Week #4: Variance

March 10, 2025

Monte Carlo Methods

Monte Carlo methods are a versatile set of computational techniques that employ random sampling to approximate solutions for problems that are difficult or impossible to solve analytically. In this document, we delve into the mathematical framework behind these methods in a single, continuous exposition.

Consider the problem of estimating an integral of a function $f(x)$ over an interval $[a, b]$. The target integral is given by

$$I = \int_a^b f(x)\, dx.$$

A straightforward Monte Carlo approach is to sample $x_1, x_2, \ldots, x_N$ independently from a uniform distribution on $[a, b]$. The probability density function (pdf) for the uniform distribution is:

$$p(x) = \frac{1}{b-a}, \quad \text{for } x \in [a, b].$$

The expected value of $f(x)$ under this distribution is

$$\mathbb{E}[f(x)] = \int_a^b f(x)\, p(x)\, dx = \frac{1}{b-a} \int_a^b f(x)\, dx = \frac{I}{b-a}.$$

Thus, the integral can be expressed as

$$I = (b-a)\, \mathbb{E}[f(x)].$$

In practice, we estimate the expected value by computing the sample mean:

$$\hat{I} = \frac{b-a}{N} \sum_{i=1}^{N} f(x_i).$$

The foundation of this method is the *law of large numbers* (LLN), which states that for independent and identically distributed (i.i.d.) random variables $x_1, x_2, \ldots, x_N$ drawn from the uniform distribution, the sample average converges almost surely to the true expected value:

$$\frac{1}{N} \sum_{i=1}^{N} f(x_i) \xrightarrow{\text{a.s.}} \mathbb{E}[f(x)] \quad \text{as } N \to \infty.$$

A common illustrative example uses indicator functions. Let $\mathbf{1}_A(x)$ denote the indicator function of an event A, defined as:

$$\mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Then the probability of the event A is given by:

$$P(A) = \int \mathbf{1}_A(x)\, p(x)\, dx.$$

By drawing samples $x_1, x_2, \ldots, x_N$ from $p(x)$ and computing the sample mean

$$\hat{P}(A) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_A(x_i),$$

the LLN assures us that

$$\hat{P}(A) \xrightarrow{\text{a.s.}} P(A) \quad \text{as } N \to \infty.$$

This simple idea of replacing integrals with sample averages forms the backbone of Monte Carlo integration.

In many practical applications, sampling uniformly is inefficient, especially in high-dimensional spaces. Instead, one often resorts to *importance sampling*. In this approach, one samples from a more convenient or efficient density $p(x)$ and rewrites the integral as:

$$I = \int_D f(x)\, dx = \int_D \frac{f(x)}{p(x)}\, p(x)\, dx = \mathbb{E}_p \left[\frac{f(x)}{p(x)} \right].$$

The corresponding Monte Carlo estimator is

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i)}{p(x_i)},$$

where the x_i are drawn according to $p(x)$.

Definition. Variance

For a random variable X with expected value $\mu = \mathbb{E}[X]$, the *variance* of X is defined as

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

Error analysis in Monte Carlo methods is critical. For a Monte Carlo estimator $\hat{I}$, the variance is expressed as:

$$\operatorname{Var}(\hat{I}) = \frac{1}{N} \operatorname{Var}\left(\frac{f(x)}{p(x)} \right),$$

which implies that the standard error decreases as $1/\sqrt{N}$. Under suitable conditions, the Central Limit Theorem guarantees that for large N,

$$\sqrt{N}\,(\hat{I} - I) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \operatorname{Var}\left(\frac{f(x)}{p(x)} \right)$. This result permits the construction of confidence intervals for the estimated value.

Several *variance reduction techniques* are commonly employed to improve the efficiency of Monte Carlo estimators:

- **Importance Sampling:** As discussed, choosing a sampling density $p(x)$ that closely resembles $f(x)$ can reduce the variance.

- **Stratified Sampling:** This technique partitions the domain into non-overlapping subdomains (strata) and samples from each, ensuring better coverage of the entire domain.
- **Control Variates:** If a function $g(x)$ with known expected value μ_g is correlated with $f(x)$, one can adjust the estimator:
$$\hat{I}_{\text{cv}} = \hat{I} + c(\mu_g - \hat{g}),$$
where c is chosen optimally and $\hat{g}$ is the sample mean of $g(x)$.
- **Antithetic Variates:** This method involves generating pairs of negatively correlated samples (e.g., x and $1 - x$ when sampling uniformly) to cancel out variance.

Rejection Sampling

In many scenarios, direct sampling from the target distribution $f(x)$ is challenging, while a proposal distribution $q(x)$ is readily available. Suppose there exists a constant $c \geq 1$ such that

$$f(x) \leq c\, q(x) \quad \text{for all } x.$$

Then, the rejection sampling algorithm can be used to generate samples from $f(x)$.

Theorem 1. Correctness of Rejection Sampling

Let $f(x)$ be the target density and $q(x)$ be a proposal density, and suppose there exists a constant $c \geq 1$ such that

$$f(x) \leq c\, q(x) \quad \text{for all } x.$$

Define the indicator random variable

$$I(X, U) = \mathbf{1}\left\{ U \leq \frac{f(X)}{c\, q(X)} \right\},$$

where $X \sim q(x)$ and $U \sim \text{Uniform}(0, 1)$ are independent. Then, the conditional distribution of X given $I(X, U) = 1$ is exactly $f(x)$.

Proof. Since X and U are independent, their joint density is

$$h(x, u) = q(x), \quad \text{for } u \in [0, 1].$$

The candidate x is accepted when

$$U \leq \frac{f(x)}{c\, q(x)}.$$

Thus, the joint density of accepted samples is

$$h_{\text{acc}}(x, u) = q(x) \cdot \mathbf{1}\left\{ u \leq \frac{f(x)}{c\, q(x)} \right\}.$$

Integrating out u gives the marginal density for accepted x:

$$g(x) = \int_0^1 h_{\text{acc}}(x, u)\, du = q(x) \frac{f(x)}{c\, q(x)} = \frac{f(x)}{c}.$$

The overall probability of acceptance is

$$P(I=1) = \int g(x)\,dx = \frac{1}{c}\int f(x)\,dx = \frac{1}{c}.$$

Hence, the conditional density of x given acceptance is

$$f_{X|I=1}(x) = \frac{g(x)}{P(I=1)} = \frac{f(x)/c}{1/c} = f(x).$$

This completes the proof. □

The rejection sampling algorithm is described as follows:

1. Sample X from the proposal density $q(x)$.
2. Independently sample $U \sim \text{Uniform}(0,1)$.
3. Define the indicator random variable
$$I(X,U) = \mathbf{1}\left\{U \leq \frac{f(X)}{c\,q(X)}\right\}.$$
4. If $I(X,U) = 1$, accept X as a sample from $f(x)$; otherwise, reject X and repeat the process.

1 Dependence & Independence

Independence is a fundamental property in probability theory, ensuring that the realization of one random variable does not alter the distribution of another. Formally, two random variables X and Y are independent if their joint density function factorizes as

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

which implies that knowing $X = x$ does not influence the probability law of Y. This definition is purely mathematical, but its justification often stems from physical intuition. If two sources of randomness arise from non-interacting systems, their outcomes are expected to be independent. However, the assumption of independence is not always evident, and distinguishing between physical and statistical independence requires careful consideration.

To illustrate this, consider the case of two fair dice rolled separately. The outcome of the first die, X, does not influence the outcome of the second die, Y, suggesting physical independence. The mathematical consequence is that their joint probability mass function satisfies

$$p_{X,Y}(i,j) = p_X(i)p_Y(j), \quad \text{for } i,j \in \{1,2,3,4,5,6\}.$$

Since each die follows a uniform distribution, we obtain

$$p_X(i) = \frac{1}{6}, \quad p_Y(j) = \frac{1}{6},$$

which leads to

$$p_{X,Y}(i,j) = \frac{1}{36}.$$

Thus, the factorization holds, confirming independence.

However, consider a modified scenario where both dice are rolled inside a closed box, and it is only revealed whether their sum is even or odd. If the sum is even, the possible outcomes are pairs (i, j) such that $i + j$ is even, meaning that knowledge of one outcome restricts the possible values of the other. In this case,

$$p_{X,Y}(i, j) \neq p_X(i)p_Y(j),$$

and independence is lost. Here, physical dependence arises due to the constraint that only certain pairs are possible, even though each die was rolled separately.

Once we have a quantitative way of defining independence, we could come to the question on how can we measure independence in qualitative scenarios. Here is where we can rigurously define an *event* or *situation*.

Definition.

An *event* or *situation* A is a random universe U. The corresponding indicator function $\mathbf{1}_A(x)$ is defined as

$$\mathbf{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

If X is a random variable with probability density function $p(x)$, then the probability of A is given by

$$P(A) = \mathbb{E}\big[\mathbf{1}_A(X)\big] = \int \mathbf{1}_A(x)\, p(x)\, dx.$$

So we can now discuss what it means the independence of two situations. Let A and B be two situations, if X is a random variable describing the underlying experiment and $p(x)$ is its probability density function, then

$$P(A \cap B) = \mathbb{E}\big[\mathbf{1}_A(X)\, \mathbf{1}_B(X)\big]. \tag{1}$$

Now note the following property.

If X and Y are independent random variables, then the expectation of their product is the product of their expectations:

$$E[XY] = E[X] \cdot E[Y] \tag{2}$$

Proof. Let X and Y be independent random variables. By the definition of expectation and independence:

$$\begin{aligned} E[XY] &= \sum_x \sum_y xy \cdot p_{X,Y}(x, y) \\ &= \sum_x \sum_y xy \cdot p_X(x)p_Y(y) \quad \text{(since } X \text{ and } Y \text{ are independent)} \\ &= \left(\sum_x x \cdot p_X(x)\right) \cdot \left(\sum_y y \cdot p_Y(y)\right) \\ &= E[X] \cdot E[Y] \end{aligned}$$

This proves that the expectation of the product of independent random variables is the product of their expectations. □

Thus following (1) can express the independence of situations as,

$$A \text{ and } B \text{ are independent} \quad \Longleftrightarrow \quad P(A \cap B) \;=\; P(A)\,P(B).$$

We discussed how to determine the independence of random variables. But what if they are dependent? Often, information about one event can help us understand another. For example, will I wear a jacket if it rains tomorrow? Conditional probability allows us to update our knowledge based on new information, helping us make better predictions about related events.

Definition. Conditional Probability

Given two random variables X and Y with $p_Y(y) > 0$, the **conditional probability** of $X = x$ given $Y = y$ is defined as:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

Intuitively, the numerator $p_{X,Y}(x,y)$ “reduces the number of possibilities” by restricting attention to those outcomes that satisfy both $X = x$ and $Y = y$. The denominator $p_Y(y)$ then “scales the measurement of the universe” to this restricted set, ensuring that the conditional probabilities sum (or integrate) to 1 over all possible x. In other words, knowing $Y = y$ limits the relevant portion of the sample space, and we normalize by $p_Y(y)$ to account for the fact that we are now only looking within this smaller region.

A direct consequence of the definition of conditional probability is the Multiplication Rule. Let X and Y be two random variables. The joint probability $p_{X,Y}(x,y)$ can be expressed in terms of conditional probabilities as:

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

It provides a foundational link between joint and conditional probabilities, allowing for systematic computation of joint probabilities.

For many random variables $X_1, X_2, \ldots, X_n$, the joint probability can be expressed as:

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1, \ldots, x_{i-1})$$

This represents the product of the conditional probabilities of each random variable occurring given the occurrence of all previous random variables.

Definition. Marginal Probability

Given a joint probability distribution $p_{X,Y}(x,y)$, the *marginal probability* $p_X(x)$ of any outcome x for the random variable X is obtained by summing the joint probabilities over all possible outcomes y for Y. Mathematically, the marginal probability $p_X(x)$ is given by:

$$p_X(x) = \sum_{y \in R(Y)} p_{X,Y}(x,y)$$

where the sum is over all possible outcomes of the random variable Y.

The connection between marginal and conditional probabilities can be understood through the **law of total probability**. The marginal probability $p_X(x)$ can be expressed in terms of conditional probabilities as follows:

$$p_X(x) = \sum_{y \in R(Y)} p_{X|Y}(x|y) \cdot p_Y(y)$$

This relationship demonstrates that the marginal probability of an outcome for a random variable can be obtained by considering all the ways that outcome can occur, weighted by the probability of each of those ways.

Suppose you want to determine whether a fruit drawn from a bag is an apple. Initially, your belief may be based on the overall proportion of apples in the bag. However, after feeling the fruit and noting its round, smooth texture, you gather new evidence. Bayes' theorem provides a systematic way to update your prior belief with this evidence. By reversing conditional probabilities, it lets you compute the probability that the fruit is an apple given its observed characteristics.

Theorem 2. Bayes' Theorem

Let X and Y be two random variables with $p_Y(y) \neq 0$. Then, the posterior probability $p_{X|Y}(x|y)$ is given by

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)\, p_X(x)}{P_Y(y)},$$

Proof. By the multiplication rule, the joint probability can be expressed as

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)\, P_X(x).$$

Since joint probabilities are symmetric, we also have

$$p_{X,Y}(x,y) = P_{X|Y}(x|y)\, P_Y(y).$$

Equating these two expressions and solving for $P(X = x|Y = y)$ gives

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)\, p_X(x)}{p_Y(y)}.$$

□

Example.

Imagine there's a rare disease, and there's a test for it. The disease affects 1% of the population, and the test is 99% accurate. If you test positive, what's the chance you actually have the disease?

Using Bayes' theorem:

Let X be the random variable representing the presence ($X = 1$) of the disease. In case the disease is not present it assumes a value 0. Let Y be the random variable representing the test result. $Y = 1$ for a positive test, while if the test is negative it is equal to 0.

We want to find: $p_{X|Y}(1|Y = 1)$, namely we want to understand what it is the probability of the presence of the disease given a positive test.

Given:

- $p_X(1) = 0.01$ (1% of the population has the disease)
- $p_{Y|X}(Y = 1|X = 1) = 0.99$ (The test is 99% accurate)
- $p_Y(1)$ is the total positive testing probability.

$$p_{X|Y}(1|Y = 1) = \frac{p_{Y|X}(Y = 1|X = 1) \cdot p_X(1)}{p_Y(1)}$$

To find $p_Y(1)$, namely the probability of a positive test, we consider the law of total probability:

$$p_Y(1) = p_{Y|X}(1|X = 1) \cdot p_X(1) + p_{Y|X}(1|X = 0) \cdot p_X(0)$$

$$p_Y(1) = (0.99)(0.01) + (0.01)(0.99) = 0.0198$$

Plugging in the numbers:

$$p_{X|Y}(1|Y = 1) = \frac{(0.99)(0.01)}{0.0198} \approx 0.5$$

So, even with a 99% accurate test, if you test positive, there's only a 50% chance you actually have the disease!

Bayes' theorem is foundational for the fields of Bayesian statistics and machine learning. It provides a mechanism to update our beliefs in light of new evidence, making it central to numerous applications, from medical diagnostics to recommendation systems. The theorem reminds us of the importance of prior knowledge and illustrates how, in a world filled with data, we can use this data to make more informed decisions and predictions.

Example. Spam Email Classification

Consider an email filter that classifies emails as spam or not spam based on certain words. Let S be the random variable representing whether an email is spam ($S = 1$) or not ($S = 0$). Let W represent the presence of certain words in the email.

Suppose we have:

- $p_S(1) = 0.4$ (40% of emails are spam)
- $p_{W|S}(1|1) = 0.8$ (Probability of certain words given the email is spam)
- $p_{W|S}(1|0) = 0.1$ (Probability of certain words given the email is not spam)

We want to find the probability that an email is spam given that it contains these words, i.e., $p_{S|W}(1|1)$.

Using Bayes' theorem:

$$p_{S|W}(1|1) = \frac{p_{W|S}(1|1) \cdot p_S(1)}{p_W(1)}$$

To find $p_W(1)$:

$$p_W(1) = p_{W|S}(1|1) \cdot p_S(1) + p_{W|S}(1|0) \cdot p_S(0)$$

$$p_W(1) = (0.8)(0.4) + (0.1)(0.6) = 0.38$$

Therefore:

$$p_{S|W}(1|1) = \frac{(0.8)(0.4)}{0.38} \approx 0.842$$

So, given that the email contains certain words, there is an 84.2% chance that it is spam.

Before defining formally what a Markov chain is let's think in the following application in finance:

Example. Market Sentiment

In financial markets, sentiment is a key indicator of future price movements. A **bullish** market ($M = 1$) is characterized by rising prices, optimism, and strong investor confidence, whereas a **bearish** market ($M = 0$) is marked by declining prices, pessimism, and caution.

Suppose an investor seeks to assess the market state based on recent news sentiment (N). Before any news is received, the investor holds a neutral view, believing that the market is equally likely to be bullish or bearish:

$$p_M(1) = 0.5 \quad \text{and} \quad p_M(0) = 0.5.$$

Historical data indicates that when the market is bullish, there is a 70% chance of receiving positive news:

$$p_{N|M}(1|1) = 0.7,$$

while if the market is bearish, the probability of positive news is only 30%:

$$p_{N|M}(1|0) = 0.3.$$

Upon receiving positive news, the investor wishes to update the probability that the market is bullish, i.e., $p_{M|N}(1|1)$. Using Bayes' theorem:

$$p_{M|N}(1|1) = \frac{p_{N|M}(1|1) \cdot p_M(1)}{p_N(1)}.$$

To calculate the evidence $p_N(1)$, we use the law of total probability:

$$p_N(1) = p_{N|M}(1|1) \cdot p_M(1) + p_{N|M}(1|0) \cdot p_M(0).$$

Substituting the values:

$$p_N(1) = (0.7)(0.5) + (0.3)(0.5) = 0.35 + 0.15 = 0.5.$$

Therefore, the updated probability that the market is bullish given the positive news is:

$$p_{M|N}(1|1) = \frac{(0.7)(0.5)}{0.5} = 0.7.$$

This means that after receiving positive news, the investor's belief that the market is bullish increases from 50% to 70%.