \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}$$
for some small Δx according to definition
  - Observation:
$$\begin{aligned} p(x) &= \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \\ &= F'(x) \\ &= \frac{\mathrm{d}F(x)}{\mathrm{d}x} \end{aligned}$$
- **Properties of PDF**: Given that X is a *continuous random variable*, then:
  - The domain of p must be the set of all possible states of X
  - $\forall x \in X, p(x) \geq 0$. Note that $p(x) \leq 1$ is not required
  - $\int p(x)dx = 1$
  - $\int_a^b p(x)dx = F(b) - F(a) = P(a \leq X \leq b)$
  - For any borel subset A of the real numbers:
$$P(X \in A) = \int_A p(x)dx$$
. This property holds even for random vectors
  - Let $g(x)$ be a function and $Y = g(X)$ be a new continuous random variable, then the PDF of Y would be:
$$p(y) = p(x = g^{-1}(y))|\frac{d}{dy}(g^{-1}(y))|$$

1.5 Expectations and Variance

1.5.1 Expectations

- **Definition**: The **expectation**, or **expected value**, of a random variable is denoted by $\mathbb{E}(X)$

  - For a discrete random variable X:

$$\mathbb{E}[X] = \sum_x xp(x)$$

    where $P(x)$ is the probability mass function and the sum is over all possible values
  - For a continuous random variable X:

$$\mathbb{E}[X] = \int_x xp(x)dx$$

    where $p(x)$ is the probability density function and the integral is over all possible values
  - Can think of the expectation as the value of the random variable that one should "expect" to get, but notice that the expectation of a random variable is usually not in the possible values of it.
- **Properties**: Let X be a random variable, then:
  - Let $g(x)$ be a function:
    - For a discrete random variable X:

$$\mathbb{E}[g(X)] = \sum_x g(x)p(x)$$

    - For a continuous random variable X:

$$\mathbb{E}[g(X)] = \int_x g(x)p(x)dx$$

  - Take any constant c:

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c$$

  - Take any constant λ:

$$\mathbb{E}[\lambda X] = \lambda \mathbb{E}[X]$$

1.5.2 Variance

- **Definition**: The **variance** measures how much the value of a function of a random variable vary as sampling different values from its probability distribution
  - For a discrete/continuous random variable X:

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

  - Can think of the variance as the spread of the random variable around its expectation
  - Square root of the variance is known as the **standard deviation**
- **Properties**:
  - $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  - Take any constant c:

$$\mathbb{V}[X + c] = \mathbb{V}[X]$$

  - Take any constant λ:

$$\mathbb{V}[\lambda X] = \lambda^2 \mathbb{V}[X]$$

1.6 Some Important Distributions

1.6.1 Bernoulli

- The *Bernoulli Distribution* is one of the most basic distribution. It is a distribution over a single binary random variable and is controlled by a single parameter $\theta \in [0, 1]$, which gives the probability of the random variable being equal to 1.
- For a *discrete random variable* X following a Bernoulli distribution with parameter θ, the notation is:

$$X \sim \text{Bernoulli}(\theta)$$

- Properties:
  - $P(X = 1) = \theta$
  - $P(X = 0) = 1 - \theta$
  - $P(X = x) = \theta^x(1 - \theta)^{1-x}$ where $x = 0$ or $x = 1$
  - $\mathbb{E}[X] = \theta$
  - $\mathbb{V}[X] = \theta(1 - \theta)$

1.6.2 Categorical Distribution

- The *Categorical Distribution*, also called *Multinoulli Distribution*, is a distribution over a single discrete random variable with k different states, where k is finite. It is parameterized by a vector $\mathbf{p} \in [0, 1]^{k-1}$, where p_i gives the probability of the $i - th$ state. The final, k-th state's probability is given by $1 - \mathbf{1}^T\mathbf{p}$, where $\mathbf{1}^T\mathbf{p}$ should be constrained to less than or equal to 1.
- For a *discrete random variable* X following a Categorical distribution with parameter $\mathbf{p}$. the notation is:

$$X \sim \text{Categorical}(p_1, p_2, \cdots, p_{k-1}, 1 - \mathbf{1}^T\mathbf{p})$$

1.6.3 Continuous Uniform Distribution

- The *Continuous Uniform Distribution*, also called *rectangular distribution*, models a random variable the takes values within an interval all with equal probability. The interval is defined by the parameters a and b, which are the minimum and maximum values. The interval can either be *closed* (i.e. $[a, b]$) or *open* (e.g. (a, b))
- For a *continuous random variable* X following a Uniform distribution in the interval $[a, b]$, the notation is:

$$X \sim U([a, b])$$

- Probability density function:
  - $p(x) = \begin{cases} c & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
  - $\int_a^b p(x)dx = 1 \Rightarrow c = \frac{1}{b-a}$
- Cumulative distribution function:
  - $F(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x - a}{b - a} & \text{if } a \leq x \leq b \\ 1 & \text{otherwise} \end{cases}$
- Expectation:
  - $\mathbb{E}[X] = \frac{a+b}{2}$

- Variation:
  - $\mathbb{V}[X] = \frac{(b-a)^2}{12}$

1.6.4 Gaussian Distribution

- The *Gaussian Distribution*, also know as the *normal distribution*, is the most commonly used distribution over real numbers. It models a random variable that take values from any real value. The distribution is controlled by two parameters $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$. The values are concentrated around μ and the variance σ^2 determines how spread out the function values are around the mean.
- For a *continuous random variable* following a Gaussian distribution with expectation μ and variance σ^2, the notation is:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

- Probability density function:
  - $p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$
  - When we need to frequently evaluate the PDF with different parameter values, a more efficient way of parameterizing the distribution is to use a parameter $\beta \in (0, \infty)$ to control the **precision**, or inverse variance, of the distribution
    - $\beta = \frac{1}{\sigma^2}$
    - $p(x) = \mathcal{N}(x|\mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x-\mu)^2\right)$
- Cumulative distribution function is not *analytically* available:
  - $F(x) = P(X \leq x) = \int_{-\infty}^{x} \mathcal{N}(\bar{x}|\mu, \sigma^2) d\bar{x}$
- Expectation:
  - $\mathbb{E}[X] = \mu$
- Variation:
  - $\mathbb{V}[X] = \sigma^2$
- Remark:
  - $P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95.45\%$
  - $P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.73\%$

1.6.5 Standard Normal Distribution

- The *Standard Normal Distribution* is the simplest case of a *Gaussian Distribution* where $\mu = 0$ and $\sigma = 1$, denoted as:

$$Z \sim \mathcal{N}(0, 1)$$

- Probability density function:
  - $\phi(z) = \mathcal{N}(z|0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$
- Cumulative distribution function is also not analytically available:
  - $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left\{-\frac{t^2}{2}\right\} dt$
- Connections between the normal and the standard normal:
  - Take a stander normal $Z \sim \mathcal{N}(0, 1)$ and two numbers μ and σ^2, make the random variable $X = \mu + \sigma Z$, then $x \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathcal{N}(\mu, \sigma^2) = \mu + \sigma Z$$

  - Take a normal $X \sim \mathcal{N}(\mu, \sigma^2)$, make the random variable $Z = \frac{X-\mu}{\sigma}$, then $Z \sim \mathcal{N}(0, 1)$

– Take a normal $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

1.6.6 Exponential and Laplace Distributions

- In the context of deep learning, a probability distribution with a sharp point at $x = 0$ is often wanted. **Exponential Distribution** can help to accomplish this.

$$p(x|\lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$$

  – The indicator function $\mathbf{1}_{x \geq 0}$ is used to assign probability zero to all negative values of x
- **Laplace distribution** allows us to place a sharp peak of probability mass at an arbitrary point μ

$$\text{Laplace}(x|\mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

1.6.7 Poisson Distribution

- The *Poisson Distribution* is a very useful distribution that deal with the arrival of events. It measures probability 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 parameterized by the average arrival rate λ.

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

- Expectation:
  – $\mathbb{E}[X] = \lambda$
- Variation:
  – $\mathbb{V}[X] = \lambda$

1.7 Collections of Random Variables

1.7.1 Joint Distributions, Marginal Distributions and Conditional Distributions

- **Joint Distributions**: Distributions over multiple random variables
  – For discrete random variables X and Y, the **joint probability mass function** of the pair (X, Y) is the function $p(x, y)$ giving the probability that $X = x$ and $Y = y$ is denoted as:

$$p(x, y) = P(X = x, Y = y)$$

    * It's non-negative: $p(x, y) \geq 0$
    * Sum over all the possible values of all random variables is 1: $\sum_x \sum_y p(x, y) = 1$
    * *Sum Rule*: Marginalizing over the values of one of the random variables gets the PMF of the other: $\sum_y p(x, y) = p(x), \sum_x p(x, y) = p(y)$. This result is called the **Marginal Distribution**.
  – For continuous random variables X and Y, the **joint probability density function** of the pair (X, Y) is the function $p(x, y)$ giving the probability that $X = x$ and $Y = y$:

$$p(x, y) = \lim_{\Delta x, \Delta y \to 0} \frac{P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y)}{\Delta x \Delta y}$$

    * It's non-negative: $p(x, y) \geq 0$
    * Sum over all the possible values of all random variables is 1: $\int \int p(x, y)dxdy = 1$
    * *Sum Rule*: Marginalizing over the values of one of the random variables gets the PDF of the other:$\int p(x, y)dy = p(x)$ and $\int p(x, y)dx = p(y)$. The result is called the **Marginal Distribution**
- **Conditional Distributions**: In many cases, the probability of some event given that some other event has happened is wanted. This called a **conditional probability**
  - For random variables X and Y, if $X = x$ is observed and the state of knowledge about Y is need to be updated, then the conditional PDF is:

$$P(Y = y|X = x) = p(y|x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x, y)}{p(x)}$$

    * The conditional probability is only defined when $P(X = x) > 0$. The conditional probability conditioned on an event that never happens is meaningless.

1.7.2 Chain Rule and Bayes Rule

- **Chain Rule** / Product Rule of Conditional Probabilities

$$P(x^{(1)}, \cdots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^{n} P(x^i|x^{(1)}, \cdots, x^{(i-1)})$$

  - Often used to evaluate the joint probability of some random variables and is especially useful when there are (conditional) independence across variables
  - Picking the right order to unravel the random variables can often make evaluating the probability much easier
  - For example:

$$\begin{aligned} p(a, b, c) &= p(a|b, c)p(b, c) \\ p(b, c) &= p(b|c)p(c) \\ p(a, b, c) &= p(a|b, c)p(b|c)p(c) \end{aligned}$$

- **Bayes Rule**

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

  - If $P(Y)$ is not given, we can always use the equation of calculating marginal probabilities first
  - For example:

$$\begin{aligned} P(X, Y|Z) &= \frac{P(Z|X, Y)P(X, Y)}{P(Z)} = \frac{P(Y, Z|X)P(X)}{P(Z)} \\ P(X|Y, Z) &= \frac{P(Y|X, Z)P(X, Z)}{P(Y, Z)} = \frac{P(Y|X, Z)P(X|Z)P(Z)}{P(Y|Z)P(Z)} = \frac{P(Y|X, Z)P(X|Z)}{P(Y|Z)} \end{aligned}$$

1.7.3 Expectations

- Take two random variables X and Y with joint PDF $p(x, y)$, the expectation of their sum is

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

1.7.4 Correlated Random Variables

- **Covariance**: A measure of the jointly variability of two random variables
  - Gives some sense of how much two values are linearly related to each other
  - For two jointly distributed random variables X and Y, the covariance is defined as:

$$\mathbb{C}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

  - Correlation is not causation
  - Two jointly distributed random variables X and Y is called:
    - positively correlated if $\mathbb{C}(X, Y) > 0$
    - negatively correlated if $\mathbb{C}(X, Y) < 0$
    - uncorrelated if $\mathbb{C}(X, Y) \approx 0$
- **Properties of the covariance**:
  - Let X be a random variable:

$$\mathbb{C}[X, X] = \mathbb{V}[X]$$

  - Let X be a random variable and take any constant λ:

$$\mathbb{C}[X, \lambda] = 0$$

  - Let X and Y be two random variables:

$$\mathbb{C}[Y, X] = \mathbb{C}[X, Y]$$

  - Let X and Y be two random variables and take any constants:

$$\mathbb{C}[\lambda X, \mu Y] = \lambda\mu\mathbb{C}[X, Y]$$

  - Let X and Y be two random variables and take any constants:

$$\mathbb{C}[X + \lambda, Y + \mu] = \mathbb{C}[X, Y]$$

  - Let X, Y and Z be three random variables:

$$\mathbb{C}[X, Y + Z] = \mathbb{C}[X, Y] + \mathbb{C}[X, Z]$$

  - Let X and Y be two random variable:

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\mathbb{C}[X, Y]$$

1.7.5 Independent Random Variables

- **Definition**: Two random variables X and Y are independent conditional on I, denoted as $X \perp Y|I$ if and only if conditioning on one does not tell anything about the other

$$p(x|y, I) = p(x|I)$$

  - When there is no ambiguity, I can be dropped
- **Properties**: Assume X and Y are two independent random variables, then:
  - $p(x, y) = p(x)p(y)$
  - $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
  - $\mathbb{C}[X, Y] = 0$
    - The reverse is not true. Uncorrelated variables do not have to be independent
  - $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$

1.8 Random Vectors

1.8.1 Definition

- A **list of random variables**, sometimes called **a multivariate random variable**
- Let $X_1, X_2, \cdots, X_N$ be N random variables, then $\mathbf{X} = [X_1, X_2, \cdots, X_N]^T$ is called a **random vector**
- The **joint probability density function** is denoted as

$$p(\mathbf{x}) = p(x_1, \cdots, x_N)$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

  - Properties:
    - $\forall \mathbf{x}, p(\mathbf{x}) \geq 0$
    - $\int p(\mathbf{x}) dx_1 dx_2 \cdots dx_N = 1$
    - $p(x_i) = \int p(\mathbf{x}) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_N$

1.8.2 Expectation and Covariance Matrix

- **Expectation** of a random vector is the vector of expectation of *each component*

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_N] \end{bmatrix}$$

- **Covariance Matrix**:
  - Let $\mathbf{X}$ be a N-dimensional random vector, $\mathbf{Y}$ be an M-dimensional random vector, then the covariance of $\mathbf{X}$ and $\mathbf{Y}$ is the $N \times M$ matrix consisting of all covariances between the components of $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbb{C}[\mathbf{X}, \mathbf{Y}] = [\mathbb{C}[X_i, Y_j]]_{i,j}$$

  - We can easily show that:

$$\mathbb{C}[\mathbf{X}, \mathbf{Y}] = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}]) (\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T \right]$$

  - Self-covariance of N-dimensional random vector $\mathbf{X}$ is a $N \times N$ matrix:

$$\mathbb{C}[\mathbf{X}, \mathbf{X}] = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}]) (\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \right]$$

1.8.3 Multivariate Normal - Diagonal Covariance Case

- Take the special case of N **independent** random variables $X_1, ..., X_N$ each distributed according to a normal with known mean and variance

$$X_i = \mathcal{N}(\mu_i, \sigma_i^2)$$

Then $\mathbf{X} = [X_1, \cdots, X_N]^T$ is called a **multivariate normal**

- Expectation:
  - $\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_N] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix} = \boldsymbol{\mu}$
- Covariance Matrix:
  - Notice that the random variables are independent to each other, which means

$$\mathbb{C}[X_i, X_j] = \begin{cases} 0 & i \neq j \\ \sigma_i^2 & i = j \end{cases}$$

  - Therefore, the covariance matrix of these matrix is

$$\mathbb{C}[\mathbf{X}, \mathbf{X}] = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{bmatrix} = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2) = \boldsymbol{\Sigma}$$

- For a multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the notation is

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{bmatrix}$$

- Joint probability density function:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

- Connection to the standard normal
  - Let $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I})$ be a collection of independent standard normal random variables

$$Z_i \sim \mathcal{N}(0, 1)$$

  Define the random vector

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{Z}$$

  Then

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

1.8.4 Multivariate Normal - Full Covariance Case

- General case of N random variables $X_1, ..., X_N$ each distributed according to a normal with known mean and variance

$$X_i = \mathcal{N}(\mu_i, \sigma_i^2)$$

  The random vector $\mathbf{X} = [X_1, \cdots, X_N]^T$ is denoted by

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- **Restriction of the covariance matrix**
  - In this general case, $\mathbf{\Sigma}$ is no longer a diagonal matrix
  - The covariance matrix has to be **positive definite**
    - For any $\mathbf{v} \neq 0$, $\mathbf{v}^T\mathbf{\Sigma}\mathbf{v} \geq 0$
    - This is so that $p(\mathbf{x})$ has a global maximum
    - Equivalently, $\mathbf{\Sigma}$ must have positive eigenvalues
- Connection to the standard normal
  - Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a collection of independent standard normal random variables

    $$Z_i \sim \mathcal{N}(0, 1)$$

    Define the random vector

    $$\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$$

    Then

    $$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^T)$$

- **Marginalization**
  - Let $\mathbf{X}$ be a random vector made out of two random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$:

    $$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

    Assume that

    $$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

    Decompose mean and covariance in blocks:

    $$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \ \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{1,1} & \mathbf{\Sigma}_{1,2} \\ \mathbf{\Sigma}_{1,2}^T & \mathbf{\Sigma}_{2,2} \end{bmatrix}$$

    To find the probability density of $\mathbf{X}_1$, marginalize

    $$p(\mathbf{x}_1) = \int p(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}, \mathbf{\Sigma}_{1,1})$$

- **Conditioning**
  - Let $\mathbf{X}$ be a random vector made out of two random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$:

    $$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

    Assume that

    $$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}), \ \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \ \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{1,1} & \mathbf{\Sigma}_{1,2} \\ \mathbf{\Sigma}_{1,2}^T & \mathbf{\Sigma}_{2,2} \end{bmatrix}$$

    The conditional PDF is

    $$p(\mathbf{x}_1 | \mathbf{x}_2) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_2)} \propto p(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_{1|\mathbf{x}_2}, \mathbf{\Sigma}_{1,1|\mathbf{x}_2})$$

    Where

    $$\boldsymbol{\mu}_{1|\mathbf{x}_2} = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_{2,2}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \ \mathbf{\Sigma}_{1,1|\mathbf{x}_2} = \mathbf{\Sigma}_{1,1} - \mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_{2,2}^{-1}\mathbf{\Sigma}_{1,2}^T$$

2 Sampling and Monte Carlo Method

2.1 Basic Sampling

2.1.1 Pseudo-random number generators

- Computers are deterministic machines and they cannot generate completely random numbers
- Pseudo-random number generator generates deterministic sequences of numbers that look random

2.1.2 Sample the uniform

- Given
  - PRNG's generate random integers from 0 to m
- Procedure
  - Sample a random integer d
  - Set $x = \frac{d}{m}$
  - Calculate the **empirical CDF** ($\hat{F}_N(x)$) with the ideal CDF

$$\hat{F}_N(x) = \frac{\text{number of elements in sample} \leq \text{x}}{N}$$

2.1.3 Sample the categorical

- To sample from a categorical distribution

$$X \sim \text{Categorical}(p_1, p_2, \cdots, p_{k-1}, 1 - \mathbf{1}^T\mathbf{p})$$

  - Draw a uniform number $u \sim U([0, 1])$
  - Find j such that

$$\sum_{k=0}^{j-1} p_k \leq u < \sum_{k=0}^{j} p_k$$

  - j-th state is the sample

2.1.4 Inverse Sampling

- Let X be an arbitrary univariate continuous random variable with CDF $F(x)$
  - Draw a uniform number $u \sim U([0, 1])$
  - Set

$$x = F^{-1}(u)$$

    and get the sample
- Proof
  - Let U be a uniform random variable

$$U \sim U([0, 1])$$

    For any CDF $F(x)$ define the random variable

$$X = F^{-1}(U)$$

The the CDF OF X is

$$\begin{aligned} P(X \leq x) &= P(F^{-1}(U) \leq x) \\ &= P(F(F^{-1}(U)) \leq F(x)) \\ &= P(U \leq F(x)) \\ &= F_U(F(x)) \\ &= F(x) \end{aligned}$$

2.2 The Monte Carlo Method for estimating expectations and variances

2.2.1 The uncertainty propagation problem

- Given a random variable $X \sim p(x)$ and a function $g(x)$, to quantify the uncertainty about the model output $Y = g(X)$
  - $\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int g(x)p(x)$
  - $\mathbb{V}[Y] = \int (g(x) - \mathbb{E}[Y])^2 p(x)dx = \mathbb{E}[[g(x)]^2] - (\mathbb{E}[g(x)])^2$
  - $P(Y \geq y) = \int \mathbf{1}_{[y,\infty]}(g(x))p(x)dx = \mathbb{E}\left[\mathbf{1}_{[u,\infty]}(g(x))\right]$
- All the statistics are essentially expectations of functions of X

2.2.2 Curse of Dimensionality

The number of samples needed to estimate an arbitrary function with a given level of accuracy grows **exponentially** with respect to the number of input variables (i.e. dimensionality) of the function

For example, take a d-dimensional uniform $X \sim U([0,1]^d)$ and a function $g(x)$, to estimate

$$\mathbb{E}[g(X)] = \int g(x)p(x)dx$$

use n equidistance points per dimension, which results in n^d boxes each with volume n^{-d}. Evaluate the integral with

$$\mathbb{E}[g(x)] \approx n^{-d} \sum_{j=1}^{n^d} g(x_j)$$

and assume it takes a millisecond to evaluate the function $g(x)$ and n = 10, it takes

- 0.1 seconds when $d = 1$
- 1 second when $d = 2$
- 100 seconds when $d = 3$
- 1000 seconds when $d = 6$
- 115 days when $d = 10$
- 3.17 billion years when $d = 20$

Blessing of dimensionality: Surprisingly and despite the expected "curse of dimensionality" difficulties, common-sense heuristics based on the most straightforward methods "can yield results which are almost surely optimal" for high-dimensional problems

2.2.3 The law of large numbers

- Take an infinite series of independent random variables $X_1, X_2, \cdots$ with the same distribution (it doesn't matter what distribution), the sample average

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \text{ a.s.}$$

  where a.s. means "almost surely" and $\mu = \mathbb{E}[X_i]$ (as $N \to \infty$)
- Take a random variable $X \sim p(x)$ and some function $g(x)$, to estimate the **expectation**:

$$I = \mathbb{E}[g(X)] = \int g(x)p(x)dx$$

  Make independent identical copies of X:

$$X_1, X_2, \cdots \sim p(x)$$

  Consider the independent identical distribution:

$$Y_1 = g(X_1), Y_2 = g(X_2), \cdots$$

  Then by the strong law of large numbers:

$$I_N = \frac{Y_1 + Y_2 + \cdots + Y_N}{N} \to \mathbb{E}[Y_i] = \mathbb{E}[g(X_i)] = I \text{ a.s.}$$

- Take a random variable $X \sim p(x)$ and some function $g(x)$, to estimate the **variance**:

$$V = \mathbb{V}[g(x)] = \mathbb{E}\left[(g(X) - \mathbb{E}[g(X)])^2\right] = \mathbb{E}\left[(g(X) - I)^2\right] = \mathbb{E}\left[g(x)^2\right] - I^2$$

  Take independent identical copies of X:

$$X_1, X_2, \cdots \sim p(x)$$

  Estimate the mean using a sample average:

$$\bar{I}_N = \frac{1}{N}\sum_{i=1}^{N} g(X_i)$$

  Estimate the variance by:

$$\bar{V}_N = \frac{1}{N}\sum_{i=1}^{N} g^2(X_i) - \bar{I}_N^2$$

2.3 Monte Carlo Estimates of various statistics

2.3.1 Estimating the cumulative distribution function

Take a random variable $X \sim p(x)$ and some function $g(x)$, consider the derived random variable $Y = g(X)$, we would like to estimate the **cumulative distribution function**:

$$F(y) = P(Y \leq y) = P(g(X) \leq y)$$

Consider the indicator function of a set A:

$$1_A(y) = \begin{cases} 1 & y \text{ in } A \\ 0 & \text{otherwise} \end{cases}$$

Use it, we can write $F(y)$ as an expectation:

$$F(y) = \mathbb{E}\left[1_{(-\infty,y]}\left(g(X)\right)\right]$$

Take $X_1, X_2, \cdots$ independent identical copies of X, estimate the CDF using a sample average:

$$\bar{F}_N(y) = \frac{1}{N}\sum_{i=1}^{N} 1_{(-\infty,y]}\left(g(X_i)\right) = \frac{\text{number of } g(X_i) \leq y}{N}$$

This estimate is called the **empirical CDF**

2.3.2 Estimating the probability density function via histograms

Take a random variable $X \sim p(x)$ and some function $g(x)$, consider the derived random variable $Y = g(X)$, we would like to estimate the **probability density function** $p(y)$ of Y

Take M small bins $[b_0, b_1], \cdots, [b_{M-1}, b_M]$ in the y space and approximate $p(y)$ with a constant inside each bin:

$$\bar{p}_M(y) = \sum_{j=1}^{M} c_j 1_{[b_{j-1},b_j]}(y)$$

The constants c_j are:

$$c_j = P(b_{j-1} \leq Y \leq b_j) = F(b_j) - F(b_j - 1)$$

So, we can approximate the constants c_j with the empirical CDF:

$$\bar{c}_{j,N} = \bar{F}_N(b_j) - \bar{F}_N(b_{j-1}) = \frac{\text{number of samples that fall in bin } [b_{j-1}, b_j]}{N} \to c_j \text{ a.s.}$$

Putting everything together the approximation becomes:

$$\bar{p}_{M,N}(y) = \sum_{j=1}^{M} \bar{c}_{j,N} 1_{[b_{j-1},b_j]}(y)$$

2.3.3 Estimating Predictive Quantiles

- **Predictive Quantiles**:
  - Take a random variable $X \sim P(X)$ and some function $g(x)$, let $Y = g(X)$ and $F(y)$ be the CDF of Y, then the q-predictive quantile of Y is the value μ_q such that:

    $$P(Y \leq \mu_q) = F_Y(\mu_q) = \frac{q}{100}$$

  - $F(\mu_{50}) = 50\%$ is called *median*
- To find the q-predictive quantile of Y:

  $$F_Y(\mu_q) = \frac{q}{100}$$

  Turn it into a root finding problem:

  $$F(\mu_q) - \frac{q}{100} = 0$$

  Take $X_1, X_2, \cdots$ independent identical copies of X, fins the sampling estimate of the CDF of $Y = g(X)$ denoted by

  $$\bar{F}_N(y)$$

  Then numerically solve the root finding problem and obtain the quantile $\bar{\mu}_{q,N}$

2.4 Quantifying uncertainties in Monte Carlo Estimates

2.4.1 The central limit theorem

- *The central limit theorem* states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed.
- Take $X_1, X_2, \cdots$ to be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2, consider their average

$$S_N = \frac{X_1 + \cdots + X_N}{N}$$

The Central Limit Theorem States that:

$$S_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) \text{ for large } N$$

2.4.2 Epistemic uncertainty of Monte Carlo Estimates

Use sampling to estimate:

$$I = \mathbb{E}[g(X)]$$

Take i.i.d. random variables:

$$X_1, X_2, \cdots$$

Consider the also i.i.d. random variables:

$$Y_1 = g(X_1), Y_2 = g(X_2), \cdots$$

Use the *law of large numbers*:

$$\bar{I}_N = \frac{g(X_1) + \cdots + g(X_N)}{N} = \frac{Y_1 + \cdots + Y_N}{N} \to I, \text{ a.s.}$$

Note that $Y_i = g(X_i)$ are i.i.d. random variables with mean:

$$\mathbb{E}[Y_i] = \mathbb{E}[g(X_i)] = I$$

Assume their variance is finite:

$$\mathbb{V}[Y_i] = \sigma^2 < +\infty$$

Then the *central limit theorem* holds and it gives:

$$\bar{I}_N = \frac{Y_1 + \cdots + Y_N}{N} \sim \mathcal{N}\left(I, \frac{\sigma^2}{N}\right) \text{ for large } N$$

Rewrite as:

$$\bar{I}_N = I + \frac{\sigma}{\sqrt{N}} \cdot Z, \ Z \sim \mathcal{N}(0,1) \Rightarrow I = \bar{I}_N - \frac{\sigma}{\sqrt{N}} \cdot Z$$

That is to say:

$$I \sim \mathcal{N}\left(\bar{I}_N, \frac{\sigma^2}{N}\right)$$

Similarly, estimate the variance by:

$$\bar{V}_N = \bar{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} g^2(X_i) - \bar{I}_N^2$$

And end up with

$$I \sim \mathcal{N}\left(\bar{I}_N, \frac{\bar{\sigma}_N^2}{N}\right) \text{ for large } N$$

3 Basics of Information Theory

3.1 Information Entropy

- In information theory, we would like to quantify information in a way that formalizes the intuition that *learning that an unlikely event has occurred is more informative than learning that a likely event has occurred*:
  - Likely events should have low information content, and in the extreme case, events that a guaranteed to happen should have no information content whatsoever
  - Less likely events should have higher information content
  - Independent events should have additive information
    - e.g. Finding out that a tossed coin has come up as head twice should convey twice as much information as finding out that a tossed coin has come up as heads once
- To satisfy those properties, **self-information** of an event $X = x$ is defined to be

$$I(x) = -\log P(X = x)$$

  As the base of logarithm is e, $I(x)$ is written in units of **nats**. One nat is the amount of information gained by observing an event of probability $\frac{1}{e}$. When the base is 2, then the unit is called **bits** or **shannons**.
- Self-information deals only with a single outcome. In 1948, **Claude Shannon** came up with a function called **Information Entropy** quantifying the amount of uncertainty in an entire probability distribution:

$$\mathbb{H}(X) = \mathbb{E}_{X\sim P(X)}[I(x)] = -\mathbb{E}_{X\sim P(X)}[\log p(x)]$$

  Where X is a **discrete random variable**. In other words, the Information Entropy of a distribution is the expected amount of information in an event drawn from that distribution
  - Distributions that are nearly deterministic (where the outcome is nearly certain) have low entropy.
  - Distributions that are closer to uniform have high entropy.
- When X is a **continuous random variable**, the naive extension of information entropy is called **differential entropy** and lose some abilities it has in discrete form. Usually we do not consider using Shannon Entropy directly for continuous random variables as it actually fails to quantify "uncertainty" over the whole distribution.
  - Is no longer invariant under change of variables
  - Can be negative
  - ...

3.2 Kullback-Leibler divergence

- Given two separate probability distributions $P(X)$ and $Q(X)$ over the same random variable X, we can measure how different there two distributions are using the *Kullback-Leibler divergence*:

$$D_{KL}(P\|Q) = \mathbb{E}_{X\sim P(X)}\left[\log\frac{P(x)}{Q(x)}\right] = \mathbb{E}_{X\sim P(X)}[\log P(x) - \log Q(x)]$$

- Can be viewed as the information loss when replacing $Q(X)$ with $P(X)$

- Formally speaking, in the case of **discrete variables**, it is the extra amount of information needed to send a message containing symbols drawn from probability distribution $P(X)$ when we use a code that was designed to minimize the length of messages drawn from probability distribution $Q(X)$
- **Important Properties**:
  - Being non-negative $D_{KL}(P\|Q) \geq 0$
  - $D_{KL}(P\|Q) = 0$ is and only if P and Q are the same distribution in the case of discrete variables, or equal "almost every where" in the case of continuous variables
  - Although often used to measuring some sort of distance between distances, it is not a true distance

$$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$$

3.3 Cross-Entropy

- Given two separate probability distributions $P(X)$ and $Q(X)$ over the same random variable X, we can measure the average number of bits (or nats, depend on the base of logarithm) to identify an event drawn from the set if a coding scheme used for an estimated probability distribution $Q(X)$ rather than the true distribution $P(X)$:

$$\mathbb{H}(P, Q) = \mathbb{H}(P) + D_{KL}(P\|Q) = -\mathbb{E}_{X\sim P(X)}\log Q(x)$$

- Notice that $Q(X)$ does not participate in the $\mathbb{H}(P)$ term, minimizing the cross-entropy with respect to Q (keep P unchanged) is **equivalent** to minimizing the KL divergence. Obviously, calculating the cross-entropy is much more convenient when the equivalence exists.
- **Important Properties**:
  - Like KL divergence, cross entropy is not symmetric

$$\mathbb{H}(P, Q) \neq \mathbb{H}(Q, P)$$

  - Cross-entropy to a distribution itself is its information entropy

$$\mathbb{H}(P, P) = \mathbb{H}(P)$$

4 Bayes Inference

4.1 Formal descriptions of Bayesian inference

4.1.1 Notations

- X: A random variable or vector.
- θ: The parameter of the random variable's distribution $X \sim P(X|\theta)$, could be a vector.
- α: The *hyperparameters* of the parameter distribution $\theta \sim P(\theta|\alpha)$, could be a vector.
- $\mathbf{x}$: Sample, a set of observations of the random variable $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$
- $\tilde{x}$: A new observation whose distributions is to be predicted

4.1.2 Bayesian inference

- **Prior Distribution**:
  - A *prior probability distribution*, often simply called the *prior*, of an uncertainty is the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account.
  - In this case, it is the *distribution of the parameter* θ before any observations is made:

$$\textbf{Prior} = p(\theta|\alpha)$$

- **Likelihood**:
  - The *distribution of the observations* conditional on its parameters.

$$\textbf{Likelihood} = \mathrm{L}(\theta|\mathbf{x}) = p(\mathbf{x}|\theta)$$

  - It measures the **goodness of fit** of a statistical model $X \sim P(X|\theta)$ to a sample of data $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ for given *values* of the unknown parameters $\theta \sim P(\theta|\alpha)$
  - When the measurements are *independent*:

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

- **Marginal likelihood**
  - The distribution of the observations marginalized over the parameters

$$p(\mathbf{x}|\alpha) = \int p(\mathbf{x}|\theta)p(\theta|\alpha)d\theta$$

  - Sometimes called *evidence*
- **Posterior Probability**
  - The *distribution of the parameters* after taking into the observations, which is determined by *Bayes' rule*

$$\begin{aligned} p(\theta|\mathbf{x}, \alpha) &= \frac{p(\theta, \mathbf{x}, \alpha)}{p(\mathbf{x}, \alpha)} \\ &= \frac{p(\mathbf{x}|\theta, \alpha)p(\theta, \alpha)}{p(\mathbf{x}|\alpha)} \\ &= \frac{p(\mathbf{x}|\theta, \alpha)p(\theta|\alpha)}{p(\mathbf{x}|\alpha)} \end{aligned}$$

  That is to say

$$\textbf{Posterior} = p(\theta|\mathbf{x}, \alpha) = \frac{p(\mathbf{x}|\theta, \alpha)p(\theta|\alpha)}{p(\mathbf{x}|\alpha)} \propto p(\mathbf{x}|\theta, \alpha)p(\theta|\alpha)$$

– This is expressed in words as *"posterior is proportional to likelihood times prior"*

$$\textbf{Posterior} \propto \textbf{Likelihood} \times \textbf{Prior}$$

– In practice, for almost all complex Bayesian models used in machine learning, the posterior distribution $p(\theta|\mathbf{x}, \alpha)$ is **not** obtained in a closed form distribution, mainly because the parameter space for θ can be very high, or the Bayesian model retains certain hierarchical structure formulated from the observations $\mathbf{x}$ and parameter θ. In such situations, approximation techniques are needed

- **Credible Intervals**
  – An interval within which an unobserved parameter θ falls with a particular probability and it is in the domain of a posterior probability distribution
    * For example, a 95% credible interval is an interval $[l,u] $where

$$\begin{aligned} p(l \leq \theta \leq u|\mathbf{x}) &= 0.95 \\ p(\theta \leq l|\mathbf{x}) &= 0.025 \\ p(\theta \leq u|\mathbf{x}) &= 0.975 \end{aligned}$$

4.1.3 Bayesian Prediction

- **Posterior Predictive Distribution**:
  – The distribution of a new observation marginalized over the posterior:

$$p(\tilde{x}|\mathbf{x}, \alpha) = \int p(\tilde{x}|\theta)p(\theta|\mathbf{x}, \alpha)d\theta$$

- **Prior Predictive Distribution**:
  – The distribution of a new observation marginalized over the prior:

$$p(\tilde{x}|\alpha) = \int p(\tilde{x}|\theta)p(\theta|\alpha)d\theta$$

- Remark:
  – In Bayesian Prediction, instead of a fixed value as a prediction, a **distribution** over possible values is returned, in which way the entire posterior distribution of the parameters is used.
  – By comparison, prediction in frequentist statistics often tries to find an optimum point estimate of the parameters (e.g. by maximum likelihood or maximum a posteriori estimation(MAP)) and then plugging this estimate into the formula for the distribution of the prediction. This leads to the disadvantage that it does not account for any uncertainty in the value of the parameter, and hence will underestimate the variance of the prediction distribution.

4.1.4 Decision-Making Problem

- To decide the best suitable parameter θ given a set of observations $\mathbf{x}$, the cost of making a mistake needs to be quantified and then a decision minimizes the cost should be made
- Denote the **loss** when we guess $\hat{\theta}$ and the true value is θ as $l(\hat{\theta}, \theta)$, then

$$\mathbb{E}\left[l(\hat{\theta}, \theta)|\mathbf{x}\right] = \int l(\hat{\theta}, \theta)p(\theta|\mathbf{x})d\theta$$

The best decision θ^* should be

$$\theta^* = \underset{\hat{\theta}}{\text{argmin}} \mathbb{E}\left[l(\hat{\theta}, \theta)|\mathbf{x}\right]$$

- The $0-1$ loss:

$$l_{01}(\hat{\theta}, \theta) = \begin{cases} 0 & \hat{\theta} = \theta \\ 1 & \hat{\theta} \neq \theta \end{cases}$$

  Then

$$\theta^* = \underset{\theta}{\text{argmax}} p(\theta|\mathbf{x})$$

- The square loss (L^2 norm):

$$l_2(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

  Then

$$\theta^* = \mathbb{E}[\theta|\mathbf{x}] = \int \theta p(\theta|\mathbf{x}) d\theta$$

- The Absolute loss (L^1 norm):

$$l_2(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$$

  Then

$$\theta^* = \text{median of the posterior}, p(\theta \leq \theta^*|\mathbf{x}) = 0.5$$

4.1.5 Posterior Predictive checking

- **Replicated data**
  - Assume that we have built model using data $\mathbf{x}$, then the replicated data $\mathbf{x}^{rep}$ is given by

$$\theta_s|\mathbf{x} \sim p(\theta|\mathbf{x})$$
$$\mathbf{x}^{rep}|\theta = \theta_s \sim p(\mathbf{x}^{rep}|\theta = \theta_s)$$

  - The idea is to sample $\mathbf{x}^{rep}$ and compare their characteristics to the observed data $\mathbf{x}$
- **Test quantiles**
  - We use test quantiles to characterize the discrepancy between the model and the data.
  - Helps to zoom into the characteristics that are of particular interest
  - Depends on *what important characteristics of the data the model is supposed to be capturing*
  - Mathematically, a quantile is a scalar function of the data and the parameters

$$T(\mathbf{x}, \theta)$$

  - The idea is to compare the histogram of the test quantity over replicated data (of $p(T(\mathbf{x}^{rep}, \theta)|\mathbf{x})$) with the observe value $T(\mathbf{x}, \theta)$
- **Bayesian p-values**
  - Check the model by evaluating the probability that the replicated data give a test quantity that is more extreme than the observed value. The probability is known as the posterior (or Bayesian) p-value and it is defined by:

$$p_B = P(T(x^{\text{rep}}_{1:n}, \theta) > T(x_{1:n}, \theta)|x_{1:n})$$

  - **Not** the probability that the model is correct
  - Values close to 0 or 1 indicate some issue with the model with regard to that particular quantity
  - Values close to 0.5 indicate no issue, but they do not mean that the model is correct

4.2 Selecting Prior Information

- Three widely accepted ways to select prior information following a systematic theoretical framework:
  - The principle of insufficient reason
  - The principle of maximum entropy
  - The principle of transformation groups

4.2.1 The principle of insufficient reason

- The *principle of insufficient reason*, also called *principle of indifference* is a rule for assigning epistemic probabilities. The principle of insufficient reason states that in the absence of any relevant evidence, agents should distribute their *degrees of belief* equally among all the possible outcomes under consideration
- Let X be a discrete random variable taking N different values $x_1, \cdots, x_N$, if that is all we know, then the principle of insufficient reason states that we should assign:

$$P(X = x_1) = P(X = x_2) = \cdots = P(X = x_N) = \frac{1}{N}$$

- Generalizations of the principle of insufficient reason
  - Principle of transformation groups (advanced)
  - Principle of maximum entropy

4.2.2 The principle of maximum entropy

- The uniform distribution required by the *principle of insufficient reason* is the most "uncertain" distribution. Mention that the *information entropy* is a measurement of uncertainty in an entire distribution, the idea is to assign the distribution that has the maximum uncertainty while being consistent with the data
- For **discrete variables**:
  - Definition of **testable information**:

    $$\mathbb{E}[f(x)] = F$$

    Where $f(x)$ is a known function and F is a known value
  - Let X be a discrete random variable and $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}$ be a set of observations while some *testable information* is available, the principle of maximum entropy states that we should assign to X the probability distribution that maximizes the entropy subject to the constraints imposed by the testable information
  - The mathematical representation for a discrete random variable with N different values is

    $$\underset{P(X)}{\operatorname{argmax}} \mathbb{H}(P(X)) = \underset{P(X)}{\operatorname{argmax}} \left\{ -\sum_{i}^{N} p(x_i) \log p(x_i) \right\}$$

    Subject to

    $$\mathbb{E}\left[f_k(x)\right] = F_k, \ k = 1, 2, \cdots, K$$

    Where K is the number of constraints and

    $$\sum_{i=1}^{N} p(x_i) = 1$$

  - The general solution to maximum entropy problem is

$$P(X = x_i) = \frac{1}{z}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(x_i)\right\}$$

    Where

$$z = \sum_{i=1}^{N}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(x_i)\right\}$$

    And λ_k needs to be determined according to

$$F_k = \frac{\partial z}{\partial \lambda_k}$$

  - Examples:
    * X takes N different values $\Rightarrow P(X = x_i) = \frac{1}{N}$
    * X takes two values 0 and 1, $\mathbb{E}[X] = \theta \Rightarrow X \sim \text{Bernoulli}(\theta)$
    * X takes values $0, 1, 2, \cdots, N$, $\mathbb{E}[X] = \mu$ and X is the number of successful trials in N potentially correlated sequential experiments $\Rightarrow X \sim \text{B}(N, \frac{\mu}{N})$ (Binomial Distribution)
    * X takes values $0, 1, 2, \cdots$, $\mathbb{E}[X] = \mu$ and X is the number of successful trials in an infinite number of potentially correlated sequential experiments $\Rightarrow X \sim \text{Poisson}(\mu)$

- For **continuous variables**:
  - Notice that naive extension of information entropy to continuous distributions is no longer invariant under change to variables. The Shannon Entropy is only defined for discrete spaces and only show the abilities to quantify "uncertainty" in discrete spaces
  - *Edwin Jaynes* proposed the following formula

$$\mathbb{H}_c = -\int p(x)\log\frac{p(x)}{q(x)}dx$$

    where $q(x)$, called the "invariant measure", is proportional to the limiting density of discrete points.
  - Then the mathematical representation for a continuous random variable is

$$\underset{P(X)}{\text{argmax}}\mathbb{H}_c(P(X)) = \underset{P(X)}{\text{argmax}} -\int p(x)\log\frac{p(x)}{q(x)}dx$$

    Subject to

$$\mathbb{E}\left[f_k(x)\right] = F_k,\ k = 1, 2, \cdots, K$$

    Where K is the number of constraints and

$$\sum_{i=1}^{N} p(x_i) = 1$$

  - The general solution to maximum entropy problem is

$$P(X = x_i) = \frac{q(x)}{z}\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(x_i)\right\}$$

    Where

$$z = \int q(x)\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(x_i)\right\}dx$$

    And λ_k needs to be determined according to

$$F_k = \frac{\partial z}{\partial \lambda_k}$$

  - Mathematical theory for finding the maximal uncertainty density $q(x)$
    - Principle of transformation groups
    - Theory of Haar measures
  - Examples:
    - X takes values in $[a, b]$,$q(x) = 1 \Rightarrow X \sim U([a, b])$
    - X takes values in $\mathbb{R}$, $q(x) = 1$, $\mathbb{E}[X] = \mu$, $\mathbb{V}[X] = \sigma^2 \Rightarrow X \sim \mathcal{N}(\mu, \sigma^2)$
- For an almost list of a commonly used maximum entropy distributions, see Maximum entropy probability distribution entry of wikipedia