Stochastic Methods

Fabian Bosshard

July 17, 2025

Contents

Preface

This document is an unofficial student-made summary of the course Stochastic Methods taught by Francisco Javier Richter Mendoza with the assistance of Jacopo Quizi in Spring 2025 at the Università della Svizzera italiana. It is based on the lecture notes and other course materials. The summary is not exhaustive and may contain errors. If you find any, please report them to fabianlucasbosshard@gmail.com or open an issue at `https://github.com/fabianbosshard/usi-informatics-course-summaries`.

1 Random Numbers

A *random number* is an unpredictable value generated independently of other numbers, lacking any discernible pattern.

1.1 Random Number Generators

Definition 1 (Random Number Generator)**.** A Random Number Generator (RNG) is an algorithm that produces a sequence of numbers that appears random. Formally, an RNG is a function

$$R : S \to T$$

where:

- S is the *seed space* (a finite set of initial states). A seed $s \in S$ is used to initialize the RNG.
- T is the *target space* (typically the interval $[0, 1)$ or a set of integers).
- The function R maps each seed $s \in S$ to a target $u \in T$ in a manner that appears random. ✍

A high-quality random number generator should have the following properties:

- **Unpredictability**: Future values cannot be deduced without knowing the seed and algorithm.
- **Reproducibility**: Given the same seed, the RNG should produce the same sequence of numbers.
- **Representation of True Randomness**: All outcomes have an equitable chance.
- **Long period**: The sequence of numbers should be long before repeating.
- **Efficiency**: The RNG should generate numbers quickly.

Example 1. The *linear congruential generator* generates a sequence of random numbers via the following linear recurrence relation:

$$X_{n+1} = (aX_n + c) \bmod m$$

where:

- X_{n+1} is the next number in the sequence.
- X_n is the current number.
- a, c, and m are constants, known as the *multiplier*, *increment*, and *modulus*, respectively.
- mod denotes the modulus operation.
- The initial or seed value $X_0 = S$ is required to start the sequence. ◀

Example 2. The *PCG64 RNG* is the default generator in newer versions of NumPy. It combines a 128-bit lLCG with an output permutation to produce high-quality 64-bit pseudorandom numbers. The algorithm proceeds in two main stages:

1. **State Update**: The 128-bit state s_n is updated via the LCG:

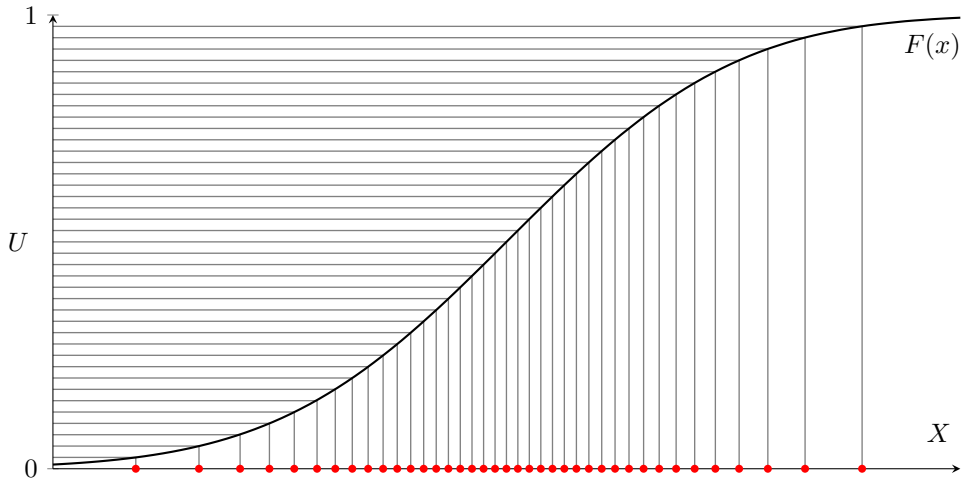
   $$s_{n+1} = (s_n \cdot a + c) \bmod 2^{128},$$

   where:
   - a is a carefully chosen 128-bit multiplier,
   - c is a 128-bit odd increment (ensuring a full period).

2. **Output Permutation**: A permutation is applied to the updated state to produce the final output.

This combination yields high-quality 64-bit pseudorandom numbers that are uniform, independent, and have an extremely long period. ◀

1.2 Inverse Transform Sampling



Theorem 1 (Inverse Transform Sampling)**.** Let U be uniformly distributed on $[0,1]$ and let $F_X(x)$ be the CDF of a random variable X with an invertible inverse $F_X^{-1}(u)$. Then the variable

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

has CDF $F_X(x)$. ◁

Proof. We show that for any $x \in \mathbb{R}$,

$$\begin{aligned} \mathrm{P}(X \le x) &= P\left(F_X^{-1}(U) \le x\right) \\ &= P\left(U \le F_X(x)\right) \quad \text{(since } F_X \text{ is strictly increasing)} \\ &= F_X(x) \quad \text{(because } U \text{ is uniformly distributed on } [0,1]). \end{aligned}$$

Thus, $X = F_X^{-1}(U)$ indeed has CDF $F_X(x)$. □

Because every number in $[0,1]$ is equally likely, the uniform distribution is ideal for generating samples from other distributions via inverse transform sampling.

Often, we require random numbers in the continuous interval $[0,1)$. A common technique to achieve this is to normalize the LCG's output. If the LCG produces integers in $\{0,1,\ldots,m-1\}$, then we define

$$\boxed{U_{n+1} = \frac{X_{n+1}}{m}}$$

This operation maps each integer to a real number in $[0,1)$.

Many experiments have outcomes defined on different sets. By running an RNG and then mapping its output (for example, via a modulus operation or parity check), we can sample from these universes. In essence, we first generate a number in $[0,1)$ and then use a suitable transformation to obtain the desired outcome.

2 Random Variables

2.1 Lebesgue Measure and Probability Measures

To provide a unified framework for probability, we define a standard measure on the interval $[0,1]$. For any subinterval $[a,b] \subseteq [0,1]$ with $0 \le a < b \le 1$, we assign

$$\mu([a,b]) = \mathrm{Leb}[a,b] = b - a.$$

This measure represents the length of the interval and, when normalized so that $\mu([0,1]) = 1$, it serves as the canonical probability measure on $[0,1]$. In practice, every set that can be constructed from intervals by countable unions, intersections, and complements (the so-called Borel sets) is measurable with respect to μ.

This construction is universal. Many experiments yield outcomes defined on various sets. Often we generate a random number in $[0,1)$ using an RNG - say, by normalizing an LCG's output-and then map that number to the desired set. More generally, if there is a measurable mapping

$$\phi : U \to [0,1],$$

we can define a probability measure P_U on U by *pulling back* μ:

$$\mathrm{P}_U(A) = \mu(\phi(A)),$$

for any measurable subset $A \subset U$. This approach transfers the well-understood properties of the Lebesgue measure on $[0,1]$ to any outcome space U, ensuring that probabilities are assigned consistently.

Definition 2. Borel Sets on $[0,1]$ The Borel σ-algebra on $[0,1]$, denoted by $\mathcal{B}([0,1])$, is the collection of all sets that can be formed from open intervals in $[0,1]$ by applying countable unions, countable intersections, and complementation. ✍

With a probability measure P established on our sample space U, we now introduce random variables as functions that assign numerical values to outcomes.

2.2 Random Variables

Definition 3 (Random Variable)**.** Let $(U, \mathcal{F}, P)$ be a probability space, where U is the sample space, $\mathcal{F}$ is a collection of measurable subsets of U (for instance, the Borel sets when $U \subset \mathbb{R}$), and P is the probability measure. A random variable is a measurable function

$$X : U \to \mathbb{R},$$

meaning that for every Borel set $B \subset \mathbb{R}$, the preimage $X^{-1}(B)$ is in $\mathcal{F}$. The set

$$R(X) = \{X(u) : u \in U\} \subset \mathbb{R}$$

is called the range or image of X. If $R(X)$ is countable, X is said to be discrete. ✍

Definition 4 (Probability Mass Function)**.** Let X be a discrete random variable with range $R(X)$. The *probability mass function* (pmf) of X is the function

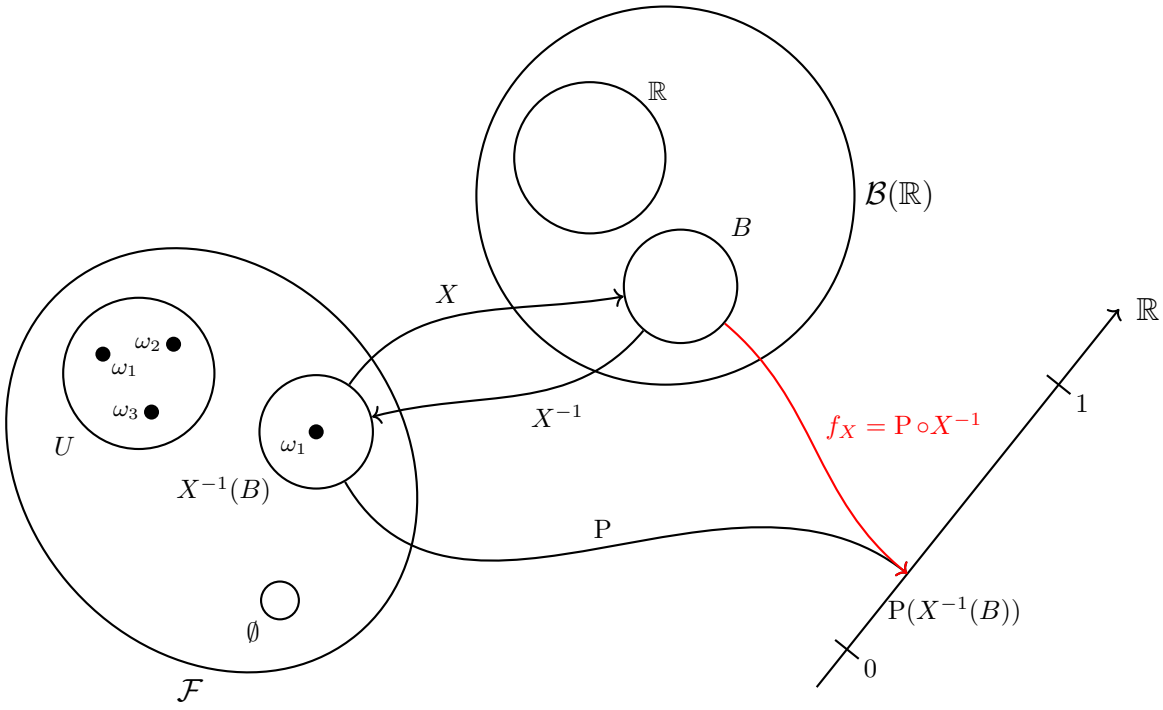
$$f_X : R(X) \to [0,1]$$

defined by

$$f_X(x) = \mathrm{P}(\{u \in U : X(u) = x\}), \quad \text{for each } x \in R(X).$$

The function f_X must satisfy:

1. $0 \leq f_X(x) \leq 1$ for all $x \in R(X)$,
2. $\sum_{x \in R(X)} f_X(x) = 1$. ✍



Definition 5 (Probability Density Function)**.** For a continuous random variable X with range $R(X) \subseteq \mathbb{R}$, the distribution is described by the *probability density function* (pdf) f_X. This function is nonnegative and satisfies

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$$

for any interval $[a, b] \subseteq R(X)$, with the normalization

$$\int_{R(X)} f_X(x)\mathrm{d}x = 1$$

✍

Definition 6. Cumulative Distribution Function The *cumulative distribution function* (CDF) of a random variable X is defined by

$$F(x) = \mathrm{P}(X \leq x).$$

✍

For a discrete random variable, this can be written as

$$F(x) = \sum_{t \leq x} f(t),$$

and for a continuous random variable with PDF $f(x)$, it is given by

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

In both cases, $F(x)$ represents the total probability that X does not exceed x.

3 Expectation and Variance

3.1 Expectation

Definition 7 (Expected Value)**.** If X is a random variable defined on a probability space $(U, \mathcal{F}, P)$, its expected value is defined by the integral

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x \, \mathrm{d}\, \mathrm{P}(x).$$

✍

For a continuous random variable with probability density function $f_X(x)$, the expected value is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x \, f_X(x) \mathrm{d}x.$$

For a discrete random variable - where the probability measure P is concentrated on a countable set $R(X) = \{x_1, x_2, \ldots\}$ - the integral reduces to the sum

$$\mathrm{E}[X] = \sum_i x_i \ \mathrm{P}(\{x_i\}) = \sum_{x \in R(X)} x \, f_X(x),$$

since the measure of a singleton $\{x\}$ is given by $\mathrm{P}(\{x\}) = f_X(x)$.

Theorem 2 (Linearity of Expectation)**.** Let $X_1, \ldots, X_n$ be random variables (not necessarily independent). For any constants $a_1, \ldots, a_n$, the expectation of their linear combination is given by

$$\mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \, \mathrm{E}\left[X_i\right].$$

◁

Proof.

$$\mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right] \ = \ \int \left(\sum_{i=1}^{n} a_i X_i\right) \mathrm{d}\, \mathrm{P} \ = \ \sum_{i=1}^{n} a_i \int X_i \, \mathrm{dP} \ = \ \sum_{i=1}^{n} a_i \, \mathrm{E}[X_i].$$

□

3.2 Variance

Definition 8 (Variance)**.** For a random variable X with expected value $\mu = \mathrm{E}[X]$, the variance is defined as

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

✍

3.3 Law of Large Numbers

The Law of Large Numbers (LLN) guarantees that, under specific conditions, the sample average converges to the theoretical expectation as the number of samples increases. This law underpins the intuitive notion that as we collect more independent observations, their average tends towards the true mean of the distribution.

Theorem 3 ((Weak) Law of Large Numbers)**.** Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with a finite expectation $\mathrm{E}[X_i]$. Then, as n approaches infinity, the sample average converges in probability to the expected value:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p} \mathrm{E}[X_i]$$

In practical terms, the average outcome from a large number of trials will approximate the expected value, and this approximation improves with more trials. ◁

3.4 Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental result in probability theory and statistics. It states that the sum of a large number of independent and identically distributed (i.i.d.) random variables, each with finite mean and variance, tends to be normally distributed, regardless of the original distribution of the variables.

Theorem 4 (Central Limit Theorem)**.** Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with expectation $\mathrm{E}[X_i] = \mu$ and variance $\mathrm{Var}[X_i] = \sigma^2$, both finite. Then, as n approaches infinity, the standardized sum converges in distribution to the standard normal distribution:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution. ◁

To understand this theorem, let's analyze the characteristics of this result.

3.4.1 Aggregation of Random Effects

When summing many independent random variables, each contributing its own randomness, individual irregularities tend to "average out," leading to predictable overall behavior.

Consider n i.i.d. random variables $X_1, \ldots, X_n$, each with mean μ and variance σ^2.

$$\begin{aligned} S_n &= \textstyle\sum_{i=1}^{n} X_i \\ \mathrm{E}[S_n] &= n\mu \\ \mathrm{Var}(S_n) &= n\sigma^2 \end{aligned}$$

To analyze the behavior as n grows, we standardize the sum:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

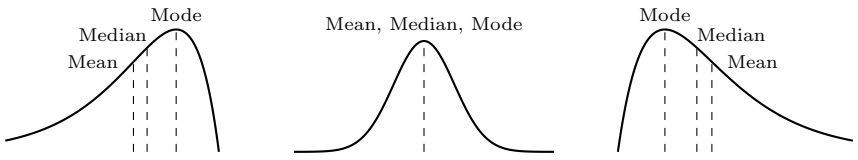
As n increases, the standardized sum Z_n becomes more stable. The "aggregation" of individual random effects leads to a reduction in relative fluctuations, making Z_n less influenced by the variability of any single X_i.

3.4.2 Symmetry Through Averaging

As more variables are added, the influence of any single variable diminishes. This averaging effect induces symmetry in the distribution of the sum.

Define the standardized individual variables:

$$Y_i = \frac{X_i - \mu}{\sigma}$$



Negative / Left Skew | Symmetric | Positive / Right Skew

Thus, the standardized sum can be expressed as:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Each Y_i has:

$$\mathrm{E}[Y_i] = 0, \quad \mathrm{Var}(Y_i) = 1$$

The Law of Large Numbers ensures that:

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty$$

If the original distribution of Y_i is skewed, the sum $\sum Y_i$ tends to balance out the skewness as positive and negative deviations cancel each other out. The skewness of Z_n diminishes as n increases:

$$\text{Skewness}(Z_n) = \frac{\mathrm{E}[Z_n^3]}{(\mathrm{Var}(Z_n))^{3/2}} = \frac{\gamma}{\sqrt{n}} \to 0 \quad \text{as} \quad n \to \infty$$

Where γ is the third central moment of Y_i.

The distribution of Z_n becomes increasingly symmetric around zero as n grows, regardless of the original distribution's symmetry. This emerging symmetry is a crucial step toward the normal distribution's characteristic bell shape.

3.4.3 Emergence of the Bell Curve

The normal distribution (bell curve) is inherently symmetric and arises naturally when multiple independent random factors contribute to a single outcome.

Proof. The characteristic function of Z_n is given by:

$$\begin{aligned}
\varphi_{Z_n}(t) &= \mathrm{E}\left[e^{\imath t Z_n}\right] \\
&= \mathrm{E}\left[e^{\imath t \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}}\right] \\
&= \mathrm{E}\left[e^{\sum_{i=1}^{n} \imath \frac{t}{\sqrt{n}} Y_i}\right] \\
&= \mathrm{E}\left[\prod_{i=1}^{n} e^{\imath \frac{t}{\sqrt{n}} Y_i}\right] \\
&= \prod_{i=1}^{n} \mathrm{E}\left[e^{\imath \frac{t}{\sqrt{n}} Y_i}\right] \quad \text{(independent)} \\
&= \left(\varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n \quad \text{(identically)}
\end{aligned}$$

The expansion of the characteristic function of Y_i around $t = 0$ is given by:

$$\begin{aligned}\varphi_{Y_i}(t) &= \mathrm{E}\left[e^{\imath t Y_i}\right] \\ &= \mathrm{E}\left[1 + \imath t Y_i - \frac{t^2 Y_i^2}{2} + o(t^2)\right] \\ &= 1 + \imath t\ \mathrm{E}\left[Y_i\right] - \frac{t^2}{2}\ \mathrm{E}\left[Y_i^2\right] + o(t^2).\end{aligned}$$

Assuming that Y_i is standardized, i.e.

$$\mathrm{E}\left[Y_i\right] = 0 \quad \text{and} \quad \mathrm{E}\left[Y_i^2\right] = 1,$$

we obtain

$$\varphi_{Y_i}(t) = 1 - \frac{t^2}{2} + o(t^2).$$

Substituting $\frac{t}{\sqrt{n}}$ for t, it follows that

$$\begin{aligned}\varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) &= 1 - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\left(\frac{t}{\sqrt{n}}\right)^2\right) \\ &= 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).\end{aligned}$$

Then,

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \xrightarrow[n\to\infty]{} e^{-t^2/2}$$

The characteristic function $e^{-t^2/2}$ uniquely corresponds to the standard normal distribution $\mathcal{N}(0, 1)$.

As n increases, the characteristic function of Z_n converges to that of the normal distribution, indicating that Z_n approaches $\mathcal{N}(0, 1)$. □

4 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. A central application is

4.1 Monte Carlo Integration

Let $f(x)$ be a real-valued function defined on a domain D with finite measure $|D|$ and suppose we wish to evaluate the integral

$$I = \int_D f(x)\mathrm{d}x.$$

If X is drawn uniformly[1] from D, then the expected value of $f(X)$ is

$$\mathrm{E}[f(X)] = \int f(x)p(x)\mathrm{d}x = \frac{1}{|D|}\int_D f(x)\,\mathrm{d}x = \frac{I}{|D|},$$

or equivalently,

$$\int_D f(x)\,\mathrm{d}x = |D|\,\mathrm{E}[f(X)] = I.$$

A Monte Carlo estimator for this integral is

$$\boxed{\hat{I}_N = |D|\frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right)}$$

where $x_1, \ldots, x_N$ are independent samples drawn uniformly from D. By the LLN (Theorem 3), as $N \to \infty$, $\hat{I}_N$ converges to I.

4.1.1 Importance Sampling

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

$$I = \int_D f(x)\mathrm{d}x = \int_D \frac{f(x)}{p(x)}p(x)\mathrm{d}x = \mathrm{E}_p\Big[\frac{f(x)}{p(x)}\Big].$$

The corresponding Monte Carlo estimator is

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

with the x_i drawn according to $p(x)$. The variance of this estimator is given by

$$\mathrm{Var}(\hat{I}) = \frac{1}{N}\,\mathrm{Var}\Big[\frac{f(x)}{p(x)}\Big],$$

which implies that the standard error decreases as $1/\sqrt{N}$. Choosing a sampling distribution $p(x)$ that closely resembles $f(x)$ can reduce the variance.

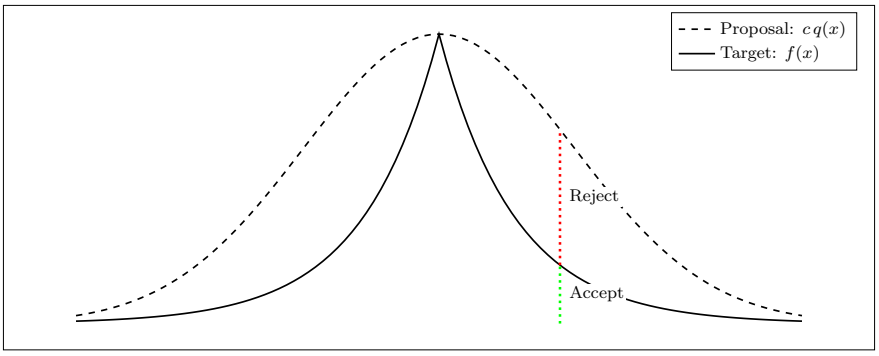
[1] i.e. with pdf $p(x) = \frac{1}{|D|}\mathbf{1}_D(x)$, where $\mathbf{1}_D(x)$ is the indicator function that is 1 if $x \in D$ and 0 otherwise

4.2 Rejection Sampling (als known as Accept-Reject Method)

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 5 (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

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

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. First we need the joint density (1) of X and I:

$$f_{X,I}(x, i) = \underbrace{f_X(x)}_{q(x)} \cdot \underbrace{f_{I|X}(i \mid x)}_{\text{Bern}\left(\frac{f(x)}{cq(x)}\right)} = q(x) \left(\frac{f(x)}{cq(x)}\right)^i \left(1 - \frac{f(x)}{cq(x)}\right)^{1-i}, \quad i \in \{0, 1\}$$

The conditional probability distribution (Def. 11) of X given $I(X, U)$ is

$$f_{X|I}(x \mid i) = \frac{f_{X,I}(x, i)}{f_I(i)}$$

For the denominator, integrate out x:

$$\begin{aligned}
f_I(i) &= \int_{-\infty}^{\infty} f_{X,I}(x, i)\, \mathrm{d}x \\
&= \int_{-\infty}^{\infty} q(x) \left(\frac{f(x)}{cq(x)}\right)^i \left(1 - \frac{f(x)}{cq(x)}\right)^{1-i} \mathrm{d}x \\
&= \begin{cases} \int_{-\infty}^{\infty} \frac{f(x)}{c}\, \mathrm{d}x, & i = 1 \\ \int_{-\infty}^{\infty} q(x) - \frac{f(x)}{c}\, \mathrm{d}x, & i = 0 \end{cases} \\
&= \begin{cases} \frac{1}{c}, & i = 1 \\ 1 - \frac{1}{c}, & i = 0 \end{cases} \\
&= \left(\frac{1}{c}\right)^i \left(1 - \frac{1}{c}\right)^{1-i}
\end{aligned}$$

So we have

$$\begin{aligned} f_{X|I}(x \mid i) &= \frac{f_{X,I}(x,i)}{f_I(i)} \\ &= \frac{q(x)\left(\frac{f(x)}{cq(x)}\right)^i \left(1-\frac{f(x)}{cq(x)}\right)^{1-i}}{\left(\frac{1}{c}\right)^i \left(1-\frac{1}{c}\right)^{1-i}} \end{aligned}$$

and the conditional distribution of X given $I(X,U)=1$ is

$$\begin{aligned} f_{X|I=1}(x \mid i=1) &= \frac{f_{X,I}(x,i=1)}{f_I(i=1)} \\ &= \frac{q(x)\left(\frac{f(x)}{cq(x)}\right)}{\frac{1}{c}} \\ &= f(x) \end{aligned}$$

□

4.3 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.

Definition 9 (Independence)**.** 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.

Definition 10 (Indicator Function of an Event)**.** For an event $A \subset \Omega$, the 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 density $p(x)$, then

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

✍

If A and B are two events, then their joint probability is

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

If X and Y are independent, then

$$\mathrm{E}[XY] = \mathrm{E}[X]\ \mathrm{E}[Y].$$

So we have the following equivalence:

$$A \text{ and } B \text{ are independent } \iff \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B).$$

Definition 11 (Conditional Probability)**.** Given two random variables X and Y with $\mathrm{P}(Y = y) > 0$, the conditional probability of $X = x$ given $Y = y$ is defined as:

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

✍

A direct consequence is the multiplication rule:

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

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

For many random variables $X_1, \ldots, X_n$, the joint probability is given by:

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

Definition 12 (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 join tprobabilities over all possible outcomes y for Y:

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

where the sum is over all possible outcomes of Y. ✍

The connection between marginal and conditional probability can be udnerstood through the law of total probability:

$$p_X(x) = \sum_{y \in \mathrm{Im}(Y)} p_{X|Y}(x|y)p_Y(y).$$

Theorem 6 (Bayes' Theorem)**.** Let X and Y be random variables with $p_Y(y) \neq 0$. Then, the posterior probability is given by

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

◁

Proof. Symmetry of joint probabilities. □

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. 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.

5 Random Networks

Random network models provide a probabilistic framework for studying the structure and dynamics of complex systems.

5.1 Fundamental Concepts

Definition 13 (Graph)**.** A graph G is defined as a pair $G = (V, E)$, where:
- V is a set of nodes (or vertices),
- E is a set of edges connecting pairs of nodes. ✍

Definition 14 (Adjacency Matrix)**.** For a graph $G = (V, E)$ with $N = |V|$ nodes, the adjacency matrix $\underline{\boldsymbol{A}}$ is an $N \times N$ matrix defined by

$$(\underline{\boldsymbol{A}})_{uv} = \begin{cases} 1, & \text{if there is an edge between nodes } u \text{ and } v, \\ 0, & \text{otherwise.} \end{cases}$$

✍

Theorem 7. Let $\underline{\boldsymbol{A}}$ be the adjacency matrix of a graph G. Then, the entry $(\underline{\boldsymbol{A}}^k)_{uv}$ equals the number of walks of length k from node u to node v. ◁

Proof. (Induction) For $k = 1$, the claim holds by definition. Assume that $(\underline{\boldsymbol{A}}^k)_{uv}$ counts the number of walks of length k from u to v. For $k + 1$, we have:

$$(\underline{\boldsymbol{A}}^{k+1})_{uv} = \sum_{w \in V} (\underline{\boldsymbol{A}}^k)_{uw} (\underline{\boldsymbol{A}})_{wv}.$$

Each term $(\underline{\boldsymbol{A}}^k)_{uw}$ counts the number of walks of length k from u to w, and $(\underline{\boldsymbol{A}})_{wv}$ indicates the presence of an edge from w to v. Summing over all w gives the total number of walks of length $k + 1$ from u to v, completing the induction. □

5.2 Random Network Models

Random network models provide a fundamental framework for generating graphs using simple probabilistic rules. These models help us understand how local random interactions can lead to the emergence of complex global network structures.

Definition 15 (Degree Distribution)**.** For a graph $G = (V, E)$, the degree of a node is defined as the number of edges incident to it. The degree distribution $P(k)$ is the probability that a randomly selected node has degree k:

$$P(k) = \frac{|\{u \in V : \deg(u) = k\}|}{|V|},$$

with the normalization condition $\sum_k P(k) = 1$. ✍

5.2.1 Erdős-Rényi Model

Definition 16 (Erdős-Rényi Model)**.** The Erdős-Rényi model, denoted by $G(n, p)$, is one of the simplest random graph models. It constructs a graph with n nodes by considering each of the $\binom{n}{2}$ possible edges and including each edge independently with probability p. That is, for any pair of distinct nodes u and v,

$$P((u, v) \in E) = p.$$

✍

Key properties include:

- **Degree Distribution:** In $G(n,p)$, the degree k of any node follows a binomial distribution,
$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$
For large n and small p (with np constant), this can be approximated by a Poisson distribution:
$$P(k) \approx \frac{(np)^k}{k!} \mathrm{e}^{-np}.$$
- **Connectivity and Phase Transition:** A critical phenomenon occurs at the threshold
$$p_c \approx \frac{\ln(n)}{n},$$
where the graph transitions from having many small disconnected components to containing a single giant component.
- **Clustering Coefficient:** Because edges are formed independently, the clustering coefficient (the probability that two neighbors of a node are connected) is low, roughly equal to p.

Due to its simplicity, the Erdős-Rényi model is mathematically tractable and provides valuable insights into the fundamental behavior of random graphs. However, its assumptions of independent and uniform edge formation limit its ability to capture clustering and degree heterogeneity observed in many real-world networks.

5.2.2 Small–World Networks: The Watts–Strogatz Model

Small–world networks feature high local clustering and short average path lengths.

Definition 17 (Regular Ring Lattice)**.** A regular ring lattice with n nodes is constructed by arranging the nodes in a circle and connecting each node to its $k/2$ nearest neighbors on each side. ✍

Definition 18 (Watts–Strogatz Model)**.** Given a regular ring lattice with n nodes and degree k, the Watts–Strogatz model introduces randomness by rewiring each edge with a probability p. For each edge (i,j) in the lattice, with probability p:

1. Remove the edge (i,j).
2. Choose a new node l uniformly at random from all nodes such that $l \neq i$ and there is no existing edge between i and l.
3. Add the edge (i,l) to the graph. ✍

Key properties include the preservation of high clustering for small p and a dramatic reduction in the average path length due to long–range shortcuts.

5.2.3 Scale–Free Networks: The Barabási–Albert Model

Scale–free networks are characterized by the presence of hubs.

Definition 19 (Barabási–Albert Model)**.** The Barabási–Albert model constructs a network via preferential attachment:

1. Start with a small, connected network of m_0 nodes.
2. At each time step, add a new node with m edges (with $m \leq m_0$). The probability $\Pi(k_i)$ that the new node attaches to an existing node i is proportional to the degree k_i:
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}.$$

✍

This "rich–get–richer" mechanism leads to a power–law degree distribution $P(k) \sim k^{-\gamma}$ (typically $\gamma \approx 3$).

5.3 The Friendship Paradox

The friendship paradox is the counterintuitive phenomenon that on average, your friends tend to have more friends than you do. Let $G = (V, E)$ be an undirected graph with N nodes, where each node i has degree d_i. Define

$$\mu = \frac{1}{N} \sum_{i=1}^{N} d_i.$$

When an edge is chosen uniformly at random, the probability that its endpoint has degree k is proportional to $k\,P(k)$. Hence, the expected degree of a node reached by following a random edge is

$$E[d_{\text{friend}}] = \frac{E[k^2]}{\mu}.$$

Since $E[k^2] \geq \mu^2$ (with equality only if all nodes have the same degree),

$$E[d_{\text{friend}}] \geq \mu.$$

The following theorem captures this formally.

Theorem 8 (Friendship Paradox)**.** In any graph whose degree distribution is not uniform, the average degree of a randomly selected neighbor is strictly larger than the average degree of a randomly selected node:

$$E[d_{\text{friend}}] = \frac{E[k^2]}{\mu} \geq \mu,$$

with equality if and only if all nodes have the same degree. ◁

Proof. Let $P(k)$ be the probability that a randomly chosen node has degree k. Since a node of degree k is k times more likely to be reached by following a random edge, the probability for a neighbor having degree k is

$$P_{\text{friend}}(k) = \frac{kP(k)}{\mu}.$$

Thus,

$$E[d_{\text{friend}}] = \sum_k k\,P_{\text{friend}}(k) = \frac{1}{\mu} \sum_k k^2 P(k) = \frac{E[k^2]}{\mu}.$$

Expressing $E[k^2]$ as $\mu^2 + \text{Var}(k)$ shows that $E[d_{\text{friend}}] \geq \mu$ with equality only if $\text{Var}(k) = 0$. □

6 Stochastic Processes and Markov Chains

Definition 20 (Stochastic Process)**.** A stochastic process is a collection of random variables $\{X_n\}_{n \geq 0}$ defined on a common probability space, where the index n typically represents time. ✍

Definition 21 (Markov Chain)**.** A stochastic process $\{X_n\}_{n \geq 0}$ is called a Markov chain if for all $n \geq 0$ and for all states $s_0, s_1, \ldots, s_{n+1}$ we have

$$\mathrm{P}\Big(X_{n+1} = s_{n+1} \mid X_n = s_n, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0\Big) = \mathrm{P}\Big(X_{n+1} = s_{n+1} \mid X_n = s_n\Big)$$

for all states $s \in S$, where S is the (countable) state space. ✍

In other words, the future depends on the past only through the current state.

Let T denote the waiting time (in number of steps) until a transition occurs (i.e., the first time the chain leaves its current state). The Markov property implies that if no transition occurs by time n, the additional waiting time is independent of the past. Thus, if no transition has occurred by time n, then for any $m \in \mathbb{N}$

$$\mathrm{P}(T > n + m \mid T > n) = \frac{\mathrm{P}(T > n + m)}{\mathrm{P}(T > n)} = \mathrm{P}(T > m).$$

Defining $G(k) = \mathrm{P}(T > k)$ (with $G(0) = 1$), we have

$$G(n+m) = G(n)G(m), \quad \forall n, m \in \mathbb{N}.$$

It is known that the only solution to this functional equation is

$$G(k) = q^k, \quad 0 < q < 1.$$

Thus, the probability mass function for T is

$$\mathrm{P}(T = k) = G(k-1) - G(k) = q^{k-1}(1-q), \quad k \geq 1.$$

This derivation shows that the only discrete distribution satisfying the memoryless property is the geometric distribution.

For a Markov chain with a finite state space of size N, the transition probabilities are represented by the transition matrix,

$$\underline{\boldsymbol{P}} = [p_{ij}] \in \mathbb{R}^{N \times N}, \quad \text{where} \quad p_{ij} = \mathrm{P}(X_{t+1} = j \mid X_t = i),$$

with each row summing to 1:

$$\sum_{j=1}^{N} p_{ij} = 1, \quad \text{for all } i.$$

Example 3 (Random Walk)**.** A random walk is a classic example of a Markov chain. In the simplest one–dimensional random walk:

1. Start at position 0.
2. At each time step, flip a fair coin:
   - If heads, move one step to the right $(+1)$.
   - If tails, move one step to the left (-1).
3. Record the position after each step.

The future position depends only on the current position and the coin flip, not on how the current position was reached. ◀

A two–dimensional random walk can be similarly defined by allowing moves in four directions (up, down, left, right).

6.1 Multi-Step Transitions

Definition 22 (n-Step Transition Probability)**.** The n-step transition probability $P_{ij}^{(n)}$ is defined as the probability of transitioning from state i to state j in n steps:

$$P_{ij}^{(n)} = \mathrm{P}(X_{k+n} = j \mid X_k = i)$$

for any $k \geq 0$ and $i, j \in S$. In particular, $P_{ij}^{(1)} = p_{ij}$ are the one-step probabilities. ✍

The n-step transition probabilities can be computed from the transition matrix $\underline{\boldsymbol{P}}$ as follows:

Theorem 9 (Chapman-Kolmogorov)**.** For any nonnegative integers $n, m \geq 0$,

$$P_{ij}^{(n+m)} = \sum_{k \in S} P_{ik}^{(n)} P_{kj}^{(m)}$$

or, in matrix form,

$$\underline{\boldsymbol{P}}^{(n+m)} = \underline{\boldsymbol{P}}^{(n)} \underline{\boldsymbol{P}}^{(m)}$$

Hence, $\underline{\boldsymbol{P}}^{(n)} = \underline{\boldsymbol{P}}^{n}$, i.e. $\underline{\boldsymbol{P}}^{(n)}$ is the n-th power of the transition matrix $\underline{\boldsymbol{P}}$. ◁

6.2 Classification of States

Understanding how a Markov chain behaves over many steps requires classifying its states and determining whether certain long-term distributions exist.

6.2.1 Communicating Classes and Irreducibility

Definition 23 (Communication)**.** States i and j communicate if $P_{ij}^{(n)} > 0$ for some n and $P_{ji}^{(m)} > 0$ for some m. A set of states C is a communicating class if every pair of states in C communicates and no state outside C communicates with a state in C. ✍

Definition 24 (Irreducibility)**.** A Markov chain is irreducible if, for every pair of states i and j, there exists an integer $n \geq 1$ such that

$$P_{ij}^{(n)} > 0$$

i.e. the entire state space S is one single communicating class. In other words, one can get from any state i to any other state j in a finite number of steps with positive probability. ✍

6.2.2 Recurrence and Transience

Definition 25 (Recurrence, Transience)**.** A state i is recurrent if, starting from i, the expected number of visits to i is infinite; equivalently, the probability of returning to i at some time in the future is 1. If that probability is less than 1, then i is transient. ✍

In finite Markov chains, irreducible classes are automatically recurrent (and at least one class may be absorbing if there is a state with $P_{ii} = 1$).

Definition 26 (Positive Recurrence)**.** A state i is positive recurrent if the expected return time to i, starting from i, is finite:

$$\mathrm{E}_i[T_i] = \sum_{n=1}^{\infty} n \ \mathrm{P}(T_i = n) < \infty$$

where T_i is the first return time to state i. A Markov chain is positive recurrent if all states are positive recurrent. ✍

6.2.3 Periodicity

Definition 27 (Period)**.** The period of a state i is

$$d(i) = \gcd\{n \geq 1 : P_{ii}^{(n)} > 0\}$$

where gcd is the greatest common divisor. If $d(i) = 1$, the state i is aperiodic. A Markov chain is aperiodic if all its states are aperiodic. In an irreducable chain, it suffices to check just one state. ✍

If a Markov chain is irreducible and aperiodic (i.e. ergodic), it has some nice properties (see 6.2.4).

6.2.4 Stationary Distributions

Definition 28 (Ergodicity)**.** A Markov chain is ergodic if it is irreducible, aperiodic, and positive recurrent. ✍

A probability vector $\vec{\boldsymbol{\pi}} = (\pi_i)_{i \in S}$ is called a stationary distribution if

$$\vec{\boldsymbol{\pi}} \underline{\boldsymbol{P}} = \vec{\boldsymbol{\pi}}, \quad \sum_j \pi_j = 1$$

For a finite irreducible aperiodic chain, there exists a unique stationary distribution $\vec{\boldsymbol{\pi}}$ and, moreover, the long-run behavior is given by

$$\lim_{n \to \infty} P_{ij}^{(n)} = \pi_j, \quad \text{for all states } i$$

This means the chain forgets its initial state in the long run and converges to $\vec{\boldsymbol{\pi}}$.

In the context of Markov chains, the terms "invariant distribution" and "stationary distribution" are usually used interchangeably.

Example 4 (Doubly Stochastic Chain)**.** Let the stace space be $S = \{1, \ldots, N\}$ and suppose the transition matrix $\underline{\boldsymbol{P}} = [p(j \mid i)]$ satisfies

$$\sum_{i=1}^{N} p(j \mid i) = 1, \quad \forall j \in S$$

i.e., the sum of each column is 1 (in addition to the usual condition that the sum of each row is 1). Assume the uniform distribution $\pi_i = \frac{1}{N}$ for $i \in S$. Then, for each $j \in S$,

$$(\vec{\boldsymbol{\pi}} \underline{\boldsymbol{P}})_j = \sum_{i=1}^{N} \pi_i \cdot p(j \mid i) = \frac{1}{N} \sum_{i=1}^{N} p(j \mid i) = \frac{1}{N} \cdot 1 = \frac{1}{N} = \pi_j$$

Thus, $\vec{\boldsymbol{\pi}} \underline{\boldsymbol{P}} = \vec{\boldsymbol{\pi}}$, and the uniform distribution is invariant. ◀

6.3 Absorbing Markov Chains

A Markov chain is absorbing if it has at least one state i with $P_{ii} = 1$ (such a state is called absorbing), and from every state in the chain, there is some way (positive-probability path) to eventually enter an absorbing state.

One typically reorders the states so that the absorbing states come last, yielding a transition matrix in the form

$$\underline{\boldsymbol{P}} = \begin{bmatrix} \underline{\boldsymbol{Q}} & \underline{\boldsymbol{R}} \\ \underline{\boldsymbol{0}} & \underline{\boldsymbol{I}} \end{bmatrix}$$

where $\underline{\boldsymbol{Q}}$ is the transition matrix among the transient states and $\underline{\boldsymbol{I}}$ is an identity matrix for the absorbing states. The fundamental matrix $\underline{\boldsymbol{N}}$ is

$$\underline{\boldsymbol{N}} = (\underline{\boldsymbol{I}} - \underline{\boldsymbol{Q}})^{-1}$$

Its $(i, j)^{\text{th}}$ entry N_{ij} is the expected number of visits to state j starting from state i before absorption occurs. The matrix $\underline{\boldsymbol{NR}}$ gives absorption probabilities into each absorbing state.

6.4 Branching Processes

Branching processes model how populations evolve when each individual reproduces independently of others. The canonical example:

Definition 29 (Galton-Watson Process)**.** Let $Z_0 = 1$. Each individual in generation n produces a random number of offspring in generation $n+1$ according to a fixed distribution $\{P_k\}_{k=0}^{\infty}$. Formally,

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i}$$

where $X_{n,i}$ are i.i.d. with $\mathrm{P}(X_{n,i} = k) = P_k$. ✍

One key question is whether the population eventually dies out (i.e., hits $Z_n = 0$ for some n). Define the generating function

$$f(s) = \sum_{k=0}^{\infty} P_k s^k$$

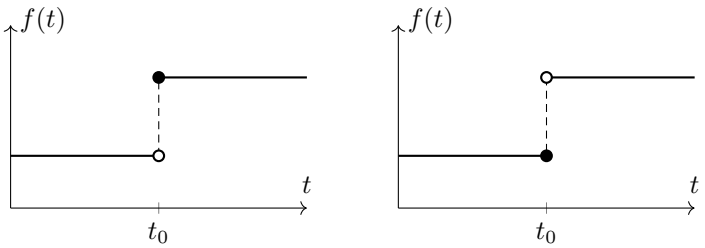
Then the extinction probability π_0 is a fixed point of f, i.e., π_0 satisfies $\pi_0 = f(\pi_0)$.

7 Counting Processes

Poisson processes are a fundamental concept in stochastic modeling, providing a rigorous mathematical framework for understanding events that occur randomly in time or space. Arrival times and counting processes (such as Poisson processes) find applications in a myriad of contexts, each with its own set of challenges and implications.

Definition 30 (Counting Process)**.** A counting process $\{N(t)\}$, where $t \geq 0$, is a stochastic process (Definition 20) that represents the total number of events that have occured up to time t. The function $N(t)$ satisfies the following properties:

- $N(0) = 0$ (initial condition)
- $N(t)$ is integer-valued for all $t \geq 0$
- $N(t)$ is non-decreasing: $t_1 < t_2 \implies N(t_1) \leq N(t_2)$
- $N(t)$ is right-continuous: $\lim_{t \to t_0^+} N(t) = N(t_0)$ ✍



right-continuous left-continuous

A counting process counts the number of times a certain event has occurred by any given time t. The count starts at zero and can only increase as time moves forward.

7.0.1 From Binomial to Poisson

Consider a time interval T divided into n smaller intervals of length $\Delta t = \frac{T}{n}$. We are interested in counting the number of occurrences of a particular event within each small time interval Δt.

Initially, let us model this as a Bernoulli process. In each small time interval Δt, the event can either occur with probability p or not occur with probability $1 - p$:

$$\mathrm{P}(\text{Event occurs in } \Delta t) = p, \quad \mathrm{P}(\text{Event does not occur in } \Delta t) = 1 - p$$

For large n and small Δt, we can relate p to a rate parameter λ by

$$p = \lambda \Delta t$$

Let X be the number of events that occur in the entire interval $[0, T]$. The variable X is a sum of n independent Bernoulli random variables, each with success probability p. Thus,

$$X \sim \text{Binomial}(n, p)$$

The probability of observing exactly k events in T is

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Substituting $p = \lambda \Delta t$ and using $\Delta t = \frac{T}{n}$, we get

$$\begin{aligned}
\mathrm{P}(X = k) &= \binom{n}{k} (\lambda \Delta t)^k (1 - \lambda \Delta t)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda T}{n}\right)^k \left(1 - \frac{\lambda T}{n}\right)^{n-k} \\
&= \frac{\prod_{i=1}^{k}(n-i+1)}{n^k} \frac{(\lambda T)^k}{k!} \left(1 - \frac{\lambda T}{n}\right)^{n} \left(1 - \frac{\lambda T}{n}\right)^{-k}
\end{aligned}$$

As $n \to \infty$ (and therefore $\Delta t \to 0$ while $n\Delta t = T$ remains constant),

$$\frac{\prod_{i=1}^{k}(n-i+1)}{n^k} \longrightarrow 1, \quad \left(1-\frac{\lambda T}{n}\right)^n \longrightarrow e^{-\lambda T}, \quad \left(1-\frac{\lambda T}{n}\right)^{-k} \longrightarrow 1$$

hence

$$\lim_{n\to\infty} \mathrm{P}(X=k) \;=\; \lim_{n\to\infty} \binom{n}{k} (\lambda\Delta t)^k (1-\lambda\Delta t)^{n-k} \;=\; \frac{(\lambda T)^k}{k!}\, e^{-\lambda T}$$

which is the Probability Mass Function (PMF) of a Poisson(λT) random variable.

Theorem 10 (Convergence of Binomial to Poisson)**.** Let X be the number of events in a time interval T divided into n smaller intervals of length $\Delta t = \frac{T}{n}$. If each interval has a success probability $p = \lambda\Delta t$, then as $n \to \infty$ with $n\Delta t = T$ fixed,

$$\lim_{n\to\infty} \mathrm{P}(X=k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \tag{2}$$

Hence, X converges to a Poisson distribution with parameter λT (see 11). ◁

7.1 Poisson Process

models scenarios where events occur randomly in continuous time (or space) at a certain average rate.

Definition 31 (Homogeneous Poisson Process)**.** A homogeneous Poisson process is a counting process $N(t)$ with the following properties:

- $N(0) = 0$
- increments are independent
- # events in an interval of length t is Poisson with mean λt ✍

A crucial insight is that the waiting times between successive events in a homogeneous Poisson process are exponentially distributed with parameter λ:

7.1.1 From Poisson to Exponential

Consider a Poisson process with rate λ, where $N(t) \sim \text{Poisson}(\lambda t)$ represents the number of events occurring in time t. Let T be the waiting time until the first event, defined as

$$T = \inf\{t > 0 \mid N(t) \geq 1\}$$

The probability of no events occurring in $[0, t]$ is given by

$$P(T > t) = P(\text{no events in } [0,t]) = P(N(t) = 0) \overset{(2)}{=} e^{-\lambda t}$$

Since the survival function of an exponential distribution with rate λ is $e^{-\lambda t}$, it follows that

$$T \sim \text{Exp}(\lambda)$$

This implies that the Poisson process also has the memoryless property, meaning that the time until the next event occurs does not depend on how much time has already passed since the last event (see Theorem 12).

Theorem 11 (Exponential Waiting Times)**.** The waiting time T until the first event is exponentially distributed:

$$\mathrm{P}(T \leq t) = 1 - e^{-\lambda t}$$ ◁

Theorem 12 (Memoryless Property)**.** For an exponentially distributed waiting time T,

$$\begin{aligned}
\mathrm{P}(T > s+t \mid T > s) &= \frac{\mathrm{P}(T > s+t \cap T > s)}{\mathrm{P}(T > s)} \\
&= \frac{\mathrm{P}(T > s+t)}{\mathrm{P}(T > s)} \quad (\text{because } \{T > s+t\} \subseteq \{T > s\}) \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} \\
&= \mathrm{P}(T > t)
\end{aligned}$$

and hence the underlying Poisson process is memoryless. ◁

In many real-world settings, the rate of arrival $\lambda(t)$ varies over time. A non-homogeneous Poisson process generalizes the basic Poisson framework by allowing $\lambda(t)$ to be a function of t. Then:

- $\{N(t)\}$ still has independent increments
- # events in $[s,t]$ is Poisson with mean $\int_s^t \lambda(u)\mathrm{d}u$

7.2 Birth-Death Process

classic family of continuous-time Markov chains often used to model population dynamics. Let $N(t)$ be the population at time t. Suppose λ_n is the birth rate when the population is n and μ_n is the death rate when the population is n. Define $P_n(t) = \mathrm{P}(N(t) = n)$. The master equation (or Kolmogorov forward equation) reads:

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \lambda_{n-1}P_{n-1}(t) - (\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t)$$

Such processes can be simulated (see 7.3) or analyzed theoretically for their steady-state behavior.

7.3 Gillepsie Algorithm

The Gillespie algorithm (also known as the stochastic simulation algorithm) is widely used for simulating discrete-event systems, particularly chemical reaction networks. It can also be applied to birth-death processes, queueing systems, and any scenario where events occur randomly in continuous time with state-dependent rates.

8 Stochastic Diffusion Processes

8.1 Brownian Motion

8.1.1 From Random Walks to Diffusions

Brownian motion sits at the interface between discrete counting processes studied earlier and continuous–time models. It emerges as the diffusive limit of a random walk with shrinking step–size and time–grid.

Definition 32 (Standard Brownian Motion)**.** A process $\{W_t\}_{t\geq 0}$ is a standard Brownian motion if

1. $W_0 = 0$ a.s.
2. Independent increments: $W_t - W_s$ is independent of the past for $0 \leq s < t$.
3. $W_t - W_s \sim \mathcal{N}(0, t-s)$.

4. Paths $t \mapsto W_t$ are continuous a.s.

✍

Key properties.

- **Martingale:** $\mathrm{E}[W_t \mid \mathcal{F}_s] = W_s$.
- **Scaling:** $c^{-1/2} W_{ct}$ is again Brownian.
- **Quadratic variation:** $[W]_t = t$.
- **Nowhere differentiable:** paths are rough on every scale.

8.2 Stochastic Differential Equations

8.3 Fokker-Planck Equation

9 Statistical Learning and Stochastic Inference

Definition 33 (Program)**.** A program is a triple $(M, \theta, P_\varepsilon)$ where

$$M : I \times \Omega \to O \quad \text{(deterministic map)}\ , \quad \theta \in \Theta \quad \text{(parameters)}\ , \quad \varepsilon \sim P_\varepsilon \quad \text{(randomness)}$$

where I is the input space, O is the output space, and Ω is the parameter space.

So that for input $X = x$,

$$Y = M_\theta(x, \varepsilon), \quad Y \mid X = x \sim p_\theta(\cdot \mid x)$$

✍

- Estimation
- Prediction

machine learning mostly focused on prediction (black box)

9.1 Maximum Likelihood Estimation

9.2 Bootstrapping

Definition 34 (Bootstrap)**.** Given data $\vec{\boldsymbol{X}} = (X_1, \ldots, X_n)$ and an estimator $\hat{\theta} = t(X_1, \ldots, X_n)$

1. For $b = 1, \ldots, B$, draw a bootstrap sample $X_1^{*(b)}, \ldots, X_n^{*(b)}$ by sampling with replacement from $\{X_i\}$
2. Compute the bootstrap estimate $\hat{\theta}^{*(b)} = t(X_1^{*(b)}, \ldots, X_n^{*(b)})$

The empirical distribution of $\{\hat{\theta}^{*(b)}\}_{b=1}^{B}$ approximates the sampling distribution of $\hat{\theta}$. ✍

The bootstrap standard error is estimated as:

$$\widehat{\mathrm{SE}}_{\mathrm{boot}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*(b)} - \overline{\hat{\theta}^*} \right)^2}, \quad \overline{\hat{\theta}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*(b)}$$

Several methods exist for constructing bootstrap confidence intervals:

10 Stochastic Optimization

10.1 Stochastic Gradient Descent

Consider the stochastic-gradient-descent (SGD) recursion

$$\theta_{k+1} = \theta_k - \eta \widehat{f'(\theta_k)}, \qquad \widehat{f'(\theta_k)} = g(\theta_k) + \epsilon_k$$

where the learning-rate is $\eta > 0$, $g : \mathbb{R} \to \mathbb{R}$ is deterministic, and the noises are i.i.d. $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$.

Theorem 13. When the step size is small ($\eta = \Delta t \ll 1$) and one rescales the variance $\sigma^2 = \tilde{\sigma}^2/\eta$, the continuous-time interpolation $\Theta_t \approx \theta_{\lfloor t/\Delta t \rfloor}$ converges (in distribution) to the Itô SDE

$$\mathrm{d}\Theta_t = -g(\Theta_t)\,\mathrm{d}t + \tilde{\sigma}\mathrm{d}W_t, \qquad \Theta_0 = 1$$

where W_t is standard Brownian motion. ◁

Proof. We can view k as an index for time, where k corresponds to the time $t_k = k \cdot \Delta t$ and thus $\Theta_{t_k} = \theta_{\lfloor t_k/\Delta t \rfloor} = \theta_k$.

We can rewrite the SGD recursion as

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \Delta t \cdot \widehat{f'(\theta_k)} \\
&= \theta_k - \Delta t \cdot (g(\theta_k) + \epsilon_k), \qquad \epsilon_k \sim \mathcal{N}(0, \sigma^2) \\
&= \theta_k - \Delta t \cdot g(\theta_k) - \Delta t \cdot \epsilon_k \\
&= \theta_k - \Delta t \cdot g(\theta_k) + \Delta t \cdot \epsilon_k \qquad \text{(symmetry of normal around 0)} \\
&= \theta_k - \Delta t \cdot g(\theta_k) + \Delta t \cdot \sigma \cdot \delta_k, \qquad \delta_k \sim \mathcal{N}(0,1) \\
&= \theta_k - \Delta t \cdot g(\theta_k) + \Delta t \cdot \tfrac{\tilde{\sigma}}{\sqrt{\Delta t}} \cdot \delta_k \\
&= \theta_k - \Delta t \cdot g(\theta_k) + \tilde{\sigma} \cdot \sqrt{\Delta t} \cdot \delta_k \\
&= \theta_k - \underbrace{\Delta t \cdot g(\theta_k)}_{\text{deterministic}} + \underbrace{\tilde{\sigma} \cdot \Delta W_k}_{\text{random}}, \qquad \Delta W_k \sim \mathcal{N}(0, \Delta t)
\end{aligned}$$

$$\implies \boxed{\Delta\theta_k = \theta_{k+1} - \theta_k = -\Delta t \cdot g(\theta_k) + \tilde{\sigma} \cdot \Delta W_k}$$

Summing over N steps, we have

$$\sum_{k=1}^{N} \Delta\theta_k = -\sum_{k=1}^{N} \Delta t \cdot g(\theta_k) + \tilde{\sigma} \sum_{k=1}^{N} \Delta W_k$$

As $\Delta t \to 0$ and $N \to \infty$, we have $\Delta W_k \to \mathrm{d}W_t$ and $\Delta t \to \mathrm{d}t$, and thus

$$\int \mathrm{d}\Theta_t = -\int g(\Theta_t)\mathrm{d}t + \tilde{\sigma} \int \mathrm{d}W_t$$

so,

$$\mathrm{d}\Theta_t = -g(\Theta_t)\mathrm{d}t + \tilde{\sigma}\mathrm{d}W_t$$

where W_t is standard Brownian motion. □

11 Important Distributions

Notation	**PDF/PMF**	**Exp.**	**Var.**
$\text{Unif}(a, b)$	$f(x) = \dfrac{1}{b-a}, \quad a \le x \le b$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
$\text{Bern}(p)$	$f(k) = p^k(1-p)^{1-k}, \quad k = 0, 1$	p	$p(1-p)$
$\text{Bin}(n, p)$	$f(k) = \binom{n}{k} p^k(1-p)^{n-k}, \quad k = 0, \ldots, n$	np	$np(1-p)$
$\text{Geom}(p)$	$f(k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
$\text{Pois}(\lambda)$	$f(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, \ldots$	λ	λ
$\text{Exp}(\beta)$	$f(x) = \dfrac{1}{\beta} e^{-x/\beta}, \quad x \ge 0$	β	β^2
$\text{N}(\mu, \sigma^2)$	$f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{1}{2}\dfrac{(x-\mu)^2}{\sigma^2}\right)$	μ	σ^2
$\text{N}_d(\vec{\boldsymbol{\mu}}, \underline{\boldsymbol{\Sigma}})$	$f(\vec{\boldsymbol{x}}) = \dfrac{1}{\sqrt{(2\pi)^d \lvert\underline{\boldsymbol{\Sigma}}\rvert}} \exp\left(-\dfrac{1}{2}(\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}})^\top \underline{\boldsymbol{\Sigma}}^{-1} (\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}})\right)$	$\vec{\boldsymbol{\mu}}$	$\underline{\boldsymbol{\Sigma}}$