40.520 Stochastic Models

Necessary Probability Background

Xiaotang Yang

Engineering Systems and Design (ESD)
Singapore University of Technology and Design (SUTD)

Spring 2026

Outline

1. Probability Review
2. Sample Path, Convergence and Average

Probability Review

Probability Model

A probability space is $(\Omega, \mathcal{F}, P)$

- Sample space Ω: set of all possible outcomes
- $\mathcal{F}$: Collection of events (σ-algebra) such that
  (a) $\emptyset \in \mathcal{F}$
  (b) $E^c \in \mathcal{F}$ whenever $E \in \mathcal{F}$
  (c) $\cup_{n \geq 1} E_n \in \mathcal{F}$ whenever $E_n \in \mathcal{F}$ for every $n \geq 1$.
- P: Probability measure $\mathcal{F} \to [0, 1]$ such that
  (a) $P(\Omega) = 1$
  (b) $P(\cup_{n \geq 1} E_n) = \sum_{n=1}^{\infty} P(E_N)$ where $E_1, E_2, \cdots \in \mathcal{F}$ are disjoint

Some Properties of A Probability Measure

1. If $E \subseteq F$, then $P(E) \leq P(F)$.
2. $P(E^c) = 1 - P(E)$.
3. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Conditional Probabilities

Definition

$$P\{E \mid F\} = \frac{P\{E \cap F\}}{P\{F\}}, \quad \text{where } P\{F\} > 0$$

Interpretation

- Probability of E given we've narrowed sample space to points in F
- Like focusing on subset of outcomes

Elementary Properties of Conditional Probabilities

- $P(E \mid E) = ?$
- $P(\varnothing \mid E) = ?$
- $P(F \mid E) = ?$, for $F \supseteq E$
- $P(F_1 \cup F_2 \mid E) = P(F_1 \mid E) + P(F_2 \mid E)$,
  where F_1 and F_2 are disjoint subsets (mutually exclusive) of E.

Mutually exclusive: $E_1 \cap E_2 = \emptyset$

Chain rule

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)\, P(E_2 \mid E_1)\, P(E_3 \mid E_1 \cap E_2) \ldots P(E_n \mid E_1 \cap E_2 \cap \cdots \cap E_{n-1})$$

Independence

- **Definition**: $P\{E \cap F\} = P\{E\} \cdot P\{F\}$
- **Implication**: $P\{E \mid F\} = P\{E\}$
- **Note:** Mutually exclusive $\neq$ Independent
  - If E and F are mutually exclusive and non-null, they cannot be independent

Conditionally Independence

- E and F are conditionally independent given G if:

$$P\{E \cap F \mid G\} = P\{E \mid G\} \cdot P\{F \mid G\}, \quad \text{where } P\{G\} > 0$$

- **Note**: Independence $\neq$ Conditional Independence

Law of Total Probability

Basic Form For any event F:

$$P\{E\} = P\{E \cap F\} + P\{E \cap F^c\} = P\{E \mid F\}P\{F\} + P\{E \mid F^c\}P\{F^c\}$$

General Form If $F_1, F_2, \ldots, F_n$ partition Ω:

$$P\{E\} = \sum_{i=1}^{n} P\{E \cap F_i\} = \sum_{i=1}^{n} P\{E \mid F_i\} \cdot P\{F_i\}$$

Warning:

- Events must:
    1. Be mutually exclusive
    2. Sum to whole sample space (partition Ω)

Example

Consider the probability that a person is late to class, which depends on the weather. The weather can be classified into three mutually exclusive and exhaustive conditions: Rainy (R), Cloudy (C), and Sunny (S). The historical probabilities for each weather type are:

$$P(R) = 0.2, \quad P(C) = 0.5, \quad P(S) = 0.3.$$

The conditional probabilities of being late (L) given the weather are:

$$P(L \mid R) = 0.6, \quad P(L \mid C) = 0.3, \quad P(L \mid S) = 0.1.$$

What is the probability that the person is late to class?

Bayes' Law

Theorem

$$P\{F \mid E\} = \frac{P\{E \mid F\} \cdot P\{F\}}{P\{E\}}$$

Extended Form (with partition) If $F_1, F_2, \ldots, F_n$ partition Ω:

$$P\{F_i \mid E\} = \frac{P\{E \mid F_i\} \cdot P\{F_i\}}{\sum_{j=1}^{n} P\{E \mid F_j\} \cdot P\{F_j\}}$$

Bayes' Law

Medical Test Example

- Disease prevalence: 1 in 10,000
- Test accuracy: 95% (both true positive and true negative rates)
- **Question**: $P\{\text{Disease} \mid \text{Test positive}\}$?

Random Variables

Definitions

- **Random Variable (r.v.)**: Real-valued function of experiment outcome
- **Discrete r.v.**: Takes countably set of values
- **Continuous r.v.**: Takes uncountable set of values

Key Insight

- "$X = k$" is an event $\rightarrow$ All probability theorems apply to r.v.'s

Discrete: Probability Mass Function (PMF)

Definition For discrete r.v. X:

$$p_X(a) = P\{X = a\}, \quad \sum_x p_X(x) = 1$$

Cumulative Distribution Functions

$$F_X(a) = P\{X \leq a\} = \sum_{x \leq a} p_X(x), \quad \bar{F}_X(a) = P\{X > a\} = 1 - F_X(a)$$

Common Discrete Distributions

1. **Bernoulli(p)**: Single trial, success prob p

$$X = \begin{cases} 1 & \text{(success) w/ prob } p \\ 0 & \text{(failure) w/ prob } 1-p \end{cases}$$

2. **Binomial(n, p)**: # successes in n independent Bernoulli(p) trials

$$p_X(i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n$$

3. **Geometric(p)**: # trials until first success

$$p_X(i) = (1-p)^{i-1} p, \quad i = 1, 2, \ldots$$

4. **Poisson(λ)**: Counts occurrences in fixed interval

$$p_X(i) = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots$$

Continuous: Probability Density Function (PDF)

Definition For continuous r.v. X:

- $f_X(x) \geq 0$
- $P\{a \leq X \leq b\} = \int_a^b f_X(x)\,dx$
- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Interpretation

- $f_X(x)dx \approx P\{x \leq X \leq x + dx\}$
- $f_X(x) \neq P\{X = x\}$ (which is 0 for continuous r.v.)

Cumulative Distribution Function (CDF)

$$F_X(a) = P\{-\infty < X \leq a\} = \int_{-\infty}^{a} f_X(x)\,dx,$$

$$f_X(x) = \frac{d}{dx} F_X(x) \quad \text{(by Fundamental Theorem of Calculus)}$$

Common Continuous Distributions

1. Uniform(a, b)

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}, \quad F_X(x) = \frac{x-a}{b-a} \ \ (\text{for } a \leq x \leq b)$$

2. Exponential(λ): Memoryless property

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}, \quad F_X(x) = 1 - e^{-\lambda x} \quad (\text{for } x \geq 0)$$

3. Pareto(α)

$$f_X(x) = \begin{cases} \alpha x^{-\alpha-1} & x \geq 1 \\ 0 & \text{otherwise} \end{cases}, \quad F_X(x) = 1 - x^{-\alpha}$$

Heavy-tailed: decays polynomially (vs exponentially)

4. Normal(μ, σ) More on this later.

Expectation and Variance

Expectation (Mean)

$$\text{Discrete: } E[X] = \sum_x x \cdot p_X(x), \quad \text{Continuous: } E[X] = \int x \cdot f_X(x)\, dx$$

Interpretation: Weighted average of possible values

Variance

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

Interpretation: Measures spread around mean

Properties of Expectation

For a nonnegative integer-valued random variable X,

$$E[X] = \sum_{n=1}^{\infty} P(X \geq n).$$

For a nonnegative continuous random variable Y:

$$E[Y] = \int_0^{\infty} \left[1 - F_Y(t)\right] dt$$

Examples

Compute $E[N]$ when $N \sim Geo(p)$

Properties of Expectation

Let X be a discrete random variable with pmf $p_X(x)$. For a real-valued function $g(X)$,

$$E[g(X)] = \sum_x g(x)\, p_X(x).$$

Let X be a continuous random variable with pdf $f_X(x)$. For a real-valued function $g(X)$,

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

Example

Let the waiting time T (in minutes) for the train follow a distribution with the following PDF.

$$f_T(t) = \frac{3}{10}\left(\frac{t}{10}\right)^2 e^{-(t/10)^3}, \quad t \geq 0$$

The **discomfort cost** is a nonlinear function of waiting time:

$$g(T) = 50\sqrt{T} + 2T^2$$

Calculate $E[C]$, the **expected discomfort cost**.

Properties of Expectation

Theorem (Linearity of Expectation) For any random variables X and Y:

$$E[X + Y] = E[X] + E[Y]$$

No independence required!

Example: Binomial Mean $X \sim \text{Binomial}(n, p) = X_1 + \cdots + X_n$ where $X_i \sim \text{Bernoulli}(p)$. Compute $E[X]$.

Example: Hat Problem n people, random hat assignment $X = \#$ people getting own hat. Compute $E[X]$.

Joint Probabilities and Independence

Joint Distributions

- **Discrete**: $p_{X,Y}(x, y) = P\{X = x, Y = y\}$
- **Continuous**: $f_{X,Y}(x, y)$ where $P\{a < X < b, c < Y < d\} = \int_c^d \int_a^b f_{X,Y}(x, y)\, dx\, dy$

Marginal Distributions

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad f_X(x) = \int f_{X,Y}(x, y)\, dy$$

Joint Probabilities and Independence

Independence

- **Discrete**: $X \perp Y$ if $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y) \ \forall x, y$
- **Continuous**: $X \perp Y$ if $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) \ \forall x, y$

Theorem If $X \perp Y$, then:

1. $E[XY] = E[X] \cdot E[Y]$
2. $E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$

Warning: $E[XY] = E[X]E[Y]$ does NOT imply $X \perp Y$

Conditional Probabilities and Expectations (Discrete)

Conditional PMF Given event A with $P\{A\} > 0$:

$$p_{X|A}(x) = P\{X = x \mid A\} = \frac{P\{(X = x) \cap A\}}{P\{A\}}$$

Conditional Expectation

$$E[X \mid A] = \sum_x x \cdot p_{X|A}(x)$$

Conditional on Random Variable
For two discrete random variables X and Y, the conditional PMF of X given $Y = y$ is

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

Conditional Probabilities and Expectations (Continuous)

Conditional PDF Given $A \subseteq \mathbb{R}$ with $P\{X \in A\} > 0$:

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P\{X \in A\}} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Conditional expectation

$$E[X \mid A] = \int_A x f_{X|A}(x)\, dx$$

Conditional on Random Variable
For two continuous random variables X and Y, the conditional PDF of X given $Y = y$ is

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

Example: Pittsburgh Supercomputing Center

Setup

- Job durations: $X \sim \text{Exp}(1/1000)$ hours
- Bin 1: jobs < 500 hours
- Bin 2: jobs ≥ 500 hours

Questions

1. $P\{\text{Job in bin 1}\}$
2. $P\{\text{Duration} < 200 \mid \text{bin 1}\}$
3. Conditional density: $f_{X|\text{bin1}}(t)$
4. $E[\text{Duration} \mid \text{bin 1}]$

Probabilities & Expectations via Conditioning

Law of Total Probability for R.V.'s

- **Discrete**: $P\{X = k\} = \sum_y P\{X = k \mid Y = y\} \cdot P\{Y = y\}$
- **Continuous**: $f_X(x) = \int f_{X|Y}(x|y) \cdot f_Y(y)\, dy$

Law of Iterated Expectations: $E[X] = E[E[X \mid Y]]$

- **Discrete**: $E[X] = \sum_y E[X \mid Y = y] \cdot P\{Y = y\}$
- **Continuous**: $E[X] = \int E[X \mid Y = y] \cdot f_Y(y)\, dy$

Probabilities & Expectations via Conditioning

Example: Which Exponential Happens First? $X_1 \sim \text{Exp}(\lambda_1), X_2 \sim \text{Exp}(\lambda_2), X_1 \perp X_2$

$$P\{X_1 < X_2\} = ?$$

Geometric Mean via Conditioning Let $N \sim \text{Geometric}(p)$, calculate $E[N]$ by conditioning on the first flip.

Polya's Urn Model

Polya's urn model supposes that an urn initially contains r red and b blue balls. At each stage a ball is randomly selected from the urn and is then returned along with m other balls of the same color. Let X_k be the number of red balls drawn in the first k selections.

(a) Find $E[X_1]$.

(b) Find $E[X_2]$.

(c) Find $E[X_3]$.

Variance and Independence

Theorem If $X \perp Y$, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

Proof

$$\begin{aligned}\text{Var}(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] \\ &\quad - (E[X]^2 + E[Y]^2 + 2E[X]E[Y]) \\ &= \text{Var}(X) + \text{Var}(Y) + 2[E[XY] - E[X]E[Y]]\end{aligned}$$

If $X \perp Y$, $E[XY] = E[X]E[Y] \rightarrow$ last term $= 0$

Without Independence

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

where $\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$

Normal (Gaussian) Distribution

Definition $X \sim \text{Normal}(\mu, \sigma^2)$ if:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

Standard Normal $Z \sim \text{Normal}(0, 1)$: $\Phi(z) = P\{Z \leq z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$

Properties

- Bell-shaped, symmetric around μ
- $E[X] = \mu$, $\text{Var}(X) = \sigma^2$
- **Linear Transformation Property**: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$

Standardization $X \sim \text{Normal}(\mu, \sigma^2) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \sim \text{Normal}(0, 1)$

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

Central Limit Theorem (CLT)

Setup $X_1, X_2, \ldots, X_n$ i.i.d. with mean μ, variance σ^2

$$S_n = X_1 + \cdots + X_n$$

Theorem

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \text{Normal}(0,1) \text{ as } n \to \infty$$

i.e., $\lim_{n\to\infty} P\{Z_n \le z\} = \Phi(z)$

Implications

- Sum of i.i.d. r.v.'s $\approx$ Normal for large n
- **Approximately**: $S_n \sim \text{Normal}(n\mu, n\sigma^2)$
- Applies to any distribution (discrete/continuous)

Applications

- $\text{Binomial}(n,p) \approx \text{Normal}(np, np(1-p))$ for large n
- $\text{Poisson}(\lambda) \approx \text{Normal}(\lambda, \lambda)$ for large λ

Sum of Random Number of Random Variables

Setup $S = \sum_{i=1}^{N} X_i$ where:

- X_i i.i.d.
- N is non-negative integer r.v.
- $N \perp X_i$

Key Results

1. $E[S] = E[N] \cdot E[X]$
2. $E[S^2] = E[N] \cdot \mathrm{Var}(X) + E[N^2] \cdot (E[X])^2$
3. $\mathrm{Var}(S) = E[N] \cdot \mathrm{Var}(X) + \mathrm{Var}(N) \cdot (E[X])^2$

Sample Path, Convergence and Average

Convergence of Random Variables

Almost Sure Convergence

$Y_n \xrightarrow{a.s.} \mu$ if

$$\forall k > 0, \quad P\left(\lim_{n\to\infty} |Y_n - \mu| > k\right) = 0$$

"Almost all sample paths eventually stay close to μ"

Convergence in Probability

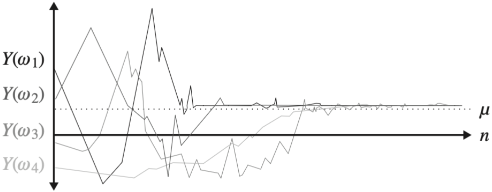
$Y_n \xrightarrow{P} \mu$ if

$$\forall k > 0, \quad \lim_{n\to\infty} P\left(|Y_n - \mu| > k\right) = 0$$

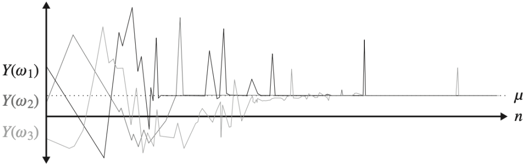
"Probability of being far from μ vanishes as n grows"

Note: Almost sure convergence $\Rightarrow$ Convergence in probability

Visualizing Convergence



Almost sure convergence
Individual paths converge



Convergence in probability
Mass of "bad" paths shrinks

Key Insight

Convergence in probability does **not** imply individual sample paths converge!

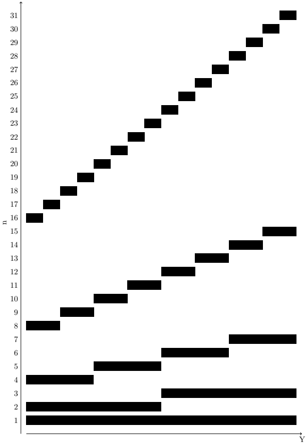
Typewriter Sequence

Construction

- Let $Y \sim U[0,1]$
- For $n = 2^k + m$, where $k = 0, 1, 2, \ldots,$ and $m = 0, 1, 2, \ldots, 2^k - 1$,

$$X_n = \begin{cases} 1 & \text{if } Y \in [\frac{m}{2^k}, \frac{m+1}{2^k}] \\ 0 & \text{otherwise} \end{cases}$$

- $X_n \to 0$ in probability
- X_n does not converge to 0 almost surely



Laws of Large Numbers

Let $X_1, X_2, \ldots$ be i.i.d. with mean $E[X]$. Define $S_n = \sum_{i=1}^{n} X_i$.

Weak Law (WLLN)

$\frac{S_n}{n} \xrightarrow{P} E[X]$. Convergence in probability.

Strong Law (SLLN)

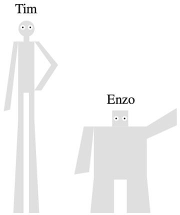
$\frac{S_n}{n} \xrightarrow{a.s.} E[X]$. Convergence almost surely (with probability 1).

Implication

SLLN $\Rightarrow$ WLLN, but not conversely

Time Average versus Ensemble Average

- Two students in our class: Tim and Enzo
- Simulate FCFS queues to determine the average number of jobs in the system



Tim's Approach	**Enzo's Approach**
One very long sample path	Many independent shorter runs
Logs system state over time	Samples at fixed time t
Computes **time average**	Averages across runs $\rightarrow$ **ensemble average**

Question. Who is "right"? Tim or Enzo?

Two Types of Averages

Time Average

Along one sample path ω:

$$\overline{N}^{\text{Time Avg}}(\omega) = \lim_{t\to\infty} \frac{1}{t}\int_0^t N(v,\omega)\,dv$$

Example: **Tim's** single long simulation

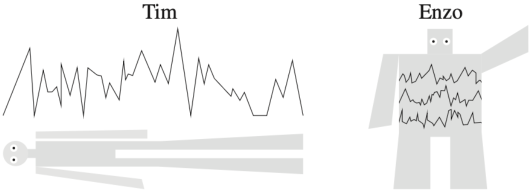
Ensemble Average

Across all sample paths:

$$\overline{N}^{\text{Ensemble}} = \lim_{t\to\infty} E[N(t)] = \sum_i i \cdot p_i$$

where $p_i = \lim_{t\to\infty} P\{N(t) = i\}$

Example: **Enzo's** many independent runs



Equivalence Under Ergodicity

Theorem (Ergodic Theorem)

For an ergodic system:

$$\overline{N}^{\text{Time Avg}} = \overline{N}^{\text{Ensemble}} \quad \textit{(with probability 1)}$$

Ergodic

- **Positive recurrent**: Finite mean time between returns to any state
- **Aperiodic**: No periodic ties to time steps
- **Irreducible**: Can reach any state from any state

Consequence

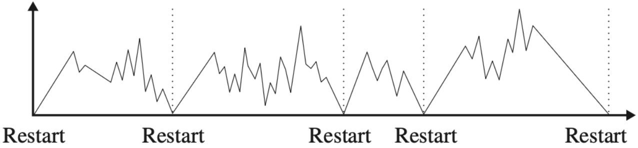
Initial conditions become irrelevant in the limit

What is Ergodicity?

- **Irreducible**: System can explore all states
- **Positive Recurrent**: Returns to states infinitely often with finite mean time
- **Aperiodic**: No fixed periodic patterns

Intuition

A single long run contains many independent "renewals"
$\Rightarrow$ behaves like many independent runs



Practical Implications for Simulation

Time Average (Tim)	**Ensemble Average (Enzo)**
One long simulation	Many independent runs
Lower overhead	Naturally parallelizable
No confidence intervals	Enables confidence intervals
Sensitive to initial transient	Must wait for steady state each run
Both converge to same value for ergodic systems	

When to use which?

- **Time average**: Quick exploration, limited resources
- **Ensemble average**: Need confidence intervals, parallel computing available