Chapter 1. Review of Probability Theory and Statistics

1 Probability Space and Rules of Probability

To any experiment we assign its **sample space**, denoted by S, consisting of all its possible outcomes (called **elementary events**, denoted by $e_i,\ i \in \mathbb{N}$).
An **event** is a subset of S (events are denoted by capital letters, $A, B, A_i,\ i \in \mathbb{N}$).

Since events are defined as sets, we use set theory in describing them.

- two special events associated with every experiment:
  - the **impossible** event, denoted by $\emptyset$ ("never happens");
  - the **sure (certain)** event, denoted by S ("surely happens").
- for events, we have the usual operations of sets:
  - **complementary** event, $\overline{A}$,
  - **union** of A and B, $A \cup B = \{e \in S \mid e \in A \text{ or } e \in B\}$, the event that occurs if either A or B or both occur;
  - **intersection** of A and B, $A \cap B = \{e \in S \mid e \in A \text{ and } e \in B\}$, the event that occurs if both A and B occur;
  - **difference** of A and B, $A \setminus B = \{e \in S \mid e \in A \text{ and } e \notin B\} = A \cap \overline{B}$, the event that occurs if A occurs and B does not;
  - A **implies (induces)** B, $A \subseteq B$, if every element of A is also an element of B, or in other words, if the occurrence of A induces (implies) the occurrence of B; A and B are equal, $A = B$, if A implies B and B implies A;
- two events A and B are **mutually exclusive (disjoint, incompatible)** if A and B cannot occur at the same time, i.e. $A \cap B = \emptyset$;
- three or more events are mutually exclusive if **any two of them are**, i.e.

$$A_i \cap A_j = \emptyset,\ \forall i \neq j;$$

- events $\{A_i\}_{i \in I}$ are **collectively exhaustive** if $\bigcup_{i \in I} A_i = S$;

- events $\{A_i\}_{i\in I}$ form a **partition** of S if the events are collectively exhaustive and mutually exclusive, i.e.
$$\bigcup_{i\in I} A_i = S, \text{ and } A_i \cap A_j = \emptyset, \forall i, j \in I,\ i \neq j.$$
- we consider all events relating to an experiment to belong to a **σ-field**, $\mathcal{K}$, a collection of events from from S, an algebraic structure that allows all the usual set operations (mentioned above) within itself (e.g. the power set $\mathcal{P}(S) = \{S' | S' \subseteq S\}$).

Definition 1.1. *Let $\mathcal{K}$ be a σ-field over S. A mapping $P : \mathcal{K} \to \mathbb{R}$ is called **probability** if it satisfies the following conditions:*

(i) $P(S) = 1$;

(ii) $P(A) \geq 0$, *for all* $A \in \mathcal{K}$;

(iii) *for any sequence $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{K}$ of mutually exclusive events,*
$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n). \tag{1.1}$$

*The triplet $(S, \mathcal{K}, P)$ is called a **probability space**.*

Theorem 1.2. *(*Rules of Probability*)*
Let $(S, \mathcal{K}, P)$ be a probability space, and let $A, B \in \mathcal{K}$. Then the following properties hold:

a) $P(\overline{A}) = 1 - P(A)$.

b) $0 \leq P(A) \leq 1$.

c) $P(\emptyset) = 0$.

d) $P(A \setminus B) = P(A) - P(A \cap B)$.

e) *If $A \subseteq B$, then $P(A) \leq P(B)$, i.e. P is monotonically increasing.*

f) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

g) *more generally,*
$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k)$$
$$+ \ldots + (-1)^{n-1} P\Big(\bigcap_{i=1}^{n} A_i\Big), \text{ for all } n \in \mathbb{N}.$$

Definition 1.3. *Let $(S, \mathcal{K}, P)$ be a probability space and let $B \in \mathcal{K}$ be an event with $P(B) > 0$. Then for every $A \in \mathcal{K}$, the* ***conditional probability of*** *A* ***given*** *B (or the* ***probability of*** *A* ***conditioned by*** *B) is defined by*

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)}. \tag{1.2}$$

Theorem 1.4. (Rules of Probability $-$ Continued)

h) $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$.

i) *Multiplication Rule*
$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1) \ldots P(A_n|A_1 \cap \ldots \cap A_{n-1})$.

j) *Total Probability Rule*

- $P(A) = P(B)P(A|B) + P\left(\overline{B}\right) P\left(A|\overline{B}\right)$.
- *in general, if $\{A_i\}_{i \in I}$ is a partition of S,*

$$P(A) \;=\; \sum_{i \in I} P\left(A_i\right) P\left(A|A_i\right). \tag{1.3}$$

Definition 1.5. *Two events $A, B \in \mathcal{K}$ are* ***independent*** *if*

$$P(A \cap B) = P(A)P(B). \tag{1.4}$$

- A, B independent $<=>$ $P(A|B) = P(A)$ $<=>$ $P(B|A) = P(B)$.
- $A = \emptyset$ or $A = S$ and $B \in \mathcal{K}$, then A, B independent.
- A, B independent $<=>$ $\overline{A}, B$ independent $<=>$ $\overline{A}, \overline{B}$ independent.

Definition 1.6. *Consider an experiment whose outcomes are finite and equally likely. Then the* ***probability*** *of the event A is given by*

$$P(A) \;=\; \frac{\text{number of favorable outcomes for the occurrence of } A}{\text{total number of possible outcomes of the experiment}} \;\overset{not}{=}\; \frac{N_f}{N_t}. \tag{1.5}$$

Remark 1.7. This notion is closely related to that of *relative frequency* of an event A: repeat an experiment a number of times N and count the number of times event A occurs, N_A. Then the relative frequency of the event A is

$$f_A = \frac{N_A}{N}.$$

Such a number is often used as an approximation to the probability of A. This is justified by the fact that

$$f_A \overset{N\to\infty}{\longrightarrow} P(A).$$

The relative frequency is used in computer simulations of random phenomena.

2 Probabilistic Models

Binomial Model

This model is used when the trials of an experiment satisfy three conditions, namely

(i) they are independent,

(ii) each trial has only two possible outcomes, which we refer to as "success" (A) and "failure" ($\overline{A}$) (i.e. the sample space for each trial is $S = A \cup \overline{A}$),

(iii) the probability of success $p = P(A)$ is the same for each trial (we denote by $q = 1-p = P(\overline{A})$ the probability of failure).

Trials of an experiment satisfying (i) $-$ (iii) are known as **Bernoulli trials**.

Model: Given n Bernoulli trials with probability of success p, find the probability $P(n;k)$ of exactly k $(0 \le k \le n)$ successes occurring.

We have

$$\begin{aligned} P(n;k) &= C_n^k p^k (1-p)^{n-k} = C_n^k p^k q^{n-k}, \quad k = 0, 1, \ldots, n \text{ and} \\ \sum_{k=0}^{n} P(n;k) &= 1. \end{aligned} \tag{2.1}$$

Pascal (Negative Binomial) Model

Model: Consider an infinite sequence of Bernoulli trials with probability of success p (and probability of failure $q = 1 - p$) in each trial. Find the probability $P(n,k)$ of the nth success occurring

after k failures ($n \in \mathbb{N},\ k \in \mathbb{N} \cup \{0\}$).
We have

$$\begin{aligned} P(n,k) &= C_{n+k-1}^{k} p^n q^k, \quad k = 0, 1, \ldots \text{ and} \\ \sum_{k=0}^{\infty} P(n;k) &= 1. \end{aligned} \tag{2.2}$$

Geometric Model

Although a particular case for the Pascal Model (case $n = 1$), the Geometric model comes up in many applications and deserves a place of its own.
Model: Consider an infinite sequence of Bernoulli trials with probability of success p (and probability of failure $q = 1 - p$) in each trial. Find the probability p_k that the first success occurs after k failures ($k \in \mathbb{N} \cup \{0\}$).
Here, we have

$$\begin{aligned} p_k &= pq^k, \quad k = 0, 1, \ldots \text{ and} \\ \sum_{k=0}^{\infty} p_k &= 1. \end{aligned} \tag{2.3}$$

3 Random Variables

3.1 Random Variables, PDF and CDF

Random variables, variables whose observed values are determined by chance, give a more comprehensive quantitative overlook of random phenomena. Random variables are the fundamentals of modern Statistics.

Definition 3.1. *Let $(S, \mathcal{K}, P)$ be a probability space. A **random variable** is a function $X : S \to \mathbb{R}$ satisfying the property that for every $x \in \mathbb{R}$, the event*

$$(X \leq x) \ := \ \{e \in S \mid X(e) \leq x\} \in \mathcal{K}. \tag{3.1}$$

- *if the set of values that it takes, $X(S)$, is at most countable in $\mathbb{R}$, then X is a **discrete random variable** (quantities that are counted);*
- *if $X(S)$ is a continuous subset of $\mathbb{R}$ (an interval), then X is a **continuous random variable** (quantities that are measured).*

For each random variable, discrete or continuous, there are two important functions associated with it:

- **PDF (probability distribution/density function)**
  - if X is discrete, then the pdf is an array

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}, \tag{3.2}$$

    where $x_i \in \mathbb{R},\ i \in I$, are the values that X takes and $p_i = P(X = x_i)$
  - if X is continuous, then the pdf is a function $f : \mathbb{R} \to \mathbb{R}$;
- **CDF (cumulative distribution function)** $F = F_X : \mathbb{R} \to \mathbb{R}$, defined by

$$F(x) \quad = \quad P(X \leq x). \tag{3.3}$$

  - if X is discrete, then

$$F(x) \quad = \quad \sum_{x_i \leq x} p_i. \tag{3.4}$$

  - if X is continuous, then

$$F(x) \quad = \quad \int_{-\infty}^{x} f(t)\, dt. \tag{3.5}$$

The pdf has the following properties:

- all values $x_i, i \in I$, are distinct and listed in increasing order;
- all probabilities $p_i > 0, i \in I$ and $f(x) \geq 0$, for all $x \in \mathbb{R}$;
- $\sum_{i \in I} p_i = 1$ and $\int_{\mathbb{R}} f(t)dt = 1$.

The cdf has the following properties:

- if $a < b$ are real numbers, then $P(a < X \leq b) = F(b) - F(a)$;
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$;

- if X is discrete, then $P(X < x) = F(x-0) = \lim\limits_{y \nearrow x} F(y)$ and $P(X = x) = F(x) - F(x-0)$;
- if X is continuous, then $P(X = x) = 0, P(X < x) = P(X \leq x) = F(x)$ and

  $$P(a < X \leq b) = P(a < X \leq b) \ = \ P(a < X < b) \ = \ P(a \leq X \leq b) = \int\limits_a^b f(t)\, dt;$$

- if X is continuous, then $F'(x) = f(x)$, for all $x \in \mathbb{R}$.

3.2 Numerical Characteristics of Random Variables

The **expectation (expected value, mean value)** of a random variable X is a real number $E(X)$ defined by

- if X is a discrete random variable with pdf $\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$,

  $$E(X) = \sum_{i \in I} x_i P(X = x_i) = \sum_{i \in I} x_i p_i, \tag{3.6}$$

  if it exists;
- if X is a continuous random variable with pdf $f : \mathbb{R} \to \mathbb{R}$,

  $$E(X) = \int\limits_{\mathbb{R}} x f(x) dx, \tag{3.7}$$

  if it exists.

The **variance (dispersion)** of a random variable X is the number

$$V(X) = E\Big(X - E(X)\Big)^2, \tag{3.8}$$

if it exists.
The **standard deviation** of a random variable X is the number

$$\sigma(X) = \text{Std}(X) = \sqrt{V(X)}. \tag{3.9}$$

Properties:

- $E(aX + b) = aE(X) + b$, for all $a, b \in \mathbb{R}$;

- $E(X+Y) = E(X) + E(Y)$;
- If X and Y are independent, then $E(X \cdot Y) = E(X)E(Y)$;
- If $X(e) \leq Y(e)$ for all $e \in S$, then $E(X) \leq E(Y)$;
- $V(X) = E(X^2) - E(X)^2$.
- If X and Y are independent, then $V(X+Y) = V(X) + V(Y)$.

Let X be a random variable with cdf $F : \mathbb{R} \to \mathbb{R}$ and $\alpha \in (0,1)$. A **quantile of order α** is a number q_α satisfying the condition $P(X < q_\alpha) \leq \alpha \leq P(X \leq q_\alpha)$, or, equivalently,

$$F(q_\alpha - 0) \leq \alpha \leq F(q_\alpha). \tag{3.10}$$

If X is continuous, then for each $\alpha \in (0,1)$, there is a *unique* quantile q_α , given by $F(q_\alpha) = \alpha$, or equivalently, $q_\alpha = F^{-1}(\alpha)$. It is the number with the property that the area to its left, under the graph of the pdf is equal to α (see Figure 1).

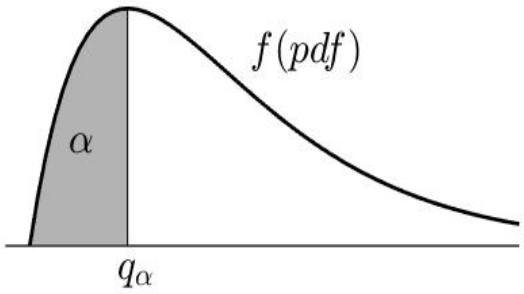


Fig. 1: Quantile q_α

Quantiles are oftenly used in various statistical procedures, such as confidence intervals, rejection regions, etc. (see Figure 2).

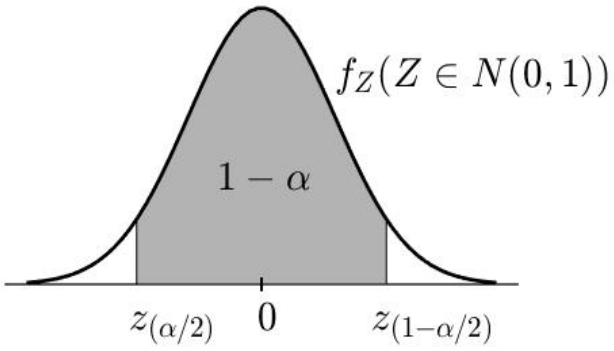


Fig. 2: Quantiles for the Normal distribution

4 Random Vectors

Everything that holds for random *variables* (one-dimensional case) can be easily generalized to any dimension, i.e. to random *vectors*. We restrict our discussion to two-dimensional random vectors $(X, Y) : S \to \mathbb{R}^2$.

Let $(S, \mathcal{K}, P)$ be a probability space. A **random vector** is a function $(X, Y) : S \to \mathbb{R}^2$ satisfying the condition

$$(X \leq x, Y \leq y) = \{e \in S \mid X(e) \leq x, Y(e) \leq y\} \in \mathcal{K},$$

for all $(x, y) \in \mathbb{R}^2$.

- if the set of values that it takes, $(X, Y)(S)$, is at most countable in $\mathbb{R}^2$, then (X, Y) is a **discrete random vector**,
- if $(X, Y)(S)$ is a continuous subset of $\mathbb{R}^2$, then (X, Y) is a **continuous random vector**.
- the function $F : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F(x, y) \;=\; P(X \leq x, Y \leq y)$$

  is called the **joint cumulative distribution function (joint cdf)** of the vector (X, Y).

The properties of the cdf of a random variable translate very naturally for a random vector, as well: Let (X, Y) be a random vector with joint cdf $F : \mathbb{R}^2 \to \mathbb{R}$ and let $F_X, F_Y : \mathbb{R} \to \mathbb{R}$ be the cdf's of X and Y, respectively. Then following properties hold:

- If $a_k < b_k,\ k = \overline{1, 2}$, then

$$\begin{aligned} P(a_1 < X \leq b_1, a_2 < Y \leq b_2) \;&=\; F(b_1, b_2) \;-\; F(b_1, a_2) \\ &-\; F(a_1, b_2) \;+\; F(a_1, a_2). \end{aligned}$$

- $\lim\limits_{x,y\to\infty} F(x, y) = 1,$
  $\lim\limits_{y\to-\infty} F(x, y) = \lim\limits_{x\to-\infty} F(x, y) = 0,\ \forall x, y \in \mathbb{R},$
  $\lim\limits_{y\to\infty} F(x, y) = F_X(x),\ \forall x \in \mathbb{R},$
  $\lim\limits_{x\to\infty} F(x, y) = F_Y(y),\ \forall y \in \mathbb{R}.$

4.1 Discrete Random Vectors

Let $(X, Y) : S \to \mathbb{R}^2$ be a two-dimensional discrete random vector. The **joint probability distribution (function)** of (X, Y) is a two-dimensional array of the form

$X \setminus Y$	y_1	$\cdots$	y_j	$\cdots$	
x_1					
$\vdots$			$\vdots$		
x_i		$\cdots$	p_{ij}	$\cdots$	p_i
$\vdots$			$\vdots$		
			q_j		

(4.1)

where $(x_i, y_j) \in \mathbb{R}^2,\ (i, j) \in I \times J$ are the values that (X, Y) takes and $p_{ij} = P(X = x_i, Y = y_j)$.

An important property is that

$$\sum_{j\in J} p_{ij} = p_i,\ \sum_{i\in I} p_{ij} = q_j \ \text{ and } \ \sum_{i\in I}\sum_{j\in J} p_{ij} = \sum_{j\in J}\sum_{i\in I} p_{ij} = 1,$$

where $p_i = P(X = x_i),\ i \in I$ and $q_j = P(Y = y_j),\ j \in J$. The probabilities p_i and q_j are called **marginal** pdf's.

For discrete random vectors, the computational formula for the cdf is

$$F(x, y) \ = \ \sum_{x_i\le x}\sum_{y_j\le y} p_{ij},\ x, y \in \mathbb{R}.$$

Operations with discrete random variables

Let X and Y be two discrete random variables with pdf's

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I} \quad \text{and} \quad Y \begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}.$$

Sum. The sum of X and Y is the random variable with pdf given by

$$X + Y \begin{pmatrix} x_i + y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}. \tag{4.2}$$

Product. The product of X and Y is the random variable with pdf given by

$$X \cdot Y \begin{pmatrix} x_i y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}. \tag{4.3}$$

Scalar Multiple. The random variable αX, $\alpha \in \mathbb{R}$, with pdf given by

$$\alpha X \begin{pmatrix} \alpha x_i \\ p_i \end{pmatrix}_{i\in I}. \tag{4.4}$$

Quotient. The quotient of X and Y is the random variable with pdf given by

$$X/Y \begin{pmatrix} x_i/y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}, \tag{4.5}$$

provided that $y_j \neq 0$, for all $j \in J$.
In general, if $h : \mathbb{R} \to \mathbb{R}$ is a function, then we can define the random variable $h(X)$, with pdf given by

$$h(X) \begin{pmatrix} h(x_i) \\ p_i \end{pmatrix}_{i\in I}. \tag{4.6}$$

Variables X and Y are said to be **independent** if

$$p_{ij} = P\left(X = x_i, Y = y_j\right) = P\left(X = x_i\right) P\left(Y = y_j\right) = p_i q_j, \tag{4.7}$$

for all $(i, j) \in I \times J$.
If X and Y are independent, then in (4.2), (4.3) and (4.5), $p_{ij} = p_i q_j$, for all $(i, j) \in I \times J$.

4.2 Continuous Random Vectors

Let (X, Y) be a continuous random vector with joint cdf $F : \mathbb{R}^2 \to \mathbb{R}$. Then F is *absolutely continuous*, i.e. there exists a real function $f : \mathbb{R}^2 \to \mathbb{R}$, such that

$$F(x, y) \;=\; \int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\, du\, dv, \tag{4.8}$$

for all $x, y \in \mathbb{R}$. The function f is called the **joint probability density function (joint pdf)** of (X, Y).

The usual properties of continuous pdf's (and their relationship with cdf's) hold for the two-dimensional case, as well: Let (X, Y) be a continuous random vector with joint cdf F and joint density function f. Let $F_X, F_Y : \mathbb{R} \to \mathbb{R}$ be the cdf's of X and Y and $f_X, f_Y : \mathbb{R} \to \mathbb{R}$ be the pdf's of X and Y, respectively. Then the following properties hold:

- $\dfrac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$, for all $(x,y) \in \mathbb{R}^2$.
- $\displaystyle\iint_{\mathbb{R}^2} f(x,y)\, dxdy = 1.$
- for any domain $D \subseteq \mathbb{R}^2$, $P\big((X,Y) \in D\big) = \displaystyle\iint_D f(x,y)\, dxdy.$
- $f_X(x) = \displaystyle\int_{\mathbb{R}} f(x,y)\, dy,\ \forall x \in \mathbb{R}$ and $f_Y(y) = \displaystyle\int_{\mathbb{R}} f(x,y)\, dx,\ \forall y \in \mathbb{R}.$

When obtained from the vector (X, Y), the pdf's f_X and f_Y are called *marginal* densities.
The continuous random variables X and Y are said to be **independent** if

$$f_{(X,Y)}(x,y) \quad = \quad f_X(x) f_Y(y), \tag{4.9}$$

for all $(x, y) \in \mathbb{R}^2$.

5 Common Distributions

5.1 Common Discrete Distributions

Bernoulli Distribution $Bern(p)$

A random variable X has a Bernoulli distribution with parameter $p \in (0, 1)$ $(q = 1 - p)$, if its pdf is

$$X \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}. \tag{5.1}$$

Then

$$\begin{aligned} E(X) &= p, \\ V(X) &= pq. \end{aligned}$$

A Bernoulli r.v. models the occurrence or nonoccurrence of an event.

Discrete Uniform Distribution $U(m)$

A random variable X has a Discrete Uniform distribution (unid) with parameter $m \in \mathbb{N}$, if its pdf is

$$X \begin{pmatrix} k \\ \frac{1}{m} \end{pmatrix}_{k=\overline{1,m}}, \tag{5.2}$$

with mean and variance

$$\begin{aligned} E(X) &= \frac{m+1}{2}, \\ V(X) &= \frac{m^2-1}{12}. \end{aligned}$$

The random variable that denotes the face number shown on a die when it is rolled, has a Discrete Uniform distribution $U(6)$.

Binomial Distribution $B(n,p)$

A random variable X has a Binomial distribution (bino) with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ ($q = 1-p$), if its pdf is

$$X \begin{pmatrix} k \\ C_n^k p^k q^{n-k} \end{pmatrix}_{k=\overline{0,n}}, \tag{5.3}$$

with

$$\begin{aligned} E(X) &= np, \\ V(X) &= npq. \end{aligned}$$

This distribution corresponds to the Binomial model. Given n Bernoulli trials with probability of success p, let X denote the number of successes. Then $X \in B(n,p)$. Also, notice that the Bernoulli distribution is a particular case of the Binomial one, for $n = 1$, $Bern(p) = B(1,p)$.

Geometric Distribution $Geo(p)$

A random variable X has a Geometric distribution (geo) with parameter $p \in (0,1)$ $(q = 1-p)$, if its pdf is given by

$$X \begin{pmatrix} k \\ pq^k \end{pmatrix}_{k=0,1,\ldots}. \tag{5.4}$$

Its cdf, expectation and variance are given by

$$\begin{aligned} F(x) &= 1 - q^{x+1}, x = 0, 1, \ldots \\ E(X) &= \frac{q}{p}, \\ V(X) &= \frac{q}{p^2}. \end{aligned}$$

If X denotes the number of failures that occurred before the occurrence of the 1^{st} success in a Geometric model, then $X \in Geo(p)$.

Remark 5.1. In a Geometric model setup, one might count the number of *trials* needed to get the 1^{st} success. Of course, if X is the number of failures and Y the number of trials, then we simply have $Y = X + 1$ (the number of failures plus the one success). The variable Y is said to have a Shifted Geometric distribution with parameter $p \in (0,1)$ $(Y \in SGeo(p))$. Its pdf is

$$X \begin{pmatrix} k \\ pq^{k-1} \end{pmatrix}_{k=1,2,\ldots} \tag{5.5}$$

and the rest of its characteristics are given by

$$\begin{aligned} F(x) &= 1 - q^{x}, x = 0, 1, \ldots \\ E(X) &= \frac{1}{p}, \\ V(X) &= \frac{q}{p^2}. \end{aligned}$$

In some books, *this* is considered to be a Geometric variable (not in Matlab, though).

Negative Binomial (Pascal) Distribution $NB(n,p)$

A random variable X has a Negative Binomial (Pascal) (nbin) distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ $(q = 1-p)$, if its pdf is

$$X \begin{pmatrix} k \\ C_{n+k-1}^{k} p^n q^k \end{pmatrix}_{k=0,1,\ldots}. \tag{5.6}$$

Then

$$\begin{aligned} E(X) &= \frac{nq}{p}, \\ V(X) &= \frac{nq}{p^2}. \end{aligned}$$

This distribution corresponds to the Negative Binomial model. If X denotes the number of failures that occurred before the occurrence of the n^{th} success in a Negative Binomial model, then $X \in NB(n,p)$. It is a generalization of the Geometric distribution, $Geo(p) = NB(1,p)$.

Poisson Distribution $\mathcal{P}(\lambda)$

A random variable X has a Poisson distribution (poiss) with parameter $\lambda > 0$, if its pdf is

$$X \begin{pmatrix} k \\ \frac{\lambda^k}{k!} e^{-\lambda} \end{pmatrix}_{k=0,1,\ldots} \tag{5.7}$$

with

$$E(X) = V(X) = \lambda.$$

Poisson's distribution is related to the concept of "rare events", or Poissonian events. Essentially, it means that two such events are *extremely unlikely* to occur simultaneously or within a very short period of time. Arrivals of jobs, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, earthquakes are examples of rare events.

A Poisson variable X counts the number of rare events occurring during a fixed time interval. The parameter λ represents the average number of occurrences of the event in that time interval.

Remark 5.2.

1. The sum of n independent $Bern(p)$ random variables is a $B(n,p)$ variable.
2. The sum of n independent $Geo(p)$ random variables is a $NB(n,p)$ variable.

5.2 Common Continuous Distributions

Uniform Distribution $U(a,b)$

A random variable X has a Uniform distribution (unif)with parameters $a, b \in \mathbb{R},\ a < b$, if its pdf is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } x \in [a,b] \\ 0, & \text{if } x \notin [a,b]. \end{cases} \tag{5.8}$$

Then its cdf is

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & \text{if } x \le a \\ \dfrac{x-a}{b-a}, & \text{if } a < x \le b \\ 1, & \text{if } x \ge b \end{cases} \tag{5.9}$$

and its numerical characteristics are

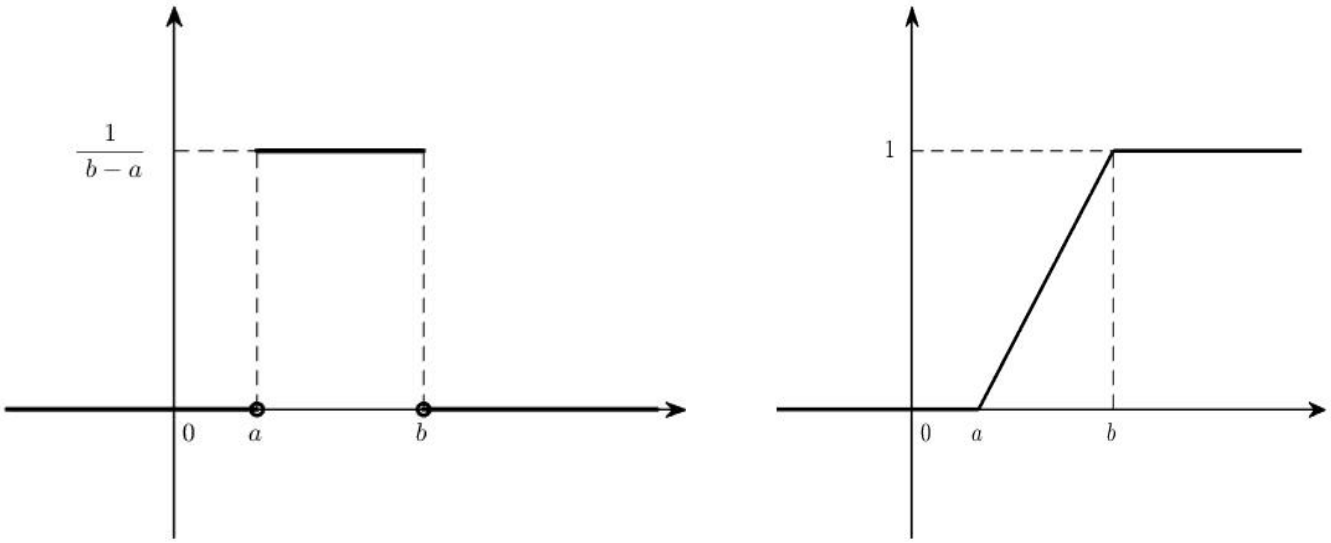
$$\begin{aligned} E(X) &= \frac{a+b}{2}, \\ V(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

The Uniform distribution is used when a variable can take *any* value in a given interval, equally probable. For example, locations of syntax errors in a program, birthdays throughout a year, arrival times of customers, etc.

A special case is that of a **Standard Uniform Distribution**, where $a = 0$ and $b = 1$. The pdf and cdf are given by

$$f_U(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}, \quad F_U(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x \le 1 \\ 1, & x \ge 1\,. \end{cases} \tag{5.10}$$

Standard Uniform variables play an important role in stochastic modeling; in fact, *any* random



(a) Density Function (pdf) (b) Cumulative Distribution Function (cdf)

Fig. 1: Uniform Distribution

variable, with any thinkable distribution (discrete or continuous) can be generated from Standard Uniform variables.

Normal Distribution $N(\mu, \sigma)$

A random variable X has a Normal distribution (norm) with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{5.11}$$

The cdf of a Normal variable is then given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt \tag{5.12}$$

and its mean and variance are

$$\begin{aligned} E(X) &= \mu, \\ V(X) &= \sigma^2. \end{aligned}$$

There is an important particular case of a Normal distribution, namely $N(0,1)$, called the **Standard (or Reduced) Normal Distribution**. A variable having a Standard Normal distribution is usually denoted by Z. The density and cdf of Z are given by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R} \quad \text{and} \quad F_Z(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \tag{5.13}$$

The function F_Z given in (5.13) is known as *Laplace's function* and its values can be found in tables or can be computed by any mathematical software. One can notice that there is a relationship between the cdf of any Normal $N(\mu, \sigma)$ variable X and that of a Standard Normal variable Z, namely,

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right).$$

Exponential Distribution $Exp(\lambda)$

A random variable X has an Exponential distribution (exp) with parameter $\lambda > 0$, if its pdf and cdf are given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{5.14}$$

respectively. Its mean and variance are given by

$$\begin{aligned} E(X) &= \frac{1}{\lambda}, \\ V(X) &= \frac{1}{\lambda^2}. \end{aligned}$$

Remark 5.3.

1. The Exponential distribution is often used to model *time*: lifetime, waiting time, halftime, inter-arrival time, failure time, time between rare events, etc. The parameter λ represents the frequency of rare events, measured in time^{-1}.

2. A word of **caution** here: The parameter μ in Matlab (where the Exponential pdf is defined as $\frac{1}{\mu}e^{-\frac{1}{\mu}x}, x \geq 0$) is actually $\mu = 1/\lambda$. It all comes from the different interpretation of the "frequency". For instance, if the frequency is "2 per hour", then $\lambda = 2/\text{hr}$, but this is equivalent to "one every half an hour", so $\mu = 1/2$ hours. The parameter μ is measured in time units.

3. The Exponential distribution is a special case of a more general distribution, namely the $Gamma(a, b)$, $a, b > 0$, distribution (gam). The Gamma distribution models the *total* time of a multistage scheme, e.g. total compilation time, total downloading time, etc.

4. If $\alpha \in \mathbb{N}$, then the sum of α independent $Exp(\lambda)$ variables has a $Gamma(\alpha, 1/\lambda)$ distribution.

5. In a Poisson process, where X is the number of rare events occurring in time t, $X \in \mathcal{P}(\lambda t)$, the time between rare events and the time of the occurrence of the first rare event have $Exp(\lambda)$ distribution, while T, the time of the occurrence of the α^{th} rare event has $Gamma(\alpha, 1/\lambda)$ distribution.

Gamma-Poisson formula

Let $T \in Gamma(\alpha, 1/\lambda)$ with $\alpha \in \mathbb{N}$ and $\lambda > 0$. Then T represents the time of the occurrence of the α^{th} rare event. Then, the event $(T > t)$ means that the α^{th} event occurs after the moment t. That means that before the time t, fewer than α rare events occur. So, if X is the number of rare events that occur before time t, then the two events

$$(T > t) = (X < \alpha)$$

are equivalent (equal). Now, X has a $\mathcal{P}(\lambda t)$ distribution. So, we have:

$$\begin{aligned} P(T > t) &= P(X < \alpha) \quad \text{and} \\ P(T \leq t) &= P(X \geq \alpha). \end{aligned} \tag{5.15}$$

Remark 5.4. This formula is useful in applications where this setup can be used (seeing a Gamma variable as a sum of times between rare events, if $\alpha \in \mathbb{N}$), as it avoids lengthy computations of Gamma probabilities. However, one should be **careful**, T is a *continuous* random variable, for which $P(T > t) = P(T \geq t)$, whereas X is a discrete one, so on the right-hand sides of (5.15) the inequality signs cannot be changed.

Remark 5.5. The Exponential distributions has the so-called "memoryless property". Suppose that an Exponential variable T represents waiting time. Memoryless property means that the fact of having waited for t minutes gets "forgotten" and it does not affect the future waiting time. Regardless of the event $(T > t)$, when the total waiting time exceeds t, the remaining waiting time still has

Exponential distribution with the same parameter. Mathematically,

$$P(T > t + x | T > t) \quad = \quad P(T > x),\ t, x > 0. \tag{5.16}$$

The Exponential distribution is the only continuous variable with this property. Among discrete ones, the Shifted Geometric distribution also has this property. In fact, there is a close relationship between the two families of variables. In a sense, the Exponential distribution is a continuous analogue of the Shifted Geometric one, one measures time (continuously) until the next rare event, the other measures time (discretely) as the number of trials until the next success.

Chapter 2. Computer Simulations and Monte Carlo Methods

1 Monte Carlo Methods and Random Number Generators

Monte Carlo (MC) methods are methods of approximation (estimation) based on computer simulations involving random numbers. And yes!, the name *does* come from the famous Monte Carlo casino in Monaco (probability distributions involved in gambling are often complicated, but they can be estimated via simulations).

The main purpose of simulations is estimating quantities whose direct computation is complicated, expensive, time consuming, dangerous, or plainly impossible (think space shuttle launch, spread of a virus or disease, performance of a medical device or procedure, etc.). MC methods are mostly used for computation of probabilities, expected values and other distribution characteristics, but they can also be used to estimate lengths, areas, volumes, integrals, irrational numbers (like $\pi, e, \sqrt{2}$), etc.

Fig. 1: Monte Carlo Casino

Recall that the probability can be defined as a *long-run proportion* (relative frequency). With the help of random number generators, computers can actually simulate a "long run". The longer

the run is simulated, the more accurate the results obtained.

Some examples include:

- **Forecasting**. We already know from Statistics that given a distribution model, it is often very difficult to make reasonably *remote predictions*. Often a one-day development depends on the results obtained during *all* the previous days. Then prediction for tomorrow may be straightforward, whereas computation of a one-month forecast is already problematic.

  On the other hand, *simulation* of such a process can be easily performed day by day (or even minute by minute). Based on present results, we simulate the next day and then the next and so on. For a time n, we estimate X_n from the (already known) variables $X_1, X_2, \ldots, X_{n-1}$. Controlling the length of this do-loop, we can obtain forecasts for the next days, weeks, months or years. Such simulations can be used to predict weather, profit, stock prices, costs, etc. Simulation of future failures reflects reliability of devices and systems. Simulation of future stock and commodity prices plays a crucial role in finance, as it allows evaluations of options and other financial deals.

- **Signal transmission (percolation)**. Consider a network of nodes (a graph), some nodes connected, others not. A signal is sent from a certain node. Once a node k receives a signal, it sends it along each of its output lines with some probability p_k. After a certain period of time, one wishes to estimate the proportion of nodes that received a signal, the probability for a certain node to receive it, etc.

  That would mean generating Bernoulli variables with parameters p_k. Line k transmits if the corresponding generated variable $X_k = 1$. In the end, we simply count the number of nodes that got the signal, or verify whether the given node received it.

  This general *percolation* model describes the way many phenomena may spread in real life. The role of a signal may be played by a computer virus spreading from one computer to another, or by rumors spreading among people, or by fire spreading through a forest, or by a disease spreading between residents.

- **Markov chain Monte Carlo**. This is a modern technique of generating random variables from rather complex, often intractable distributions, as long as *conditional distributions* have a reasonably simple form. In semiconductor industry, for example, the joint distribution of good and defective chips on a produced wafer (which is a thin slice of semiconductor) has a rather complicated correlation structure. As a result, it can only be written explicitly for rather simplified artificial models. On the other hand, the quality of each chip is predictable based

on the quality of the surrounding, neighboring chips. Given its neighborhood, conditional probability for a chip to fail can be written, and thus, its quality can be simulated by generating a corresponding Bernoulli random variable with $X_i = 1$ indicating a failure.

According to the *Markov chain Monte Carlo* (MCMC) methodology, a long sequence of random variables is generated from conditional distributions. A wisely designed MCMC will then produce random variables that have the desired *unconditional* distribution, no matter how complex it is.

- **Queuing systems (server facilities)**. A *queuing system* is described by a number of random variables. It involves spontaneous arrivals of jobs, their random waiting time, assignment to servers, their random service and departure time; some jobs may exit prematurely, others may not enter the system if it appears full or busy, also, intensity of the incoming traffic and the number of servers on duty may change during the day. One wants to be able to evaluate a queuing system's vital performance characteristics, such as a job's average waiting time, average length of a queue, the proportion of customers who had to wait, the proportion of "unsatisfied customers" (that exit prematurely or cannot enter), the proportion of jobs spending more than a certain time in the system, the expected usage of each server, the average number of available (idle) servers at the time when a job arrives, and so on.

In all these examples, we saw how different types of phenomena can be computer-simulated. However, *one* simulation is not enough for estimating probabilities and expectations. After we understand how to program the given phenomenon once, we can embed it in a do-loop and repeat similar simulations a large number of times, generating *a long run*. Since the simulated variables are random, we will generally obtain a number of different *realizations*, from which we calculate probabilities and expectations as long-run frequencies and averages.

Implementation of MC methods reduces to generation of random variables from given distributions. All mathematical and statistical software packages (Matlab, Maple, Mathematica, SAS, R, Splus, SPSS, Minitab, Excel, etc.) have built-in procedures for generating random variables from the most common (discrete or continuous) distributions. As it turns out, the main job is that of generating random numbers from a Uniform distribution, in fact, from a *Standard Uniform* distribution. These can then be used to generate random numbers from *any* given distribution.

However, generating *truly random* Uniformly distributed numbers is not an easy task and is an ongoing research area in Modern Statistics and Stochastic Analysis. How do we know that the numbers obtained are "truly random" and do not have any undesired patterns? For example, quality

random number generation is so important in coding and password creation that people design special tests to verify the "randomness" of generated numbers.

More often, *pseudo-random* (deterministic) numbers are generated, i.e. a long list of numbers. The user specifies a random number *seed* that points to the location from which the list will be read. Often each seed is generated within the system, to improve the quality of random numbers.

2 Discrete Methods

These are methods for generating some simple discrete variables, most being specific to certain distributions, by using their interpretation and relationship with other variables, rather than their definition (pdf or cdf).

From here on, we denote by $U \in U(0,1)$ a Standard Uniform variable. Let us recall the pdf and cdf (equation (5.10) in Chapter 1, Lecture 2):

$$f_U(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}, \quad F_U(x) = P(U \le x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x \le 1 \\ 1, & x \ge 1\,. \end{cases} \tag{2.1}$$

Bernoulli Distribution $Bern(p), p \in (0,1)$

Recall that a Bernoulli distribution models the occurrence (or nonoccurrence) of an event (success), with a given probability p.

Let U be a Standard Uniform variable. That means its value is in $(0,1)$, just like the value of p. Then define

$$X \;=\; \begin{cases} 1, & U < p \\ 0, & U \ge p \end{cases}, \tag{2.2}$$

i.e. define "success" as $(U < p)$ (and, obviously, "failure" as $(U \ge p)$). Let us check that this is indeed a Bernoulli variable with parameter p. We have

$$P(X = 1) = P(U < p) = F_U(p) = p,$$

by (2.1), since $p \in (0,1)$. So X has the desired distribution.

Algorithm 2.1.
Read $p \in (0,1)$

$U = rand$;
$X = (U < p)$;

Now we can use this simple way of simulating success/failure with a given probability, for all the variables that are defined in terms of that.

Binomial Distribution $B(n,p), n \in \mathbb{N}, p \in (0,1)$

Recall (Remark 5.2 in Chapter 1, Lecture 2) that a Binomial $B(n,p)$ variable is the sum of n independent $Bern(p)$ variables.

Algorithm 2.2.
Read $n \in \mathbb{N}, p \in (0,1)$
$U = rand(n,1)$;
$X = sum(U < p)$;

Geometric Distribution $Geo(p), p \in (0,1)$

A Geometric variable counts the number of failures that occurred before the 1^{st} success. We can simulate that.

Algorithm 2.3.
Read $p \in (0,1)$
$X = 0$; % initial number of failures
while $rand \geq p$ % while there are failures
  $X = X + 1$; % count number of failures
end % stop at the first success

Remark 2.4. Obviously, the same algorithm can be used to simulate a $Y \in SGeo(p)$ variable, as well. Y is the number of *trials* needed to get the 1^{st} success, so simply let $Y = X + 1$ in Algorithm 2.3.

Negative Binomial (Pascal) Distribution $NB(n,p), n \in \mathbb{N}, p \in (0,1)$

Again, we use the same Remark 5.2 (in Chapter 1, Lecture 2), which states that a $NB(n,p)$ variable is the sum of n independent $Geo(p)$ random variables.

Algorithm 2.5.
Read $n \in \mathbb{N}, p \in (0,1)$

$X = zeros(1, n)$;
for $i = 1 : n$ % generate n independent $Geo(p)$ variables
  while $rand \geq p$ % while there are failures
    $X(i) = X(i) + 1$; % count number of failures
  end % stop at the first success
end
$Y = sum(X)$; % the sum is a $NBin(n, p)$ variable

Arbitrary Discrete Distribution

Let X be an arbitrary discrete random variable with pdf

$$X \begin{pmatrix} x_0 & x_1 & \dots \\ p_0 & p_1 & \dots \end{pmatrix}, \quad \sum p_i = 1. \tag{2.3}$$

We generalize the idea that was used to generate a $Bern(p)$ variable. There, we basically divided the interval $[0, 1]$ into the disjoint subintervals $[0, p)$ and $[p, p + (1 - p)]$. We can do the same for *any* number of subintervals, as seen in Figure 2.

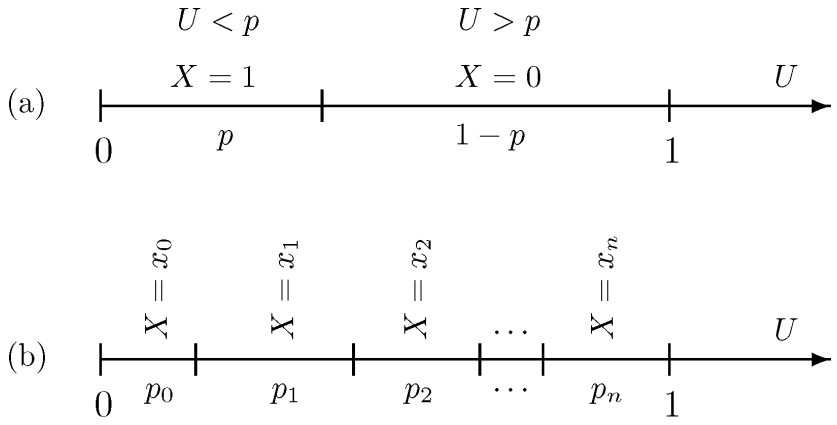


Fig. 2: Generating arbitrary discrete distributions

Algorithm 2.6.

- Read x_i, p_i, $i = 1, 2, \dots$

- Divide the interval $[0, 1]$ into subintervals

$$\begin{aligned} A_0 &= [0, p_0) \\ A_1 &= [p_0, p_0 + p_1) \\ A_2 &= [p_0 + p_1, p_0 + p_1 + p_2) \\ &\dots \\ A_i &= [p_0 + \dots + p_{i-1}, p_0 + \dots + p_i) \\ &\dots \end{aligned}$$

  Then length$(A_i) = p_i$, for a finite or countably infinite number of intervals.

- Get $U \in U(0, 1)$.

- If $U \in A_i$, then let $X = x_i$.

Then, we have

$$\begin{aligned} P(X = x_i) &= P(U \in A_i) = P\Big(p_0 + \dots + p_{i-1} \le U < p_0 + \dots + p_{i-1} + p_i\Big) \\ &= F_U(p_0 + \dots + p_{i-1} + p_i) - F_U(p_0 + \dots + p_{i-1}) \\ &= (p_0 + \dots + p_{i-1} + p_i) - (p_0 + \dots + p_{i-1}) \\ &= p_i, \end{aligned}$$

so X has indeed pdf (2.3).

Example 2.7. Let us use Algorithm 2.6 to generate a variable with pdf

$$X \begin{pmatrix} 0 & 1 & 2 & \dots \\ p_0 & p_1 & p_2 & \dots \end{pmatrix}, \quad \sum_{i \in \mathbb{N}} p_i = 1.$$

Recall that for a discrete random variable, the cdf is computed as $F(x) = P(X \le x) = \sum_{x_i \le x} p_i$, so in this case,

$$F(k) = \sum_{i \le k} p_i = p_0 + p_1 + \dots + p_k, \quad k = 0, 1, \dots,$$

so to check if $U \in A_k$, we check that $F(k-1) \le U < F(k)$.

Algorithm

Read $p_0, p_1, \dots$

Get $U \in U(0,1)$
$i = 0$; % initial value of X
$F = p_0$; % initial value of cdf $F(0)$
while $U \geq F$ % check if $U \in A_i$
$i = i + 1$;
$F = F + p_i$; % new value of cdf, $F(i+1)$
end % the while loop ends when $U < F(i)$
$X = i$;

We can use this to generate a $\mathrm{Poiss}(\lambda)$ variable, with pdf

$$X \begin{pmatrix} i \\ \frac{\lambda^i}{i!} e^{-\lambda} \end{pmatrix}_{i=0,1,\ldots} = \begin{pmatrix} 0 & 1 & 2 & \ldots \\ e^{-\lambda} & \lambda e^{-\lambda} & \frac{\lambda^2}{2} e^{-\lambda} & \ldots \end{pmatrix}, \ \lambda > 0.$$

Algorithm 2.8.
Read $\lambda > 0$
$U = rand$;
$i = 0$;
$F = exp(-\lambda)$;
while $U >= F$
$i = i + 1$;
$F = F + exp(-\lambda) * \lambda\hat{}i/factorial(i)$;
end
$X = i$;

3 Inverse Transform Method

This is a method used when we want to generate a random variable whose cdf F does not have a very complicated form. It is based on the following result:

Theorem 3.1. *Let X be a continuous random variable with cdf $F : \mathbb{R} \to \mathbb{R}$. Then $U = F(X) \in U(0,1)$.*

Proof. So, F is the cdf of X, i.e. $F(x) = P(X \leq x)$, for all $x \in \mathbb{R}$.
We will show that U has the $U(0,1)$ pdf, by starting with its cdf and then taking its derivative.
First off, let us notice that, being a cdf (i.e. a *probability*), $F(x) \in [0,1]$, for all $x \in \mathbb{R}$ and, thus, all

the values of U are in $[0,1]$.

Secondly, X being a continuous random variable, there exists an interval $[a,b] \subseteq \mathbb{R}$ such that :

- $F : [a,b] \to [0,1]$ is strictly increasing (therefore one-to-one),
- $F(x) = 0, \forall x < a$ and
- $F(x) = 1, \forall x > b$.

Hence, its inverse $F^{-1} : [0,1] \to [a,b]$ exists.

Now, let us consider the cdf, F_U. Let $x \in \mathbb{R}$.

If $x < 0$, then $F_U(x) = P(U \leq x) = P(\text{impossible event}) = 0$.

Hence, $f_U(x) = F_U'(x) = 0$.

If $x > 1$, then $F_U(x) = P(U \leq x) = P(\text{sure event}) = 1$ and thus, $f_U(x) = F_U'(x) = 0$.

For $x \in [0,1]$, we have

$$F_U(x) = P(U \leq x) = P(F(X) \leq x) = P\left(X \leq F^{-1}(x)\right) = F\left(F^{-1}(x)\right) = x.$$

Then $f_U(x) = F_U'(x) = 1$ and $U \in U(0,1)$. □

As a consequence, to generate a continuous random variable with given cdf F, we generate a variable $U \in U(0,1)$ and let

$$X = F^{-1}(U). \tag{3.1}$$

Indeed, then the cdf of X is

$$F_X(x) = P(X \leq x) = P\left(F^{-1}(U) \leq x\right) = P(U \leq F(x)) = F_U(F(x)) = F(x),$$

for all $x \in \mathbb{R}$, the last assertion following from (2.1) and the fact that $F(x) \in [0,1]$. Thus X has the desired cdf F.

Example 3.2. Use the ITM to generate a random variable X with pdf

$$f(x) = \frac{1}{2}(x+1),\ x \in [-1,1]. \tag{3.2}$$

Then use the value $U = 0.16$ to generate a value for X.

Solution. First, we find the cdf $F(x) = \int_{-\infty}^{x} f(t)dt$.

If $x < -1$, obviously $F(x) = 0$ (the integrand is 0).

If $x \in [-1, 1]$, we have

$$\begin{aligned} F(x) &= \frac{1}{2}\int_{-1}^{x}(t+1)dt = \frac{1}{2}\left(\frac{1}{2}t^2 + t\right)\Big|_{-1}^{x} \\ &= \frac{1}{2}\left(\frac{1}{2}x^2 + x + \frac{1}{2}\right) = \frac{1}{4}(x+1)^2. \end{aligned}$$

If $x > 1$, then $F(x) = \frac{1}{2}\int_{-1}^{1}(t+1)dt = \int_{\mathbb{R}} f(t)dt = 1$.

So,

$$F(x) = \begin{cases} 0, & x < -1 \\ \frac{1}{4}(x+1)^2, & -1 \le x \le 1 \\ 1, & x > 1 \end{cases}.$$

The graph of the cdf F is shown in Figure 3. One can see that $F : [-1, 1] \to [0, 1]$ is one-to-one and surjective, so the inverse $F^{-1} : [0, 1] \to [-1, 1]$ exists. We find it by solving $F(x) = y$ for x, i.e. $x = F^{-1}(y)$.

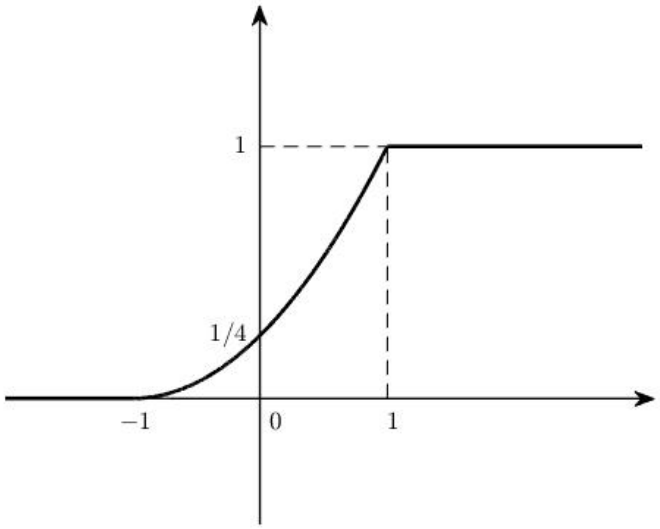


Fig. 3: Function F in Example 3.2

We have

$$\begin{aligned}\frac{1}{4}(x+1)^2 &= y,\\ (x+1)^2 &= 4y,\\ x+1 &= \sqrt{4y},\\ x &= 2\sqrt{y}-1,\end{aligned}$$

so, $F^{-1}(y) = 2\sqrt{y} - 1$, for $y \in [0, 1]$.
Then we generate X from U by

$$X = F^{-1}(U) = 2\sqrt{U} - 1.$$

For the value $U = 0.16$, we get $X = 2 \cdot 0.4 - 1 = -0.2$. ■

Example 3.3. Use the ITM to generate $X \in Exp(\lambda), \lambda > 0$.

Solution. For $X \in Exp(\lambda)$, the pdf and cdf are given by

$$f(x) \;=\; \lambda e^{-\lambda x}, x \geq 0 \text{ and } \; F(x) \;=\; 1 - e^{-\lambda x}, x \geq 0,$$

respectively (see equation (5.14), Chapter 1, Lecture 2).

So, we find the inverse of the cdf

$$\begin{aligned}F(X) &= U,\\ 1 - e^{-\lambda X} &= U,\\ e^{-\lambda X} &= 1 - U,\\ -\lambda X &= \ln(1-U),\\ X_1 &= -\frac{1}{\lambda}\ln(1-U).\end{aligned} \tag{3.3}$$

Now, algebraically, we cannot simplify this expression, but we can notice the following:

$$U \in U(0,1) \Longleftrightarrow 1 - U \in U(0,1).$$

Indeed, we see that for $x \in [0,1]$,

$$\begin{aligned} F_{1-U}(x) &= P(1-U \leq x) = P(U \geq 1-x) = 1 - P(U < 1-x) \\ &= 1 - F_U(1-x) \overset{(2.1)}{=} 1-(1-x) = x = F_U(x). \end{aligned}$$

Then, by taking the derivative, $f_{1-U}(x) = f_U(x)$, so we can replace U by $1-U$ in (3.3) and get

$$X_2 = -\frac{1}{\lambda} \ln(U). \tag{3.4}$$

Notice that since $U, 1-U \in (0,1)$, we have that both $\ln(U), \ln(1-U) < 0$ and, thus, $X_1, X_2 > 0$, as they should be. ■

Discrete Inverse Transform Method

We can see that the previous algorithm seems to have one major fault, namely, that it is not applicable to discrete random variables, since in this case, the cdf F is neither injective, nor surjective and, thus, not invertible. This problem can be overcome, by adjusting the algorithm the following way. In equation (3.1), we will take

$$X = F^{-1}(U) = \min\{x \mid F(x) \geq U\}. \tag{3.5}$$

This is known as the *generalized inverse* function.

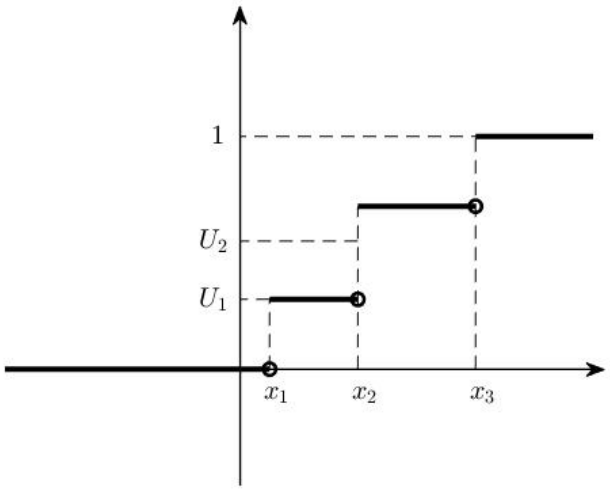


Fig. 4: Generalized inverse

So, in Figure 4, we have

$$\begin{aligned} X_1 &= F^{-1}(U_1) = \min\{x \mid F(x) \geq U_1\} = x_1, \text{ here } F(x_1) = U_1, \\ X_2 &= F^{-1}(U_2) = \min\{x \mid F(x) \geq U_2\} = x_2, \text{ here } F(x_2) > U_2. \end{aligned}$$

Example 3.4. Let us revisit Geometric and Shifted Geometric variables and generate them using the DITM.

Solution. To keep computations simple, we generate a $SGeo(p)$ variable first and then adjust it accordingly to get a $Geo(p)$ variable.
For $X \in SGeo(p), p \in (0,1)$, recall the cdf (Chapter 1, Lecture 2):

$$F(x) = 1 - q^x = 1 - (1-p)^x, \; x \in \mathbb{N}.$$

We use (3.5) to find X:

$$\begin{aligned} 1 - (1-p)^x &\geq U, \\ (1-p)^x &\leq 1 - U, \\ x \ln(1-p) &\leq \ln(1-U), \\ x &\geq \frac{\ln(1-U)}{\ln(1-p)}, \text{ since } \ln(1-p) < 0. \end{aligned}$$

The smallest integer value satisfying this is the *ceiling function* value. Also, as before, $1-U$ can be replaced by U. Thus, a variable $X \in SGeo(p)$ is generated by

$$X = \left\lceil \frac{\ln(U)}{\ln(1-p)} \right\rceil. \tag{3.6}$$

For a $X \in Geo(p)$ variable, with cdf $F(x) = 1 - (1-p)^{x+1}$, the same computations lead to

$$X = \left\lceil \frac{\ln(U)}{\ln(1-p)} - 1 \right\rceil. \tag{3.7}$$

■

Remark 3.5. Notice how similar formula (3.6) is to (3.4), which gives the simulation of an $Exp(\lambda)$ variable. If $\lambda = -\ln(1-p)$, then the generated $SGeo(p)$ variable is just the ceiling of the $Exp(\lambda)$

one. In other words, the ceiling of an Exponential variable has Shifted Geometric distribution. This just shows, again, the strong analogy between the two distributions.

4 Rejection Method

The inverse transform method works well when the cdf F of a random variable does not have a complicated expression and its inverse F^{-1} is relatively easy to find. But when that is not the case, we need other methods with larger applicability. We present next a method that uses the pdf f instead.

Remark 4.1. Before we get started, let us briefly review some properties of Uniformly distributed random variables and random vectors.

1. Recall that for Uniform $U(a,b)$ variables, the pdf is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases} = \begin{cases} \dfrac{1}{\text{length}[a,b]}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}.$$

A vector (X,Y) has a Uniform distribution over a domain $D \subseteq \mathbb{R}^2$ if its (joint) pdf is of the form

$$f(x,y) = \begin{cases} \text{const}, & (x,y) \in D \\ 0, & (x,y) \notin D \end{cases}.$$

What value must that constant be? It should be the value that makes the total integral of f over $\mathbb{R}^2$ equal to 1, as that represents the probability of the sure event. Then

$$1 = \iint\limits_{\mathbb{R}^2} f(x,y)\,dxdy = \iint\limits_{D} \text{const}\,dxdy = \text{const} \iint\limits_{D} dxdy = \text{const} \cdot \text{area}(D),$$

so that constant must be $1/\text{area}(D)$. Thus, a vector (X,Y) has a Uniform distribution over $D \subseteq \mathbb{R}^2$ if its pdf is

$$f(x,y) = \begin{cases} \dfrac{1}{\text{area}(D)}, & (x,y) \in D \\ 0, & (x,y) \notin D \end{cases}. \tag{4.1}$$

2. From the joint pdf of a vector (X,Y), we can always get the *marginal* pdf's of its components by

$$f_X(x) = \int\limits_{\mathbb{R}} f(x,y)\,dy,\ \forall x \in \mathbb{R}, \quad f_Y(y) = \int\limits_{\mathbb{R}} f(x,y)\,dx,\ \forall y \in \mathbb{R}. \tag{4.2}$$

3. If $U \in U(0,1)$, then $X = \alpha + (\beta - \alpha)U \in U(\alpha,\beta),\ \forall \alpha < \beta$.

Theorem 4.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a pdf. Let the vector (X, Y) be Uniformly distributed over the region*

$$D = \{(x, y) \in \mathbb{R}^2 \mid 0 \le y \le f(x)\} \tag{4.3}$$

(see Figure 1). Then X has pdf f, i.e. $f_X = f$.

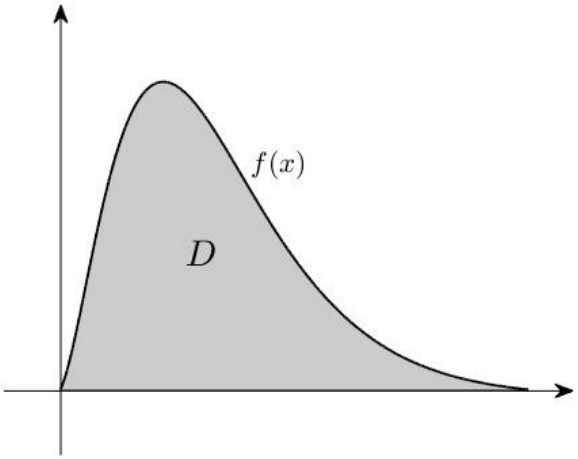


Fig. 1: Domain D

Proof. First, let us determine the joint pdf of the vector (X, Y). By (4.1), it is

$$f_{(X,Y)}(x, y) = \frac{1}{\text{area}(D)}, \text{ for } (x, y) \in D$$

and 0 everywhere else. But, since f is a pdf, that area is $\int_{\mathbb{R}} f(x)\, dx = 1$.

So, the joint pdf of (X, Y) is

$$f_{(X,Y)}(x, y) = \begin{cases} 1, & (x, y) \in D \\ 0, & (x, y) \notin D. \end{cases}$$

Then, using (4.2), we find the (marginal) pdf of its first component X. Fix $x \in \mathbb{R}$. We have

$$f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy = \int_{(x,y)\in D} dy = \int_{y=0}^{y=f(x)} dy = f(x). \tag{4.4}$$

Thus, X indeed has the function f as its pdf. □

So, to generate a variable with given pdf f, we generate points (X, Y) that are Uniformly distributed in D. In order to have $(X, Y) \in D$, we must have $Y \leq f(X)$. If that is not the case, we *reject* the value, hence the name of the method.

Algorithm 4.3.

1. Find numbers $a, b \in \mathbb{R}, c \in \mathbb{R}_+$ such that $f(x) \in [0, c]$ for $x \in [a, b]$ (this is always possible, since D is a bounded region in $\mathbb{R}^2$, having an area of 1). The rectangle $[a, b] \times [0, c]$ is called a *bounding box*.

2. Generate $U, V \in U(0, 1)$.

3. Let $X = a + (b - a)U$ and $Y = cV$. Then $X \in U(a, b)$, $Y \in U(0, c)$ and $(X, Y) \in U([a, b] \times [0, c])$.

4. If $Y > f(X)$, reject the point and return to step 2. If $Y \leq f(X)$, then X has the desired pdf, f.

The idea of the rejection method is displayed graphically in Figure 2.

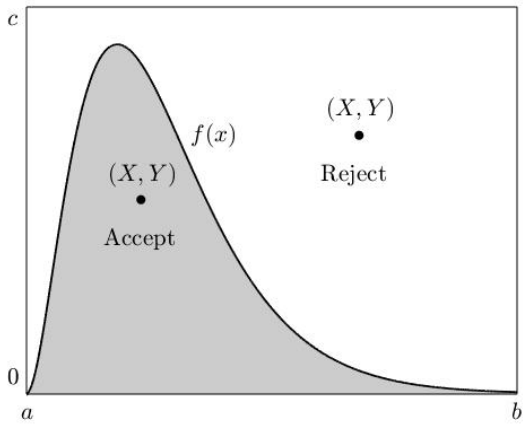


Fig. 2: Rejection Method

Example 4.4. Recall Example 3.2 (Lecture 3). Now let us use the rejection method to generate a random variable X with that same pdf

$$f(x) = \frac{1}{2}(x + 1), \ x \in [-1, 1]. \tag{4.5}$$

Then test it for the values $(U_1, V_1) = (0.12, 0.45)$ and $(U_2, V_2) = (0.91, 0.37)$.

Solution. The graph of f is shown in Figure 3. We can see that a bounding box is $[-1,1] \times [0,1]$. Then by Algorithm 4.3, we find

$$\begin{aligned} X &= 2U - 1 \text{ and} \\ Y &= V. \end{aligned}$$

For (U_1, V_1), we find $(X, Y) = (-0.76, 0.45)$, for which $Y = 0.45 > 0.12 = f(X)$, so this point is *rejected*.
For (U_2, V_2), we have $(X, Y) = (0.82, 0.37)$ and $Y = 0.37 < 0.91 = f(X)$, so this point is *accepted*.
Thus we generated the value 0.82 for the random variable X with pdf (4.5).

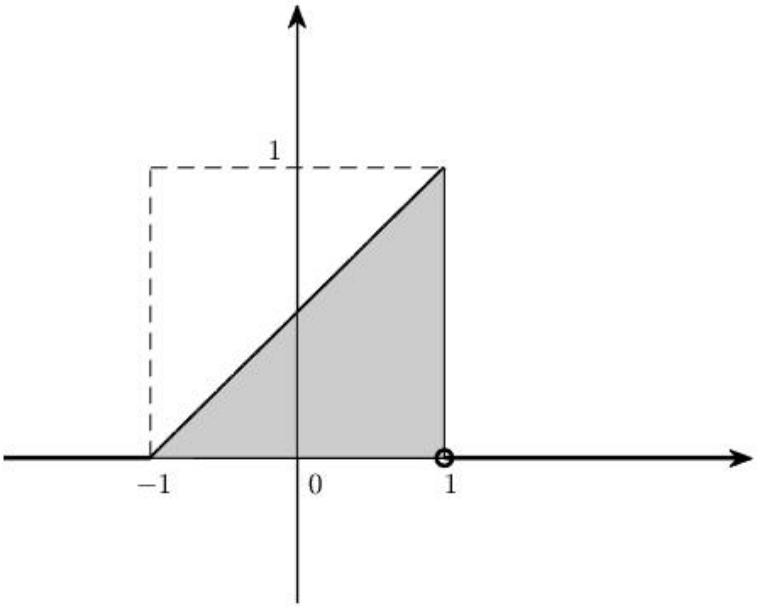


Fig. 3: Function f in Example 4.4

■

Remark 4.5. The rejection method can be used to generate $n-$dimensional random vectors having a desired joint pdf $f : \mathbb{R}^n \to \mathbb{R}$. A bounding box now becomes an $(n+1)-$dimensional cube, where we generate a Uniformly distributed random vector $(X_1, X_2, \ldots, X_n, Y)$, which will be accepted only if $Y \leq f(X_1, X_2, \ldots, X_n)$. Then, the generated vector $(X_1, X_2, \ldots, X_n)$ will have the desired joint density f.

5 Special Methods

These are methods aimed at particular variables, using specific properties of certain distributions. They are a good alternative of simulation, when the more general methods presented so far, are too complicated to implement.

Poisson Distribution $\mathcal{P}(\lambda)$, $\lambda > 0$

Let us recall: In a Poisson process, where X is the number of rare events occurring in time t, X has a Poisson distribution, $X \in \mathcal{P}(\lambda t)$, while the time between rare events and the time of the occurrence of the first rare event have an $Exp(\lambda)$ distribution. Then to generate a $\mathcal{P}(\lambda)$ variable, we count the number of rare events that occur during *one unit* of time ($t = 1$) and generate the Exponential times between events by formula (3.4) (Lecture 3), using the Inverse Transform Method. So, each such time is generated by $T_i = -\frac{1}{\lambda}\ln(U_i)$, for $U_i \in U(0,1)$ and then we count the number of events that occurred in one unit of time:

$$X = \max\{n \mid T_1 + \ldots + T_n \leq 1\}. \tag{5.1}$$

We can simplify the above formula:

$$\begin{aligned} T_1 + \ldots + T_n &= -\frac{1}{\lambda}\ln(U_1) + \ldots - \frac{1}{\lambda}\ln(U_n) \\ &= -\frac{1}{\lambda}\Big(\ln(U_1) + \ldots + \ln(U_n)\Big) \\ &= -\frac{1}{\lambda}\ln\big(U_1 \cdot \ldots \cdot U_n\big). \end{aligned}$$

Then in (5.1) we have, equivalently,

$$\begin{aligned} -\frac{1}{\lambda}\ln\big(U_1 \cdot \ldots \cdot U_n\big) &\leq 1 \\ \ln\big(U_1 \cdot \ldots \cdot U_n\big) &\geq -\lambda \\ U_1 \cdot \ldots \cdot U_n &\geq e^{-\lambda}. \end{aligned}$$

So, a Poisson variable X can be generated by

$$X = \max\{n \mid U_1 \cdot \ldots \cdot U_n \geq e^{-\lambda}\}. \tag{5.2}$$

Algorithm 5.1.

1. Generate $U_1, U_2, \ldots \in U(0,1)$.
2. $X = \max\{n \mid U_1 \cdot U_2 \cdot \ldots \cdot U_n \geq e^{-\lambda}\}$.

Normal Distribution $N(\mu, \sigma),\ \mu \in \mathbb{R}, \sigma > 0$

We present an algorithm called the **Box-Muller transform**, that converts a pair of independent Standard Uniform variables (U, V) into a pair of independent Standard Normal variables (Z_1, Z_2).

Algorithm 5.2.

1. Generate $U, V \in U(0,1)$.
2. Let

$$\begin{aligned} Z_1 &= \sqrt{-2\ln(U)}\cos(2\pi V), \\ Z_2 &= \sqrt{-2\ln(U)}\sin(2\pi V). \end{aligned} \tag{5.3}$$

   Then Z_1, Z_2 are independent $N(0,1)$ random variables.

3. Let $X = \sigma Z + \mu$ (for either Z from above). Then $X \in N(\mu, \sigma)$.

Without going into too many details, this transform is based on the following idea:
For $Z_1, Z_2 \in N(0,1)$, the variable

$$D^2 = Z_1^2 + Z_2^2 \in Exp(1/2),$$

so D^2 can be generated by

$$\begin{aligned} D^2 &= -\frac{1}{1/2}\ln(U) = -2\ln(U) \text{ and} \\ D &= \sqrt{-2\ln(U)}, \end{aligned}$$

for some $U \in U(0,1)$. From here, it is just a matter of getting the two sides of a rectangular triangle from its hypotenuse (see Figure 4). We have

$$\begin{aligned} Z_1 &= D\cos\omega \\ Z_2 &= D\sin\omega, \end{aligned}$$

where the angle ω can take any value in $[0, 2\pi]$ (a complete rotation), i.e. it has a Uniform distribution $U(0, 2\pi)$, so $\omega = 2\pi V$ for some other Standard Uniform variable $V \in U(0, 1)$. Thus, we get (5.3).

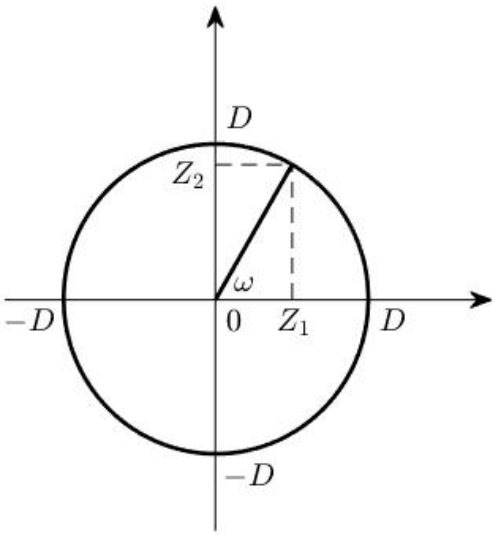


Fig. 4: Box-Muller Transform

6 Accuracy and Size of a Monte Carlo Study

Now, using the methods of simulation presented so far, we perform a Monte Carlo study, meaning that we put the chosen algorithm in a loop and simulate a "long run", i.e. generate a number of such variables, $X_1, \ldots, X_N$.

Recall from Statistics that when a parameter θ is approximated by an estimator (a function of sample variables) $\overline{\theta}$, a desired quality of that estimator is to be *unbiased*, i.e. that

$$E(\overline{\theta}) = \theta, \tag{6.1}$$

so that, in the long run, we know its values will stabilize at the right point. We also want that its variance $V(\overline{\theta})$ be small, approaching 0, as the sample size $N \to \infty$.

Estimating Probabilities, Means and Variances

We estimate probabilities by long run relative frequencies. For a random variable X, we generate variables $X_1, \ldots, X_N$ with the same distribution and approximate $p = P(X \in A)$ by

$$\overline{p} = \frac{\text{number of } X_1, \ldots, X_N \in A}{N}. \tag{6.2}$$

The mean value $E(X) = \mu$, the variance $V(X) = \sigma^2$ and the standard deviation $\sigma = \sqrt{V(X)}$ of a random variable X are estimated by

$$\begin{aligned} \overline{X} &= \frac{X_1 + \ldots + X_N}{N}, \\ s^2 &= \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2, \; s = \sqrt{s^2}, \end{aligned} \tag{6.3}$$

respectively. Since the simulations are independent, the number at the numerator in (6.2) has Binomial $B(N, p)$ distribution (the number of successes in N trials) and, hence, expected value Np and variance $Np(1-p)$. Then, we have

$$\begin{aligned} E(\overline{p}) &= \frac{1}{N} Np \;=\; p, \\ V(\overline{p}) &= \frac{1}{N^2} Np(1-p) \;=\; \frac{p(1-p)}{N}. \end{aligned} \tag{6.4}$$

Thus, $\overline{p}$ is an unbiased estimator for p and its standard deviation $\sigma(\overline{p}) = \sqrt{\dfrac{p(1-p)}{N}}$ decreases with N at the rate of $1/\sqrt{N}$.
The same is true for the estimators in (6.3), but we omit the details.

Accuracy of a Monte Carlo Study

When we conduct a Monte Carlo study, the question arises about its size. What would be a suitable size in order to get a certain accuracy? Given a tolerable error $\varepsilon > 0$ and a significance level (probability of error) $\alpha \in (0, 1)$, we want to determine the size N so that

$$P\left(|\overline{p} - p| > \varepsilon\right) \leq \alpha. \tag{6.5}$$

It is known that for moderate values of p ($0.05 \leq p \leq 0.95$) and large values of n, a Binomial variable can be approximated by a Normal one:

$$B(n,p) \approx N\left(\mu = np,\ \sigma = \sqrt{np(1-p)}\right) \tag{6.6}$$

Also, recall that for a Normal variable $X \in N(\mu, \sigma)$, its *reduced* variable $\dfrac{X - E(X)}{\sigma(X)} = \dfrac{X-\mu}{\sigma}$ has a Standard Normal $N(0,1)$ distribution.

Then for the variable $N\overline{p}$ for large values of N, we use the Normal approximation of the Binomial distribution of $N\overline{p}$, to get

$$\frac{N\overline{p} - E(N\overline{p})}{\sqrt{V(N\overline{p})}} \;=\; \frac{N(\overline{p}-p)}{N\sqrt{\dfrac{p(1-p)}{N}}} \;=\; \frac{\overline{p}-p}{\sqrt{\dfrac{p(1-p)}{N}}} \;\approx\; N(0,1).$$

We can use this to estimate the probability in (6.5). We have

$$P\big(|\overline{p}-p| > \varepsilon\big) \;=\; P\left(\frac{|\overline{p}-p|}{\sqrt{\dfrac{p(1-p)}{N}}} > \frac{\varepsilon}{\sqrt{\dfrac{p(1-p)}{N}}}\right) = 2\Phi\left(-\frac{\varepsilon\sqrt{N}}{\sqrt{p(1-p)}}\right),$$

where Φ is Laplace's function (the cdf of a $N(0,1)$ variable) described in equation (5.13) (Lecture 2).

Still, this contains the unknown value p. We can manage that using the fact that for any $p \in (0,1)$,

$$\begin{aligned} p(1-p) &\leq \frac{1}{4}, \\ \sqrt{p(1-p)} &\leq \frac{1}{2}, \\ \frac{1}{\sqrt{p(1-p)}} &\geq 2, \\ -\frac{1}{\sqrt{p(1-p)}} &\leq -2, \end{aligned}$$

so,

$$-\frac{\varepsilon\sqrt{N}}{\sqrt{p(1-p)}} \;\leq\; -2\varepsilon\sqrt{N}.$$

Since Φ is an increasing function,

$$\Phi\left(-\frac{\varepsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \leq \Phi\big(-2\varepsilon\sqrt{N}\big).$$

Then to ensure (6.5), we take $\Phi\big(-2\varepsilon\sqrt{N}\big) \leq \alpha/2$, or, equivalently, $-2\varepsilon\sqrt{N} \leq \Phi^{-1}(\alpha/2) = z_{\alpha/2}$, i.e.

$$N \geq \frac{1}{4}\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2, \tag{6.7}$$

where $z_{\alpha/2}$ is the quantile (inverse of the cdf Φ) of order $\alpha/2$ for the $N(0,1)$ distribution.

7 Other Applications of Monte Carlo Methods

Estimating lengths, areas, volumes

Let $A \subseteq [0,1]$. How to estimate length(A)? Recall that a variable $U \in U(0,1)$ has pdf $f_U(x) = 1,\ x \in [0,1]$. Then

$$P(U \in A) \;=\; \int_A f_U(x)\,dx \;=\; \int_A dx \;=\; \text{length}(A). \tag{7.1}$$

But MC methods can be used to estimate the probability on the left-hand side. So, we generate $U_1, \ldots, U_N \in U(0,1)$, compute the proportion of U_i that lie in A and estimate the length of A by that proportion.
What if $A \nsubseteq [0,1]$, but in some other interval $A \subseteq [a,b]$? Then for a variable $X \in U(a,b)$, with pdf $f(x) = 1/(b-a), x \in [a,b]$,

$$P(X \in A) \;=\; \int_A f(x)\,dx \;=\; \frac{1}{b-a}\int_A dx \;=\; \frac{1}{b-a}\text{length}(A). \tag{7.2}$$

Then we generate $X_1, \ldots, X_N \in U(a,b)$ and estimate the length of A by $(b-a)P(X \in A)$.

For estimating areas, we do exactly the same things, but with double integrals.
Let $A \subseteq [0,1] \times [0,1]$ and $U, V \in U(0,1)$. That means $(U,V) \in U\big([0,1] \times [0,1]\big)$, so its joint pdf is $f_{(U,V)}(x,y) = 1,\ (x,y) \in [0,1] \times [0,1]$. Then

$$P\big((U,V) \in A\big) \;=\; \iint_A f_{(U,V)}(x,y)\,dxdy \;=\; \iint_A dxdy \;=\; \text{area}(A). \tag{7.3}$$

Again, if $A \subseteq [a,b] \times [c,d]$ and $(X,Y) \in U\big([a,b] \times [c,d]\big)$, then

$$P\big((X,Y) \in A\big) \;=\; \iint\limits_A f_{(X,Y)}(x,y)\,dxdy \;=\; \frac{1}{(b-a)(d-c)}\text{area}(A) \tag{7.4}$$

and the area of A can be approximated by $(b-a)(d-c)P\big((X,Y) \in A\big)$.

Algorithm 7.1.

1. Generate $X_i \in U(a,b), Y_i \in U(c,d), i = 1,\ldots,N$.
2. Compute the number of pairs (X_i, Y_i) that belong to A, say N_A.
3. Estimate $\text{area}(A) \approx (b-a)(d-c)\dfrac{N_A}{N}$.

Example 7.2. Approximate π by MC methods.

Solution. The number π is the area of the unit disk $x^2+y^2 \le 1$. Cover the unit disk by the rectangle $[-1,1] \times [-1,1]$, i.e. find a bounding box (see Figure 5).

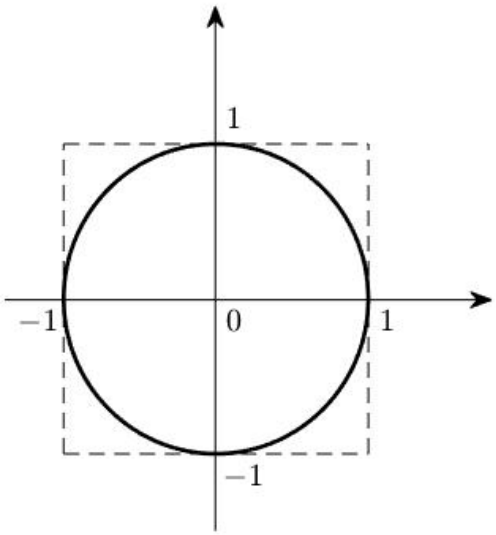


Fig. 5: Bounding box for the unit disk

Apply Algorithm 7.1:

1. Generate $X_1,\ldots,X_N, Y_1,\ldots,Y_N \in U(-1,1)$.
2. Compute the number of pairs (X_i, Y_i) for which $X_i^2 + Y_i^2 \le 1$, say N_π.

3. Approximate $\pi \approx 4\dfrac{N_\pi}{N}$.

■

Remark 7.3. Notice that estimation of areas (or lengths, volumes) by MC methods **does not require** knowing the *exact* boundaries of the region to be estimated. All that is necessary is covering that region by a bounding box (rectangle) and generating points Uniformly distributed in that rectangle.

Example 7.4. An emergency is reported at a nuclear power plant and it is necessary to assess the size of the region exposed to radioactivity. Boundaries of the region cannot be determined, but it is known that it is covered by a rectangle of 15 by 20 km and the level of radioactivity can be measured at any given location. Suppose that 100 measurements are taken at random points in that rectangular area and radioactivity is found above normal in 43 locations. Estimate the area of the exposed region A.

Solution. We have a bounding box $[0, 15] \times [0, 20]$ and the 100 locations represent 100 Uniform variables $(X_i, Y_i) \in U\Big([0, 15] \times [0, 20]\Big)$. Of those, 43 pairs are found to belong to the region A. Thus, we estimate

$$\text{area}(A) \approx 15 \cdot 20 \cdot \frac{43}{100} = 129 \text{ km}^2.$$

■

Monte Carlo integration

Recall that the definite integral of a nonnegative function represents the area of the region underneath the graph of that function. Then MC methods can be used to approximate definite integrals by estimating areas below or above the graphs of corresponding functions.

Suppose we want to approximate the integral

$$I \;=\; \int_a^b g(x)\, dx,$$

for some function $g : [a, b] \to [0, c]$. Then we cover the area that is I by $[a, b] \times [0, c]$ and estimate it with the rejection method.

Algorithm 7.5.

1. Generate $U_i, V_i \in U(0, 1), i = 1, \ldots, N$.

2. Let $X_i = a + (b-a)U_i$ and $Y_i = cV_i, i = 1, \ldots, N$ (or generate directly $X_i \in U(a,b)$ and $Y_i \in U(0,c)$).

3. Compute the number of pairs (X_i, Y_i) for which $Y_i \leq g(X_i)$, say N_I.

4. Estimate the integral by
$$I \approx \overline{I} = (b-a)\, c\, \frac{N_I}{N}.$$

Finally, for the general case $g : [a,b] \to [c,d]$, where g can take both positive and negative values, take each subinterval separately and on those subintervals where $g(x) \leq 0$, consider $|g(x)|$. Then the areas above the $x-$axis are added and those below are subtracted (see Figure 6). Since the estimation of the integral still uses (long-run) proportions, the accuracy of the approximation is the same, i.e.

$$\begin{aligned} E(\overline{I}) &= I \text{ (it is unbiased) and} \\ \sigma(\overline{I}) &= \sqrt{\frac{I(1-I)}{N}}. \end{aligned}$$

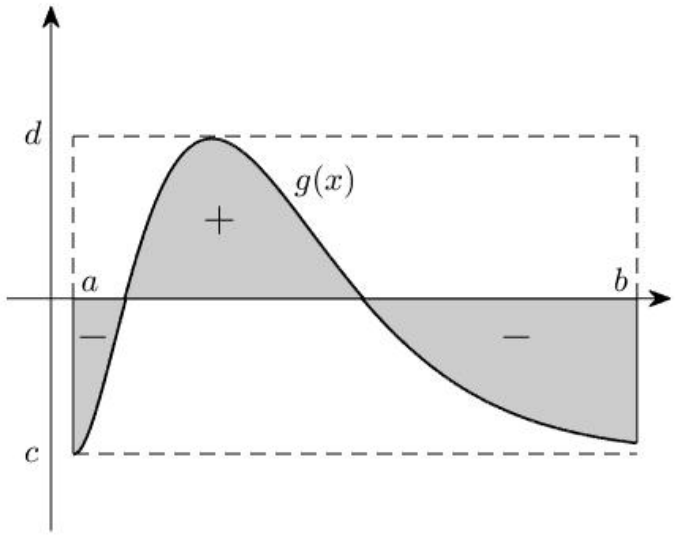


Fig. 6: MC Integration

Chapter 3. Stochastic Processes

So far, when discussing random variables, random vectors and their distributions, we described the situation at a particular moment of time, as if someone had said "Freeze!" and everything stood still. But the real world is dynamic and many random variables develop and change in time (think stock prices, air temperatures, interest rates, football scores, CPU usage, the speed of internet connection, popularity of politicians, and so on).

Basically, stochastic processes are random variables that *evolve and change in time*.

1 Basic Notions

Definition 1.1. *A **stochastic process** is a random variable that also depends on time. It is denoted by $X(t, e)$ or $X_t(e)$, where $t \in \mathcal{T}$ is time and $e \in S$ is an outcome. The values of $X(t, e)$ are called **states**.*

If $t \in \mathcal{T}$ is fixed, then X_t is a random variable, whereas if we fix $e \in S$, X_e is a function of time, called a **realization** or **sample path** or **trajectory** of the process $X(t, e)$.

Definition 1.2. *A stochastic process is called **discrete-state** if $X_t(e)$ is a discrete random variable, for all $t \in \mathcal{T}$ and **continuous-state** if $X_t(e)$ is a continuous random variable, for all $t \in \mathcal{T}$.*
*Similarly, a stochastic process is said to be **discrete-time** if the set $\mathcal{T}$ is discrete and **continuous-time** if the set of times $\mathcal{T}$ is a (possibly unbounded) interval in $\mathbb{R}$.*

Example 1.3.
1. Available memory, CPU usage, in percents, is a continuous-state, continuous-time process.
2. The CPU usage *per hour* is continuous-state, discrete-time.
3. In a printer shop, $X_n(e)$, the amount of time required to print the n^{th} job, is a discrete-time, continuous-state stochastic process, because $n = 1, 2, \ldots$ and $X \in (0, \infty)$.
4. On the other hand, $Y_n(e)$, the number of pages of the n^{th} printing job, is discrete-time and discrete-state. In this case, $Y = 1, 2, \ldots$, which is a discrete set.
5. The *actual* air temperature $X_t(e)$ at time t is a continuous-time, continuous-state stochastic process. Indeed, it changes smoothly and never jumps from one value to another.
6. However, $Y_t(e)$, the temperature reported *every hour* on radio or TV, is a discrete-time process. Moreover, since the reported temperature is usually rounded to the nearest degree, it is also a discrete-state process.

Throughout the rest of the course, we will omit writing e as an argument of a stochastic process (as it is customary when writing random variables).

2 Markov Processes and Markov Chains

2.1 Transition Probability Matrix

Definition 2.1. *A stochastic process X_t is **Markov** if for any times $t_1 < t_2 < \ldots < t_n < t$ and any sets $A_1, A_2, \ldots, A_n; A$,*

$$P(X_t \in A \mid X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = P(X_t \in A \mid X_{t_n} \in A_n). \tag{2.1}$$

What this means is that the conditional distribution of X_t given observations of the process *at several moments in the past*, is the same as the one given *only the latest* observation. In other words, knowing the *present*, we get no information from the *past* that can be used to predict the *future*:

$$P(\text{future} \mid \text{past},\ \text{present}) = P(\text{future} \mid \text{present}).$$

Then, for the future development of a Markov process, only its present state is important, and it does not matter *how* the process arrived to this state.

Some processes satisfy the Markov property, others don't.

Example 2.2.
1. Let X_t be the total number of internet users registered by some internet service provider by the time t. If, say, there were 999 users connected by 10 o'clock, then their total number will be or exceed 1000 during the next hour *regardless* of when and how those 999 users connected to the internet in the past. The number of connections in an hour will only depend on the current number. This process *is* Markov.
2. Let Y_t be the value of some stock or some market index at time t. If we know $Y(t)$, do we also want to know $Y(t-1)$ in order to predict $Y(t+1)$? One may argue that if $Y(t-1) < Y(t)$, then the market is rising, therefore, $Y(t+1)$ is likely (but not certain) to exceed $Y(t)$. On the other hand, if $Y(t-1) > Y(t)$, we may conclude that the market is falling and may expect $Y(t+1) < Y(t)$. It looks like knowing the past in addition to the present did help us to predict the future. In this case, to make predictions about the future, we need a history (so the past, too, not just the present). Then, this process is *not* Markov.

Due to a well-developed theory and a number of simple techniques available for Markov processes, it is important to know whether a stochastic process is Markov or not.

Remark 2.3. The idea of Markov dependence was proposed and developed by Andrey A. Markov $(1856 - 1922)$ who was a student of P. L. Chebyshev at St. Petersburg University (Russia).

Definition 2.4. *A discrete-state, discrete-time Markov stochastic process is called a* ***Markov chain****.*

To simplify the writing, we use the following notations: Since a Markov chain is a discrete-time process, we can consider the time set as $\mathcal{T} = \{0, 1, 2, \dots\}$ and the Markov chain as a sequence of random variables

$$\{X_0, X_1, \dots\},$$

where X_k describes the situation at time $t = k$.

It is also a discrete-state process, so we can denote the states by $1, 2, \dots, n$. Sometimes we will start enumeration from state 0, and sometimes we might deal with a Markov chain with infinitely many (discrete) states, then we will have $n = \infty$.
Then the random variable X_k has the pdf

$$X_k \begin{pmatrix} 1 & 2 & \dots & n \\ P_k(1) & P_k(2) & \dots & P_k(n) \end{pmatrix}, \tag{2.2}$$

where

$$\begin{aligned} P_k(1) &= P(X_k = 1), \\ P_k(2) &= P(X_k = 2), \\ &\dots, \\ P_k(n) &= P(X_k = n). \end{aligned}$$

Since the states (the values of the random variable X_k) are the same for each k, one only needs the second row to describe the pdf. Then let

$$P_k = [P_k(1) \;\; P_k(2) \;\; \dots \;\; P_k(n)] \tag{2.3}$$

denote the vector on the second row of the pdf (2.2). Obviously,

$$\sum_{i=1}^{n} P_k(i) = 1.$$

So, in short, we can write the pdf of X_k as

$$X_k \begin{pmatrix} 1 & \dots & n \\ & P_k & \end{pmatrix}.$$

The Markov property (2.1) means that in predicting the value of X_{t+1}, i.e. in which state j it is and with what probability $P_{t+1}(j)$, only the value i of X_t matters. So (2.1) can now be written as

$$P(X_{t+1} = j \mid X_t = i, X_{t-1} = l, \dots) \quad = \quad P(X_{t+1} = j \mid X_t = i), \text{ for all } t \in \mathcal{T}. \tag{2.4}$$

We summarize this information in a matrix.

Definition 2.5.

- *The conditional probability*

$$p_{ij}(t) \quad = \quad P(X_{t+1} = j \mid X_t = i) \tag{2.5}$$

is called a ***transition probability****; it is the probability that the Markov chain transitions from state i to state j, at time t. The matrix*

$$P(t) \quad = \quad [p_{ij}(t)]_{i,j=\overline{1,n}} \tag{2.6}$$

is called the ***transition probability matrix*** *at time t.*

- *Similarly, the conditional probability*

$$p_{ij}^{(h)}(t) \quad = \quad P(X_{t+h} = j \mid X_t = i) \tag{2.7}$$

*is called an h-**step transition probability**, i.e. the probability that the Markov chain moves from state i to state j in h steps, and the matrix*

$$P^{(h)}(t) \quad = \quad \left[p_{ij}^{(h)}(t)\right]_{i,j=\overline{1,n}} \tag{2.8}$$

*is the h-**step transition probability matrix** at time t.*

Definition 2.6. *A Markov chain is* ***homogeneous (or stationary)*** *if all transition probabilities are*

independent of time,

$$\begin{aligned} p_{ij}(t) &= p_{ij}, \\ P(t) &= P = [p_{ij}]_{i,j=\overline{1,n}}, \\ p_{ij}^{(h)}(t) &= p_{ij}^{(h)}, \\ P^{(h)}(t) &= P^{(h)} = \left[p_{ij}^{(h)}\right]_{i,j=\overline{1,n}}. \end{aligned}$$

Being homogeneous means that transition from i to j has the same probability *at any time*.

By the Markov property, each next state can be predicted from the previous state *only*.
So, when working with Markov chains, we will need to know:

- X_0, its initial situation, i.e. the distribution of its initial state, P_0;
- the mechanism of transitions from one state to another, i.e. the matrix P.

Based on this, we want to find:

- h-step transition probabilities $p_{ij}^{(h)}$ and $P^{(h)}$;
- the distribution of states at time h, X_h, i.e. P_h, which will be our forecast;
- possibly the limit of $P^{(h)}$ and P_h as $h \to \infty$, i.e. a *long-term* forecast; as we will see, when making forecasts for *many* transitions ahead, computations will become rather lengthy, and thus, it will be more efficient to take the limit.

In order to better understand the ideas and the computations, let us start with a simple example and then discuss the general formulas.

Example 2.7. In Rainbow City, each day is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7, while a rainy day is followed by a sunny day with probability 0.4. Suppose it rains on Monday. Make forecasts for Tuesday.

Solution. This process has two states, $1 =$ "sunny" and $2 =$ "rainy", so it is **discrete-state**. The time set $\{$Monday, Tuesday, $\ldots\}$ is also discrete, so it is **discrete-time**.
Since the weather forecast for each day depends *only* on the weather the previous day, it is a **Markov** process and, hence, a **Markov chain**.
Finally, since transition probabilities are the same for *any* two consecutive times (days), it is also **homogeneous**.

Thus, X_k, the weather situation on day k, is a homogeneous Markov chain with 2 states.
The initial situation (on Monday) is

$$X_0 \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad P_0(1) = 0, \; P_0(2) = 1, \quad P_0 = [0 \quad 1].$$

The transition probability matrix is

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

This can also be seen in a *transition diagram* (Figure 1). Arrows represent all possible one-step transitions, along with the corresponding probabilities.

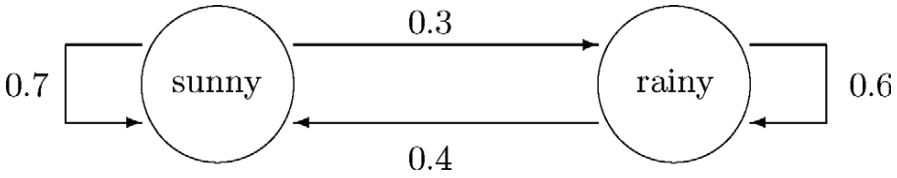


Fig. 1: Transition diagram for Example 2.7

Now, what is the prognosis for Tuesday ($t = 1$)? Since it rains on Monday, we only need to look at the second row in matrix P, the transition probabilities from state 2. Then the forecast for Tuesday is "sunny" with probability $p_{21} = 0.4$ (making a transition from a rainy to a sunny day) and "rainy" with probability $p_{22} = 0.6$. So for X_1, we have

$$X_1 \begin{pmatrix} 1 & 2 \\ 0.4 & 0.6 \end{pmatrix}, \quad P_1(1) = 0.4, \; P_1(2) = 0.6, \quad P_1 = [0.4 \quad 0.6].$$

■

Now, before we go any further with our forecast, we need a little review. Recall the Total Probability Rule (Theorem 1.4j, in Lecture 1):

$$P(E) \quad = \quad \sum_{i \in I} P\left(E|A_i\right) P\left(A_i\right),$$

for any partition $\{A_i\}_{i \in I}$.

The same formula holds for a *conditional* probability, i.e.

$$P(E|B) \;=\; \sum_{i\in I} P\left(E|A_i\right) P\left(A_i|B\right), \tag{2.9}$$

if $\{A_i\}_{i\in I}$ is a partition of S and $P(B)\neq 0$.

Example 2.8. Assuming the same situation as before, make forecasts for Wednesday and Thursday.

Solution. To make forecasts for Wednesday, we need the 2-step transition probability matrix $P^{(2)}$, making one transition from Monday to Tuesday, X_0 to X_1, and another one from Tuesday to Wednesday, X_1 to X_2. We'll have to *condition* on the weather situation on Tuesday and use formula (2.9). Notice that the events $\Big\{\{\text{Tuesday is sunny}\}, \{\text{Tuesday is rainy}\}\Big\}$ form a partition. That is, $\{(X_1=1),(X_1=2)\}$ form a partition.

So, let us proceed:

$$\begin{aligned}
p_{21}^{(2)} &= P\left(\text{Wednesday is sunny} \mid \text{Monday is rainy}\right)\\
&= P(X_2=1|X_0=2)\\
&= P(X_2=1|X_1=1)P(X_1=1|X_0=2)\\
&+ P(X_2=1|X_1=2)P(X_1=2|X_0=2)\\
&= p_{11}\cdot p_{21}+p_{21}\cdot p_{22}\\
&= 0.7\cdot 0.4+0.4\cdot 0.6 \;=\; 0.52.
\end{aligned}$$

Obviously,

$$\begin{aligned}
p_{22}^{(2)} &= P\left(\text{Wednesday is rainy} \mid \text{Monday is rainy}\right)\\
&= 1-P\left(\text{Wednesday is sunny} \mid \text{Monday is rainy}\right)\\
&= 1-p_{21}^{(2)} \;=\; 0.48.
\end{aligned}$$

Thus, we have the second row of $P^{(2)}$, which is *all* we need to know in order to make forecasts for Wednesday:

$$X_2\begin{pmatrix} 1 & 2\\ 0.52 & 0.48\end{pmatrix},\;\; P_2(1)=0.52,\; P_2(2)=0.48,\;\; P_2=[0.52\quad 0.48].$$

So, for Wednesday there is 52% chance of sun and 48% chance of rain.

For the Thursday forecast, we need to compute 3-step transition probabilities $p_{ij}^{(3)}$, because it takes *three* transitions to move from Monday to Thursday. We have to use the Total Probability Rule conditioning on *both* Tuesday and Wednesday. This corresponds to a sequence of states

$$2 \ \to \ i \ \to \ j \ \to \ 1.$$

Luckily, we have already computed the 2-step transition probabilities $p_{21}^{(2)}$ and $p_{22}^{(2)}$, describing transition from Monday to Wednesday. It remains to add *one* transition to Thursday. Thus,

$$\begin{aligned} p_{21}^{(3)} &= p_{21}^{(2)} \cdot p_{11} + p_{22}^{(2)} \cdot p_{21} \\ &= 0.52 \cdot 0.7 + 0.48 \cdot 0.4 \ = \ 0.556 \end{aligned}$$

and then,

$$p_{22}^{(3)} \ = \ 1 - p_{21}^{(3)} \ = \ 0.444.$$

So, for Thursday, we predict a 55.6% chance of sun and a 44.4% chance of rain.

■

Remark 2.9. Obviously, more remote forecasts require more lengthy computations. For a t-day ahead forecast, we have to account for *all* t-step paths on diagram Figure 1. Or, we use the of Total Probability Rule, conditioning on *all* the intermediate states $X_1, X_2, \ldots, X_{t-1}$. To simplify the task, we will employ matrices.

Recall multiplication of matrices. For two $n \times n$ matrices, $A = [a_{ij}]_{i,j=\overline{1,n}}$, $B = [b_{ij}]_{i,j=\overline{1,n}}$, the product is computed by

$$[A \cdot B]_{ij} \ = \ \underbrace{[a_{i1} \ \ldots \ a_{in}]}_{i^{th} \text{ row of } A} \cdot \underbrace{\begin{bmatrix} b_{1j} \\ \ldots \\ b_{nj} \end{bmatrix}}_{j^{th} \text{ col. of } B} \ = \ \sum_{k=1}^{n} a_{ik} \cdot b_{kj}.$$

Let us notice that

$$\underline{P_0 \cdot P} \ = \ [0 \ \ 1] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \ = \ [0.4 \ \ 0.6] \ = \ \underline{P_1}. \tag{2.10}$$

Now, to get back to our task, from all the previous computations, let us notice that

$$\underline{P_0 \cdot P^{(2)}} \;=\; [0 \quad 1] \begin{bmatrix} \cdots & \cdots \\ 0.52 & 0.48 \end{bmatrix} \;=\; [0.52 \quad 0.48] \;=\; \underline{P_2}. \tag{2.11}$$

Even though it wasn't necessary for the Wednesday forecast, let us still compute the first row of $P^{(2)}$, in order to draw some conclusions. We proceed in a similar way (but write fewer details). We have

$$\begin{aligned} p_{11}^{(2)} &= P(X_2 = 1 | X_0 = 1) \\ &= P(X_2 = 1 | X_1 = 1) P(X_1 = 1 | X_0 = 1) \\ &+ P(X_2 = 1 | X_1 = 2) P(X_1 = 2 | X_0 = 1) \\ &= p_{11} \cdot p_{11} + p_{21} \cdot p_{12} \\ &= (0.7)^2 + 0.3 \cdot 0.4 \;=\; 0.61 \end{aligned}$$

and, of course,

$$p_{12}^{(2)} \;=\; 1 - p_{11}^{(2)} \;=\; 0.39.$$

So, we notice that

$$\begin{aligned} p_{11}^{(2)} &= p_{11} \cdot p_{11} + p_{21} \cdot p_{12} \\ &= [p_{11} \quad p_{12}] \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}, \\ p_{21}^{(2)} &= p_{11} \cdot p_{21} + p_{21} \cdot p_{22} \\ &= [p_{21} \quad p_{22}] \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}. \end{aligned}$$

If we had computed the other two probabilities *directly*, we would have found that

$$\begin{aligned} p_{12}^{(2)} &= p_{11} \cdot p_{12} + p_{12} \cdot p_{22} \\ &= [p_{11} \quad p_{12}] \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix} \text{ and} \\ p_{22}^{(2)} &= p_{21} \cdot p_{12} + p_{22} \cdot p_{22} \\ &= [p_{21} \quad p_{22}] \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix}. \end{aligned}$$

So, in fact, we see that

$$P^{(2)} = \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} = P^2,$$

the *second power* of P.
Also, from (2.10) and (2.11), we notice that

$$P_0 \cdot P^{(i)} = P_i, \quad i = 1, 2.$$

Now, we can state the general result.

Proposition 2.10 (Chapman-Kolmogorov). *Let $\{X_0, X_1, \dots\}$ be a Markov chain. Then the following relations hold:*

$$P^{(h)} = P^h \ (= \underbrace{P \cdot P \cdot \ldots \cdot P}_{h \text{ times}}), \text{ for all } h = 1, 2, \dots \tag{2.12}$$

$$P_i = P_0 \cdot P^{(i)} = P_0 \cdot P^i, \text{ for all } i = 0, 1, \dots \tag{2.13}$$

Proof.
The proof of (2.12) goes by induction.
Obviously, relation (2.12) is true for $h = 1$. Assume $P^{(h-1)} = P^{h-1}$.

For a matrix M, we use the notation $[M]_{ij} = M(i, j)$ and, similarly, for a vector v, $(v)_i = v(i)$.

Since the events $\{(X_{h-1} = k)\}_{k=\overline{1,n}}$ form a partition, using the Total Probability Rule (2.9) with $E = (X_h = j), B = (X_0 = i), A_k = (X_{h-1} = k), k = \overline{1, n}$, for $[P^{(h)}]_{ij} = p_{ij}^{(h)}$ (the (i, j)-entry in matrix $P^{(h)}$), we have

$$\begin{aligned} p_{ij}^{(h)} &= P(X_h = j \mid X_0 = i) \\ &= \sum_{k=1}^{n} \underbrace{P(X_h = j \mid X_{h-1} = k)}_{p_{kj}} \cdot \underbrace{P(X_{h-1} = k \mid X_0 = i)}_{p_{ik}^{(h-1)}} \\ &= \sum_{k=1}^{n} p_{ik}^{(h-1)} \cdot p_{kj} = \left[P^{(h-1)} \cdot P\right]_{ij} \\ &\overset{\text{ind. hyp.}}{=} \left[P^{h-1} \cdot P\right]_{ij}, \text{ for all } i, j = \overline{1, n}, \end{aligned}$$

so

$$P^{(h)} = P^h.$$

To prove the second relation (2.13), for each $j = \overline{1,n}$, we have $[P_i]_j = P_i(j) = P(X_i = j)$. Again, using the Total Probability Rule for the partition $\{(X_0 = k)\}_{k=\overline{1,n}}$, with $E = (X_i = j)$ and $A_k = (X_0 = k)$), we get for $[P_i]_j$

$$\begin{aligned} P(X_i = j) &= \sum_{k=1}^{n} \underbrace{P(X_i = j \mid X_0 = k)}_{p_{kj}^{(i)}} \cdot \underbrace{P(X_0 = k)}_{[P_0]_k} \\ &= \sum_{k=1}^{n} [P_0]_k \cdot p_{kj}^{(i)} \\ &= \left[P_0 \cdot P^{(i)}\right]_j, \end{aligned}$$

so, by the previous relation proved, (2.12), we obtain

$$P_i = P_0 \cdot P^i.$$

□

Example 2.11. Assume the same situation as before, except for Monday the forecast is 80% chance of rain. Make forecasts for Wednesday and Friday.

Solution. What is different from the previous situation? The transition probability matrices P and $P^{(h)} = P^h$ are the same. What changes is the *initial* situation. Now, a sunny Monday (state 1) is *also possible* and the pdf of X_0 is

$$X_0 \begin{pmatrix} 1 & 2 \\ 0.2 & 0.8 \end{pmatrix}, \quad P_0 = [0.2 \quad 0.8].$$

So, for Wednesday ($t = 2$), we have

$$P_2 = P_0 \cdot P^{(2)} = P_0 \cdot P^2 = [0.2 \quad 0.8] \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} = [0.538 \quad 0.462],$$

that means 53.8% chance of sun and 46.2% chance of rain.

For Friday, four days after Monday (so, at $t = 4$), we have

$$P_4 = P_0 \cdot P^{(4)} = P_0 \cdot P^4 = [0.2 \quad 0.8] \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix} = [0.5684 \quad 0.4316],$$

i.e. 56.84% chance of sun and 43.16% chance of rain.

■

Remark 2.12. Notice that in matrices P and $P^{(h)}(= P^h)$, the sum of all the probabilities on each row is 1. That is because from each state, a Markov chain makes a transition to *one and only one* state, i.e. state destinations are mutually exclusive and exhaustive events, thus forming a partition. Such matrices are called **stochastic**. **Caution!** In general, this property does not hold for column totals. Some states may be "more favorable" than others, then they are visited more often than others, thus their column total will be larger. In our weather example, that is the case for the state "sunny".

2.2 Simulation of Markov Chains

Many important characteristics of stochastic processes require lengthy complex computations.
Thus, it is preferable to estimate them by means of Monte Carlo methods.

For Markov chains, to predict its future behavior, all that is required is the distribution of X_0, i.e. P_0 (the initial situation) and the pattern of change at each step, i.e. the transition probability matrix P.

Once X_0 is generated, it takes some value $X_0 = i$ (according to its pdf P_0). Then, at the next step, X_1 is a discrete random variable taking the values $j, j = 1, \ldots, n$ with probabilities p_{ij} from row i of the matrix P. Its pdf will be

$$X_1 \begin{pmatrix} 1 & 2 & \ldots & n \\ p_{i1} & p_{i2} & \ldots & p_{in} \end{pmatrix}$$

The next steps are simulated similarly.
Since, at each step, the generation of a discrete random variable is needed, we can use the algorithm that simulates an arbitrary discrete distribution, Algorithm 2.6 in Lecture 3.

Algorithm 2.13.
1. Given:

$N_M =$ sample path size (length of Markov chain),

$P_0 = [P_0(1) \ \ldots \ P_0(n)]$,
$P = [p_{ij}]_{i,j=\overline{1,n}}$.

2. Generate X_0 from its pdf P_0.

3. Transition: if $X_t = i$, generate X_{t+1}, with probabilities $p_{ij}, j = \overline{1, n}$ (i.e. the i^{th} row of P), using Algorithm 2.6 (L3).

4. Return to step 3 until a Markov chain of length N_M is generated.

2.3 Steady-State Distribution; Regular Markov Chains

It is sometimes necessary to be able to make *long-term* forecasts, meaning we want

$$\lim_{h\to\infty} P_h,$$

so we need to compute $\lim_{h\to\infty} p_{ij}^{(h)}$.

Definition 2.13. *Let X be a Markov chain. The vector $\pi = [\pi_1, \ldots, \pi_n]$, consisting of the limiting probabilities*

$$\pi_k = \lim_{h\to\infty} P_h(k), k = 1, \ldots, n, \tag{2.13}$$

if it exists, is called a ***steady-state (stationary, limiting) distribution*** *of X.*

When this limit exists, it can be used as a forecast of the distribution of X after *many* transitions.

In order to find it, let us notice that

$$P_h P \;=\; (P_0 P^h)P \;=\; P_0 P^{h+1} \;=\; P_{h+1}.$$

Taking the limit as $h \to \infty$ on both sides, we get

$$\pi P \;=\; \pi. \tag{2.14}$$

System (2.14) is an $n \times n$ *singular* linear system (multiplication by a constant on each side leads to infinitely many solutions). However, since π must also be a *stochastic* matrix, the sum of its components must equal 1. Thus, we add one more condition,

$$\pi_1 + \pi_2 + \ldots \;=\; 1,$$

called the *normalizing* equation. If a solution of system (2.14) *exists*, then this extra condition will make it *unique*.

We state the following result, without proof.

Proposition 2.14. *The steady-state distribution of a homogeneous Markov chain X, $\pi = [\pi_1, \ldots, \pi_n]$, if it exists, is unique and is the solution of the $(n+1) \times n$ linear system*

$$\begin{cases} \pi P & = & \pi \\ \sum\limits_k \pi_k & = & 1. \end{cases} \tag{2.15}$$

Example 2.15. Let us find the steady-state distribution of the Markov chain in Example 2.10 (Lecture 5). What is the weather forecast in Rainbow City for Christmas Day next year?

Solution. Recall that in Example 2.10 we had a homogeneous Markov chain with two states, (1-sunny, 2-rainy), the initial situation (on Monday) was 80% chance of rain, i.e.

$$P_0 = [0.2 \;\; 0.8]$$

and the transition probability matrix was

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

We write system (2.14). We have

$$[\pi_1 \;\; \pi_2]P = [\pi_1 \;\; \pi_2] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.7\pi_1 + 0.4\pi_2 \\ 0.3\pi_1 + 0.6\pi_2 \end{bmatrix},$$

so system (2.14) becomes

$$\begin{cases} 0.7\pi_1 + 0.4\pi_2 &= \pi_1 \\ 0.3\pi_1 + 0.6\pi_2 &= \pi_2 \end{cases} \iff 0.3\pi_1 - 0.4\pi_2 = 0 \iff \pi_2 = \frac{3}{4}\pi_1.$$

We see that two equations in the system reduced to one. This will *always* happen, i.e., one equation will follow from the others, and this is because the system $\pi P = \pi$ is *singular*. It remains to use the normalizing equation, to get

$$\begin{cases} 3\pi_1 - 4\pi_2 &= 0 \\ \pi_1 + \pi_2 &= 1, \end{cases}$$

with solution

$$[\pi_1 \;\; \pi_2] = [4/7 \;\; 3/7].$$

Interpretation: in the long-run, in the future, $4/7 \approx 57\%$ of days are sunny and $3/7 \approx 43\%$ of days are rainy. Recall that the forecast for Wednesday was $53.8\%/46.2\%$ and for Friday, $56.84\%/43.16\%$, which is already getting close to the steady-state distribution.

Since Christmas Day next year is *many* steps from now, we use the steady-state distribution instead. So that would be the forecast for Christmas Day next year, too!

■

Remark 2.16.
1. Just as we did in the previous example, when we need to make predictions after a large number of steps, instead of the lengthy computation of P_h (i.e., P^h), it may be easier to try to find the steady-state distribution, π, directly.
2. If a steady-state distribution exists, then it can be shown that the matrix $P^{(h)} = P^h$ also has a limit, as $h \to \infty$, and the limiting matrix is given by

$$\mathbf{\Pi} = \lim_{h\to\infty} P^{(h)} = \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 & \dots & \pi_n \\ \vdots & \vdots & \dots & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_n \end{bmatrix}.$$

3. Notice that π and $\mathbf{\Pi}$ *do not* depend on the initial state X_0. Actually, in the long run, the probabilities of transitioning from any state to a given state are the same, $p_{ik} = p_{jk},\ \forall i, j, k = \overline{1, n}$ (all the rows of $\mathbf{\Pi}$ coincide). Then, it is just a matter of "reaching" a certain state (from anywhere), rather than "transitioning" to it (from another state). That should, indeed, depend only on the pattern of changes, i.e. only on the transition probability matrix.
4. What is actually the "steady" state of a Markov chain? Suppose the system has reached its steady state, so that the current distribution of states is

$$P_t = \pi.$$

Then the system makes one more transition, and the distribution becomes

$$P_{t+1} = \pi P.$$

But $\pi P = \pi$ and thus,

$$P_t = P_{t+1}.$$

We see that in a steady state, transitions do not affect the distribution. A system may go from one state to another, but the *distribution* (the pdf) of states *does not change*. In this sense, it is *steady*.

Now, a natural question arises: does a steady-state distribution always exist? The answer is **no**! Here is a simple example:

Example 2.17. In a game of chess, a knight (în rom. "calul") can only move to a field of different color (white-to-black or black-to-white) at any time. Then the transition probability matrix of the

color of its field is

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

For this matrix, a simple computation yields

$$P^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I,$$

so,

$$\begin{aligned} P^{2k} &= I \text{ and} \\ P^{2k+1} &= P, \forall k \in \mathbb{N}. \end{aligned}$$

These are the *only* possible values. Thus,

$$\lim_{h \to \infty} P^h$$

does not exist and neither does

$$\lim_{h \to \infty} P_h.$$

This is a **periodic** Markov chain with period 2,

$$X_t = X_{t+2}.$$

Periodic Markov chains *do not* have a steady-state distribution.

There are other situations when steady-state probabilities cannot be found. So, when *does* a steady-state distribution exist? This is an ongoing research problem. We mention (without proof) one case, which is really easy to check, when such a distribution does exist.

Definition 2.18. *A Markov chain is called **regular** if there exists $h \geq 0$, such that*

$$p_{ij}^{(h)} > 0, \tag{2.16}$$

for all $i, j = 1, \ldots, n$.

This is saying that at some step h, $P^{(h)}$ has *only* non-zero entries. meaning that h-step transitions

from any state to any state are possible.

Proposition 2.19. *Any regular Markov chain has a steady-state distribution.*

Remark 2.20. Regularity of Markov chains does not mean that *all* $p_{ij}^{(h)}$ should be positive, for *all* h. The transition probability matrix P, or some of its powers, may have some 0 entries, but there must exist some power h, for which $P^{(h)}$ has all non-zero entries.

Example 2.21. The Markov chain in Example 2.15 is regular because all transitions are possible for $h = 1$ already, and matrix P does not contain any zeros. Indeed, it has a steady-state distribution.

Example 2.22. A Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.9 & 0 & 0 & 0.1 \end{bmatrix}$$

is also regular. Matrix P contains zeros and so do P^2, P^3, P^4 and P^5. However, the 6-step transition probability matrix

$$P^{(6)} = \begin{bmatrix} .009 & .090 & .900 & .001 \\ .001 & .009 & .090 & .900 \\ .810 & .001 & .009 & .180 \\ .162 & .810 & .001 & .027 \end{bmatrix}$$

contains no zeros and shows regularity of this Markov chain.

In fact, computation of all P^h up to $h = 6$ is not even required in this case. Regularity can also be seen from the transition diagram in Figure 1. We can see that any state $j = 1, 2, 3, 4$ can be reached in 6 steps from any state $i = 1, 2, 3, 4$. Indeed, moving counterclockwise through this figure, we can reach state 4 from any state i in at most 3 steps. Then, we can reach any state j from state 4 again in at most 3 additional steps, for the total of at most 6 steps. If we can reach a state i from a state j in *fewer* than 6 steps, we just use the remaining steps circling around state 4. For example, state 2 is reached from state 1 in 6 steps as follows:

$$1 \to 2 \to 3 \to 4 \to 4 \to 1 \to 2.$$

Then, indeed all $p_{ij}^{(6)}$ are positive and the chain is regular. This goes to show that we don't have to *actually compute* all $p_{ij}^{(h)}$. We only need to verify that they are all positive for some h.

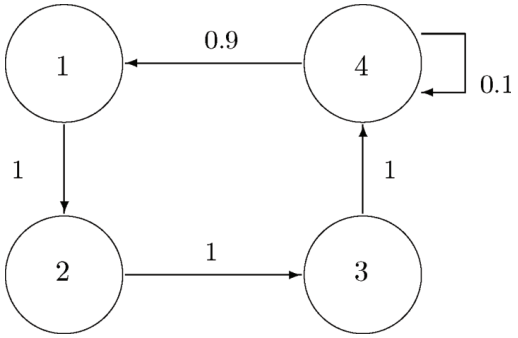


Fig. 1: Transition diagram for the regular Markov chain in Example 2.22

Absorbing states

If there exists a state i with $p_{ii} = 1$, then that Markov chain *cannot* be regular. There is no exit (no transition possible) from state i. Such a state is called an **absorbing state**. For example, state 4 in Figure 2(a) is absorbing, therefore, the Markov chain is irregular.

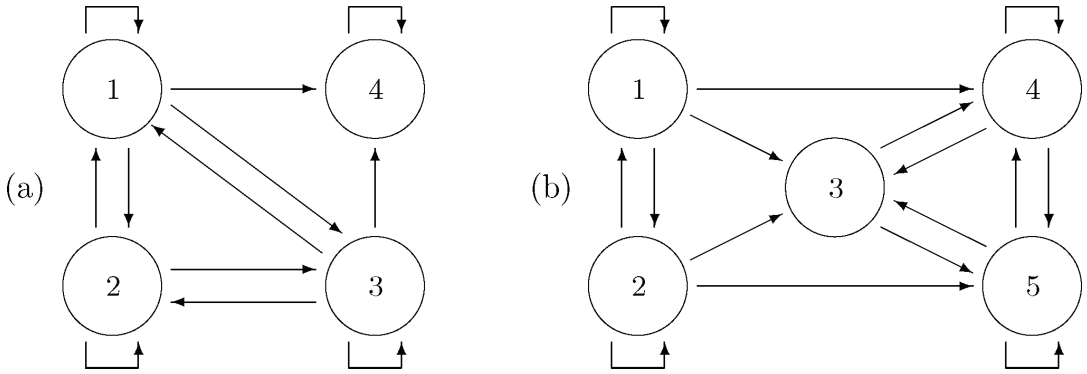


Fig. 2: Absorbing states and absorbing zones

There may be several absorbing states or an entire *absorbing zone*, from which the remaining states can never be reached. For example, states $3, 4$ and 5 in Figure 2(b) form an absorbing zone, some kind of a "Bermuda triangle". When this process finds itself in the set $\{3, 4, 5\}$, there is no route from there to the set $\{1, 2\}$. As a result, e.g. probability $p_{31}^{(h)}$ is 0 for *all* h.

However, notice that both Markov chains *do* have steady-state distributions. The first process will eventually reach state 4 and will stay there for good. Therefore, the limiting distribution of X_h is

$$\pi \;=\; \lim_{h\to\infty} P_h \;=\; [0\;\;0\;\;0\;\;1].$$

The second Markov chain will eventually leave states 1 and 2 for good, thus its limiting (steady-state) distribution has the form

$$\pi \;=\; [0\;\;0\;\;\pi_3\;\;\pi_4\;\;\pi_5].$$

This goes to show that the converse of Proposition 2.19 is *not true*, there are irregular Markov chains that have a steady-state distribution.

Remark 2.23. The study of Markov chains gives us an important method of analyzing rather complicated stochastic systems. Once the Markov property of a process is established, it only remains to find its one-step transition probabilities. Then, the steady-state distribution can be computed, and thus, we obtain the distribution of the process *at any time*, after a sufficient number of transitions. This methodology will be our main working tool in the next chapter, when we study queuing systems and evaluate their performance.

3 Counting Processes

A special case of stochastic processes are the ones where one needs to count the occurrences of some types of events over time. These are described by *counting processes*.

Definition 3.1. *A **counting process** $X(t), t \geq 0$, is a stochastic process that represents the number of items counted by the time t.*

Counting processes deal with the number of occurrences of something over time, such as customers arriving at a supermarket, completed tasks, transmitted messages, detected errors, scored goals, number of job arrivals to a queue, holding times (in renewal processes), etc.

In general, we refer to the occurrence of each event that is being counted as an "arrival". As time passes, one can count additional items. Therefore, sample paths (values) of a counting process are always *non-decreasing, non-negative integers* $\{0, 1, \ldots\}$.

Thus, **all** counting processes are **discrete-state** stochastic processes. They can be discrete-time or continuous-time.

Example 3.2. Figure 3 shows sample paths of two counting processes, $X(t)$ being the number of transmitted e-mails by the time t and $A(t)$ being the number of transmitted attachments. According

to the graphs, e-mails were transmitted at times $t = 8, 22, 30, 32, 35, 40, 41, 50, 52$ and 57 min. The e-mail counting process $X(t)$ increments by 1 at each of these times. Only 3 of these e-mails contained attachments. One attachment was sent at time $t = 8$, five more at $t = 35$, making the total of $A(35) = 6$, and two more attachments at $t = 50$, making the total of $A(50) = 8$.

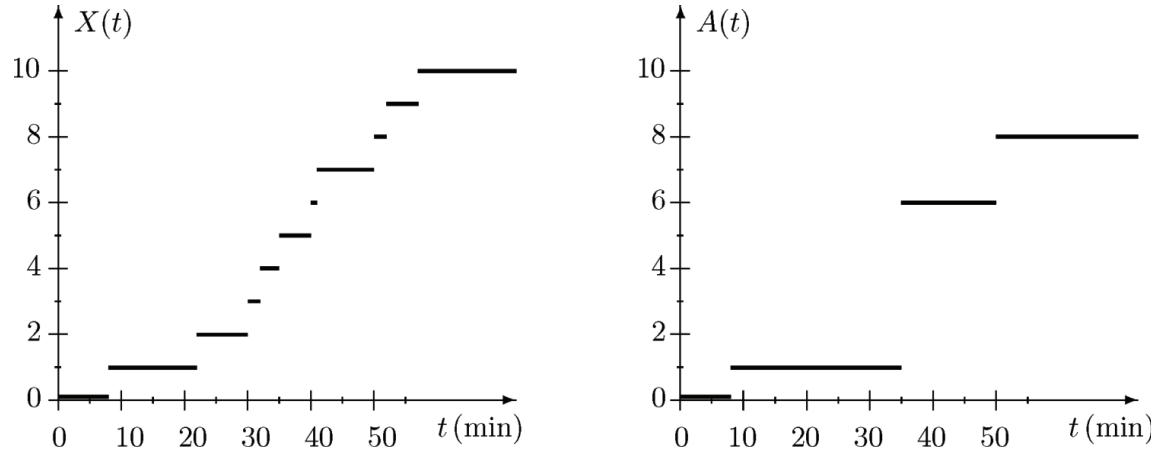


Fig. 3: Counting processes in Example 3.2

Next, we consider the most widely used examples, Binomial (discrete-time) and Poisson (continuous-time) counting processes.

3.1 Binomial Counting Process

Consider a sequence of Bernoulli trials with probability of success p and count the number of "successes".

Definition 3.3. *A **Binomial counting process** $X(n)$ is the number of successes in n Bernoulli trials, $n = 0, 1, \ldots$.*

Remark 3.4.

1. Obviously, a Binomial process $X(n)$ is a discrete-state, discrete-time stochastic process, "time" being measured discreetly, by the number of trials, n.
2. The pdf of $X(n)$ is Binomial $B(n, p)$ at any time n (see Figure 4). Recall that

$$E(X(n)) = np.$$

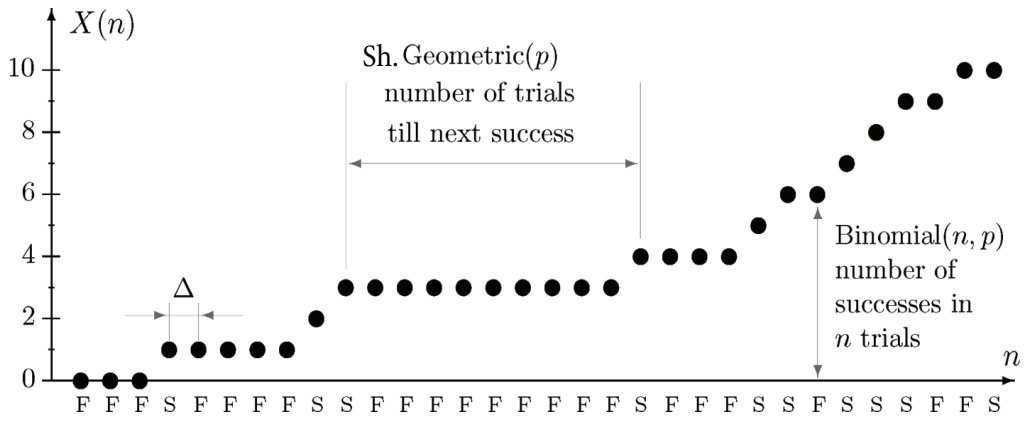


Fig. 4: Binomial process sample path (S = success, F = failure)

3. The number of trials between two consecutive successes, Y, is the *number of trials needed to get the next (first) success*, so it has a $SGeo(p)$ pdf (see Figure 4). Recall that

$$E(Y) = \frac{1}{p},\ V(Y) = \frac{q}{p^2}.$$

Relation to real time, frames

It is important to make the distinction between real time and the "time" variable n ("time" as in a stochastic process). Variable n is not measured in time units, it measures the number of trials.

Suppose that Bernoulli trials occur at equal time intervals, say every Δ seconds (see Figure 4). That means that n trials occur during time $t = n\Delta$. The value of the process at time t has Binomial pdf with parameters $n = \frac{t}{\Delta}$ and p. Then the expected number of successes *during* t *seconds* is

$$E(X(n)) \ = \ E\left(X\left(\frac{t}{\Delta}\right)\right) \ = \ np \ = \ \frac{t}{\Delta}p \ = \ t\frac{p}{\Delta},$$

so the expected number of successes *per second* is

$$\lambda \ = \ \frac{p}{\Delta}.$$

Definition 3.5.

- *The quantity* $\lambda = \dfrac{p}{\Delta}$ *is called the* ***arrival rate****, i.e. the average number of successes per one unit of time.*
- *The quantity* Δ *is called a* ***frame****, i.e the time interval of each Bernoulli trial.*
- *The* ***interarrival time*** *is the time between successes.*

We can now rephrase:

– p is the probability of *arrival* (success) during one *frame* (trial),

– $n = \dfrac{t}{\Delta}$ is the number of *frames* during time t,

– $X\left(\dfrac{t}{\Delta}\right)$ is the number of *arrivals* by time t.

The concepts of *arrival rate* and *interarrival time* deal with modeling arrival of jobs in discrete-time queuing systems by Binomial counting processes. The key assumption in such models is that no more than 1 arrival can occur during each Δ-second frame (otherwise, a smaller Δ should be considered), so each frame is a Bernoulli trial.

The interarrival period, Y, measured in number of frames, has a $SGeo(p)$ pdf (as mentioned earlier). Since each frame takes Δ seconds, the interarrival time is

$$T = \Delta Y,$$

a *rescaled* $SGeo(p)$ variable, whose expected value and variance are given by

$$\begin{aligned} E(T) &= \Delta E(Y) = \Delta\frac{1}{p} = \frac{1}{\lambda}, \\ V(T) &= \Delta^2 V(Y) = \Delta^2\frac{q}{p^2} = \frac{1-p}{\lambda^2}. \end{aligned} \tag{3.1}$$

Example 3.6. Messages arrive at a communications center at the rate of 6 messages per minute. Assume arrivals of messages are modeled by a Binomial counting process.
a) What frame size should be used to guarantee that the probability of a message arriving during each frame is 0.1?
b) Using the chosen frames, find the probability of no messages arriving during the next 1 minute.

c) Compute the probability of more than 35 messages arriving during the next 6 minutes.

d) Find the probability of more than 350 messages arriving during the next hour.

e) What is the average interarrival time and its standard deviation?

f) Compute the probability that the next message does not arrive during the next 20 seconds.

Solution.

a) We have $\lambda = 6$ / min. and $p = 0.1$. Thus,

$$\Delta \;=\; \frac{p}{\lambda} \;=\; \frac{1}{60}\,\text{min.} \;=\; 1\,\text{sec.}$$

b) So $\Delta = 1$ sec. In $t = 1$ minute $= 60$ seconds, there are $n = \dfrac{t}{\Delta} = 60$ frames. The number of messages arriving during 1 minute (i.e. 60 frames), $X(60)$, has a Binomial distribution with parameters $n = 60$ and $p = 0.1$. So the desired probability is

$$\begin{aligned} P\big(X(60) = 0\big) &= \text{pdf}_{X(60)}(0) \\ &= \text{binopdf}(0, 60, 0.1) \\ &= 0.0018. \end{aligned}$$

c) Similarly, in $t = 6$ minutes $= 360$ seconds, there are $n = \dfrac{t}{\Delta} = 360$ frames. So, the number of messages arriving during the next 6 minutes, $X(360)$, has Binomial distribution with parameters $n = 360$ and $p = 0.1$. Then the probability of more than 35 messages arriving during the next 6 minutes is

$$\begin{aligned} P\big(X(360) > 35\big) &= 1 - P\big(X(360) \le 35\big) \\ &= 1 - \text{cdf}_{X(360)}(35) \\ &= 1 - \text{binocdf}(35, 360, 0.1) \\ &= 0.5257. \end{aligned}$$

d) Again, in $t = 1$ hour $= 3600$ seconds, there are $n = \dfrac{t}{\Delta} = 3600$ frames. Thus, the number of messages arriving during one hour, $X(3600)$, has Binomial distribution with parameters $n = 3600$

and $p = 0.1$. Then the probability of more than 350 messages arriving during the next hour is

$$\begin{aligned} P\big(X(3600) > 350\big) &= 1 - P\big(X(3600) \leq 350\big) \\ &= 1 - \mathrm{cdf}_{X(3600)}(350) \\ &= 1 - \mathrm{binocdf}(350, 3600, 0.1) \\ &= 0.6993. \end{aligned}$$

Notice that "more than 35 messages in 6 minutes" is **not** the same as "more than 350 messages in 60 minutes"!! These are *random* variables ...

e) By (3.1), we have

$$\begin{aligned} E(T) &= \frac{1}{\lambda} = \frac{1}{6} \text{ minutes} = 10 \text{ seconds}, \\ \mathrm{Std}(T) &= \sqrt{V(T)} = \sqrt{\frac{1-p}{\lambda^2}} = \sqrt{0.0250} \text{ minutes} \approx 9.5 \text{ seconds}. \end{aligned}$$

f) Recall that the interarrival time $T = \Delta Y$, where Y has a $SGeo(p)$ distribution and, hence, $Y - 1$ has a $Geo(p)$ pdf. The next message does not arrive during the next 20 seconds, if $T > 20$. So,

$$\begin{aligned} P(T > 20) &= P(\Delta Y > 20) = P\left(Y > 20/\Delta\right) = P\left(Y > 20\right) \\ &= 1 - P(Y \leq 20) = 1 - P(Y - 1 \leq 19) \\ &= 1 - \mathrm{cdf}_{Y-1}(19) = 1 - \mathrm{geocdf}(19, 0.1) \\ &= 0.1216. \end{aligned}$$

Alternatively, this is also the probability of 0 arrivals during the next $t = 20$ seconds, i.e. during $n = \dfrac{t}{\Delta} = 20$ frames. The number of messages arriving during the next 20 seconds, $X(20)$, has a Binomial distribution with parameters $n = 20$ and $p = 0.1$. Thus, the probability that no messages arrive during the next 20 seconds is

$$P\big(X(20) = 0\big) = \mathrm{pdf}_{X(20)}(0) = \mathrm{binopdf}(0, 20, 0.1) = 0.1216.$$

■

Markov property of Binomial counting processes

It is clear that the number of successes in n trials depends *only* on the number of successes in $n-1$ trials (not on previous values $n-2, n-3, \dots$), so a Binomial process has the Markov property. Thus, it is a **Markov chain**.

Let us find the transition probability matrix. At each trial (i.e. during each frame), the number of successes $X(n)$ either increases by 1 (in case of success), or stays the same (in case of failure). Then,

$$p_{ij} = \begin{cases} p, & j = i+1 \\ q = 1-p, & j = i \\ 0, & \text{otherwise} \end{cases} . \tag{3.2}$$

Obviously, transition probabilities are constant over time and independent of past values of $X(n)$. Hence, $X(n)$ is a **homogeneous Markov chain** with transition probability matrix given by

$$P = \begin{bmatrix} 1-p & p & 0 & \dots & 0 & \dots \\ 0 & 1-p & p & \dots & 0 & \dots \\ 0 & 0 & 1-p & \dots & 0 & \dots \\ \vdots & \vdots & \vdots & & \vdots & \end{bmatrix} \tag{3.3}$$

and transition diagram depicted in Figure 5.

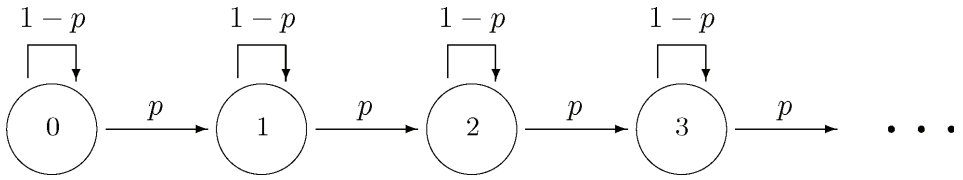


Fig. 5: Transition diagram for a Binomial counting process

Notice that it is an *irregular* Markov chain. Since $X(n)$ is non-decreasing, e.g. $p_{10}^{(h)} = 0$, for all $h \geq 0$ (once we have a success, the number of successes will *never* go back to 0). A Binomial counting process *does not* have a steady-state distribution.

Another interesting fact: the h-step transition probabilities simply form a Binomial distribution.

Indeed, $p_{ij}^{(h)}$ is the probability of going from i to j successes in h transitions, i.e.,

$$\begin{aligned} p_{ij}^{(h)} &= P\big((j-i) \text{ successes in } h \text{ trials}\big) \\ &= \begin{cases} C_h^{j-i} p^{j-i} q^{h-j+i}, & 0 \le j-i \le h \\ 0, & \text{otherwise} \end{cases} . \end{aligned}$$

Simulation of Binomial counting processes

This is straightforward, a sequence of Bernoulli trials, where we count the number of successes.

Algorithm 3.7.

1. Given: $N_B =$ sample path length of the Binomial counting process
2. Generate $U \in U(0,1)$, let $Y = (U < p)$, let $X(1) = Y$.
3. At each time t, let $Y = (U < p)$, let $X(t) = X(t-1) + Y$.
4. Return to step 3 until length N_B is achieved.

3.2 Poisson Counting Process

Now we want to consider a continuous-time counting process. The time variable t runs continuously through an interval, and thus, $X(t)$ changes at infinitely many moments. We can obtain a continuous-time process as a limit of some discrete-time process whose frame size (time between trials) Δ approaches 0 (thus allowing more frames during any fixed period of time). We will let

$$\Delta \to 0, \quad \text{as} \quad n \to \infty,$$

while keeping the arrival rate $\lambda = \text{const}$.

Think of movies, i.e. of video camera exposures. Although all motions on the screen seem continuous, we realize that an infinite amount of information could not be stored by any video device. Instead, what we see is a discrete sequence of exposures that run so fast that each motion seems continuous and smooth. Early-age video cameras shot exposures rather slowly; the interval Δ between successive shots was pretty long ($\sim 0.2 - 0.5$ sec). As a result, the quality of recorded video was rather low. Movies were "too discrete". Modern camcorders can shoot more than 200 exposures per second attaining $\Delta \leq 0.005$. With such a small Δ, the resulting movie seems perfectly continuous. A shorter frame Δ results in a "more continuous" process (see Figure 1).

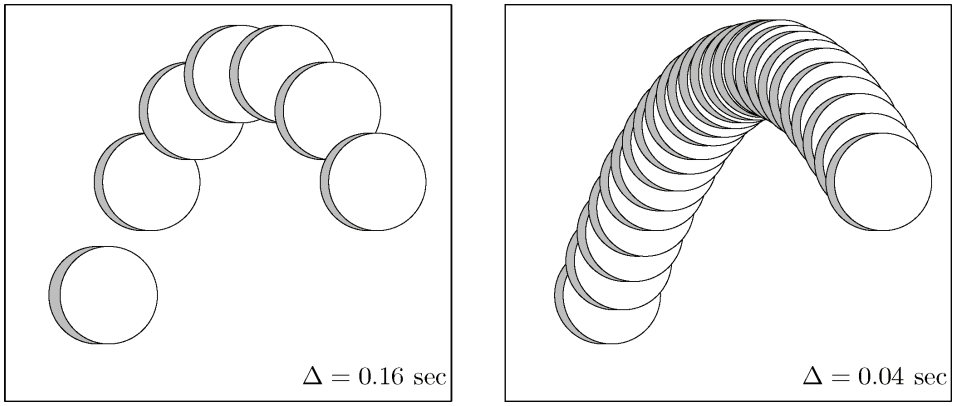


Fig. 1: Reducing the frame size Δ

So, let us take the limiting case of a Binomial counting process as $\Delta \to 0$.

Consider a Binomial counting process that counts arrivals occurring at a rate of λ / time unit. $X(t)$ denotes the number of arrivals occurring during time t.

– The *arrival rate* λ remains constant. Arrivals occur at the same rate, regardless of our choice of frame Δ.

– The *number of frames* during time t, $n = \dfrac{t}{\Delta} \to \infty$, as $\Delta \to 0$.

– The *probability of an arrival* during each frame, $p = \lambda \cdot \Delta \to 0$, as $\Delta \to 0$.

– $X(t)$, the *number of arrivals* during time t has a $B(n, p)$ distribution with expectation

$$E\big(X(t)\big) \;=\; np \;=\; \frac{t}{\Delta}p \;=\; \frac{p}{\Delta}t \;=\; \lambda t.$$

The pdf of $X(t)$ is

$$X(t) \begin{pmatrix} k \\ C_n^k p^k (1-p)^{n-k} \end{pmatrix}_{k=\overline{0,n}}.$$

So, as $n \to \infty$, the values will be $k = 0, 1, \ldots$, all the way to ∞. What about the corresponding probability $P(X = k) = C_n^k p^k (1-p)^{n-k}$? Let us see what this becomes.

$$\begin{aligned} P(X=k) &= \frac{n(n-1)\ldots(n-k+1)}{k!} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\ldots(n-k+1)}{k!} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \\ &= \frac{(\lambda t)^k}{k!} \cdot \frac{n(n-1)\ldots(n-k+1)}{n^k} \left(1 - \frac{\lambda t}{n}\right)^{-k} \left(1 - \frac{\lambda t}{n}\right)^{n} \\ &\overset{n\to\infty}{\longrightarrow} \frac{(\lambda t)^k}{k!} \cdot 1 \cdot 1 \cdot e^{-\lambda t}. \end{aligned}$$

So, the limiting pdf is

$$X(t) \begin{pmatrix} k \\ \dfrac{(\lambda t)^k}{k!} e^{-\lambda t} \end{pmatrix}_{k=0,1,\ldots}, \tag{3.1}$$

which means $X(t)$ has a Poisson $\mathcal{P}(\lambda t)$ distribution. This is a **Poisson counting process**.

Let us analyze what happens to the other characteristics.

Recall that the interarrival time $T = \Delta Y$, where Y has $SGeo(p)$ pdf. For its cdf, we have

$$\begin{aligned} F_T(t) &= P(T \le t) = P(\Delta Y \le n\Delta) \\ &= P(Y \le n) = F_Y(n) \\ &= 1-(1-p)^n = 1-\left(1-\frac{\lambda t}{n}\right)^n \\ &\overset{n\to\infty}{\longrightarrow} 1-e^{-\lambda t}. \end{aligned}$$

Hence,

$$f_T(t) = F_T'(t) = \lambda e^{-\lambda t},\ t>0, \tag{3.2}$$

so T has an $Exp(\lambda)$ pdf. Then its expectation and variance are given by

$$E(T) = \frac{1}{\lambda},\ V(T) = \frac{1}{\lambda^2}. \tag{3.3}$$

Furthermore, the time T_k of the k-th arrival is the sum of k independent $Exp(\lambda)$ interarrival times, so has $Gamma(k, 1/\lambda)$ distribution, with

$$E(T_k) = k\frac{1}{\lambda},\ V(T_k) = k\frac{1}{\lambda^2}. \tag{3.4}$$

Remark 3.1. From here, we immediately find the already-known Gamma-Poisson formula: if we want the k-th arrival to be *before or at* time t, then that means the number of arrivals by time t (X, having a Poisson $\mathcal{P}(\lambda t)$ distribution) should be *at least* k.

$$\begin{aligned} P(T_k \le t) &= P(X \ge k), \text{ or, equivalently,} \\ P(T_k > t) &= P(X < k). \end{aligned} \tag{3.5}$$

A sample path of some Poisson process is shown in Figure 2.

Example 3.2. The number of hits to a certain web site follows a Poisson process with the intensity parameter $\lambda = 7$ hits per minute. On the average, how much time is needed to get $10,000$ hits? What is the probability that this will happen within 24 hours?

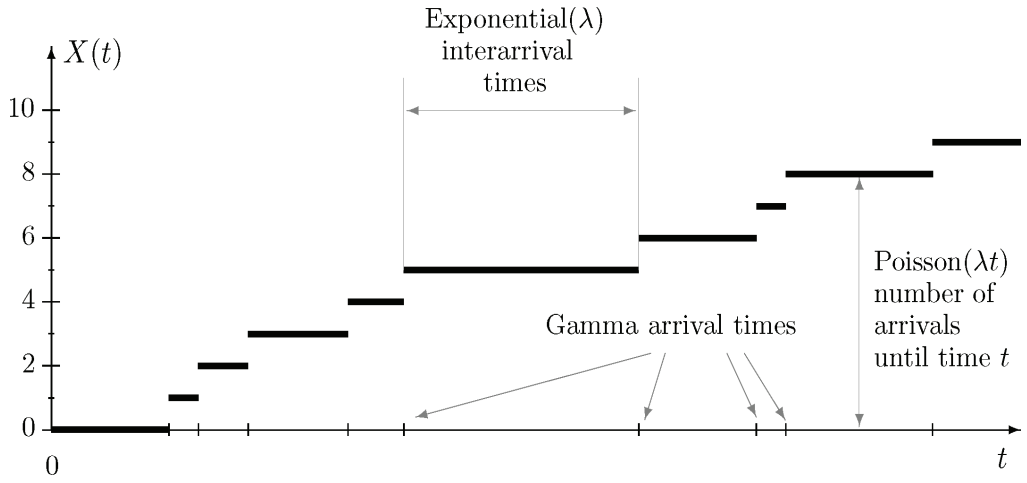


Fig. 2: Poisson process sample path

Solution. The time of the $10,000$-th hit T_k has $Gamma(k, 1/\lambda)$ distribution with parameters $k = 10,000$ and $\lambda = 7$ min^{-1}. Then, the expected time of the k-th hit is

$$\mu \;=\; E(T_k) \;=\; k \cdot \frac{1}{\lambda} \;=\; 1,428.6 \text{ minutes} \;=\; 23.8 \text{ hours}.$$

Also, the standard deviation is

$$\sigma \;=\; \text{Std}(T_k) \;=\; \sqrt{V(T_k)} \;=\; \sqrt{k} \cdot \frac{1}{\lambda} \;=\; 14.3 \text{ minutes}.$$

The probability that the $10,000$-th hit will happen within 24 hours is

$$P(T_k < 24 \cdot 60) \;=\; P(T_k < 1440) \;=\; \text{gamcdf}(1440, 10000, 1/7) \;=\; 0.7885.$$

Alternatively, with the Gamma-Poisson formula, (3.5),

$$\begin{aligned} P(T_k < 1440) \;&=\; P(T_k \le 1440) \;=\; P(X \ge k) \\ &=\; 1 - P(X < k) \;=\; 1 - P(X \le k-1) \\ &=\; 1 - \text{poisscdf}(9999, 7 \cdot 1440) \;=\; 0.7885, \end{aligned}$$

where X has a Poisson $\mathcal{P}(7 \cdot 1440)$ distribution.

■

Rare events

It can be shown that for a Poisson counting process, the following hold

$$\begin{aligned} a)\ P\Big(X(t+\Delta)-X(t)=1\Big) &= \lambda\Delta + o(\Delta), \\ b)\ P\Big(X(t+\Delta)-X(t)>1\Big) &= o(\Delta),\ \text{as}\ \Delta\to 0. \end{aligned} \tag{3.6}$$

The term $o(\Delta)$ denotes a *negligible* term, a quantity converging to 0 faster than Δ, i.e. $\dfrac{o(\Delta)}{\Delta}\to 0$, as $\Delta\to 0$.

For a Binomial counting process, the probability in part a) is the probability of 1 arrival during 1 frame and it equals $p=\lambda\Delta$, whereas the probability of *more* than 1 arrival is 0 (part b)).

For a Poisson process, these probabilities may be different, but only by "a little". The differences $X(t+\Delta)-X(t)$ are called *increments*. For a Poisson process, an increment is the number of arrivals during the time interval $(t, t+\Delta]$.

Relations (3.6) formally describe the concept of **rare events**. These events occur at random times and the probability of a new event occurring during a short interval of time is proportional to the length of that interval. Probability of *more* than 1 event during that time is much smaller, compared to the length of the interval. For such sequences of events, a Poisson process is a suitable stochastic model. Examples of rare events include telephone calls, message arrivals, virus attacks, errors in codes, traffic accidents, natural disasters, network blackouts, and so on. Relations (3.6) can also be considered as the definition of a Poisson process.

Example 3.3. Let us revisit the example from last time and model the arrivals of messages with a Poisson counting process, keeping the same arrival rate of 6 messages per minute.
a) Find the probability of no messages arriving during the next 1 minute.
b) Compute the probability of more than 35 messages arriving during the next 6 minutes.
c) Find the probability of more than 350 messages arriving during the next hour.
d) What is the average interarrival time and its standard deviation?
e) Compute the probability that the next message does not arrive during the next 20 seconds.

Solution.
a) We have $t=1$ minute and $\lambda=6$ / minute. The number of messages arriving during 1 minute, $X(1)$, has a Poisson distribution with parameter $\lambda t=6$. So the desired probability is

$$P\big(X(1)=0\big) \;=\; \text{pdf}_{X(1)}(0) \;=\; poisspdf(0,6) \;=\; 0.0025.$$

b) Similarly, the number of messages arriving in $t = 6$ minutes, $X(6)$, has a Poisson distribution with parameter $\lambda t = 36$. Then the probability of more than 35 messages arriving during that time is

$$\begin{aligned} P\big(X(6) > 35\big) &= 1 - P\big(X(6) \le 35\big) \\ &= 1 - \mathrm{cdf}_{X(6)}(35) \\ &= 1 - \mathrm{poisscdf}(35, 36) \\ &= 0.5222. \end{aligned}$$

c) Again, in $t = 1$ hour $= 60$ minutes, the number of arriving messages, $X(60)$, has Poisson distribution with parameter $\lambda t = 360$. So, the probability of more than 350 messages arriving during the next hour is

$$\begin{aligned} P\big(X(60) > 350\big) &= 1 - P\big(X(60) \le 350\big) \\ &= 1 - \mathrm{cdf}_{X(60)}(350) \\ &= 1 - \mathrm{poisscdf}(350, 360) \\ &= 0.6894. \end{aligned}$$

Notice that again, as in the case of a Binomial process, "more than 35 messages in 6 minutes" is **not** the same as "more than 350 messages in 60 minutes".

d) The interarrival time, T, now has an $Exp(\lambda) = Exp(6)$ distribution, so

$$\begin{aligned} E(T) &= \frac{1}{\lambda} = \frac{1}{6} \text{ minutes} = 10 \text{ seconds}, \\ Std(T) &= \sqrt{V(T)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{6} \text{ minutes} = 10 \text{ seconds}. \end{aligned}$$

Notice that the average interarrival time has not changed. This is to be expected, since jobs (messages) arrive at the same rate, λ, regardless of whether their arrivals are modeled by a Binomial or a Poisson process.

However, the standard deviation is slightly increased (it was 9.5 seconds before). That is because a Binomial process has a restriction on the number of arrivals during each frame, thus reducing variability.

e) Either we work with seconds (so $\lambda = \dfrac{1}{10}$ / second) and compute the probability $P(T > 20)$, where T has an $Exp(1/10)$ distribution), or in minutes ($\lambda = 6$ / minute, 20 seconds $= 1/3$ minutes)

and compute the probability $P(T > 1/3)$, where T has an $Exp(6)$ distribution. Either way, we have

$$\begin{aligned} P(T > 20) &= 1 - P(T \le 20) = 1 - \mathrm{cdf}_T(20) = 1 - \mathit{expcdf}(20, 10) = 0.1353, \\ P(T > 1/3) &= 1 - P(T \le 1/3) = 1 - \mathrm{cdf}_T(1/3) = 1 - \mathit{expcdf}(1/3, 1/6) = 0.1353. \end{aligned}$$

Again, this is the same as 0 arrivals in 20 seconds, where the number of arriving messages, $X(20)$, has a Poisson distribution with parameter $\lambda t = 2$ (or 0 arrivals in $1/3$ minutes, where the number of arriving messages, $X(1/3)$, has a Poisson distribution with parameter $\lambda t = 2$).

$$P\big(X(20) = 0\big) = \mathrm{pdf}_{X(20)}(0) = \mathit{poisspdf}(0, 2) = 0.1353.$$

■

Simulation of a Poisson counting process

Simulation of continuous-time processes has a clear problem. The time t runs continuously through the time interval, taking infinitely many values in this range. However, we cannot store an infinite number of random variables in the memory of our computer! For most practical purposes, it suffices to generate a discrete-time process with a rather short frame Δ (discretization).

But, Poisson processes can be generated without discretization. Indeed, although they are continuous-time, the value of $X(t)$ can change only a finite number of times during each interval. The process changes every time a new "rare event" or arrival occurs, which happens a Poisson $\mathcal{P}(\lambda t)$ number of times during an interval of length t. Then, it suffices to generate these moments of arrival. As we know, the first arrival time has $Exp(\lambda)$ distribution, and each interarrival time is $Exp(\lambda)$ distributed, too. So, we generate them using the ITM $(-\frac{1}{\lambda} \ln U)$ and then, generate a segment of a Poisson process during a time interval $[0, M]$ by counting the number of such times in that interval.

Algorithm 3.4.

1. Given:

   T_{max} time period,

   λ arrival rate.

2. Initial arrival time:

$T = -1/\lambda \cdot \ln U$; growing array containing arrival times,

$last = T$; last (most recent) arrival time,

3. Count number of arrivals until time T_{max}:

while $last \leq T_{max}$

$last = last - 1/\lambda \cdot \ln U$; new arrival time

$T = [T,\ last]$; array of arrival times extended

end

Chapter 4. Queuing Systems

1 Basic Notions; Main Components

Definition 1.1. *A **queuing system** is a facility consisting of one or several servers designed to perform certain tasks or process certain jobs, and a queue of jobs waiting to be processed.*
*A queuing system is called **stationary** if its distribution characteristics do not change over time.*

Jobs arrive at the queuing system at some arrival rate, wait for an available server, get processed by this server, and leave.

Example 1.2. Examples of queuing systems are:

- a computer executing tasks sent by its users;
- a printer processing jobs sent to it from different computers;
- cars at a toll booth, gas station, or auto service facility;
- an internet service provider whose customers connect to the internet, browse, and disconnect;
- people waiting in line at a cafeteria, or a bank;
- a medical office serving patients;
- airplanes waiting to take off or land, at an airport;
- a TV or radio channel viewed (listened to) by many people at various times;
- a customer service with one or several representatives on duty answering calls from their customers;
- people connecting to Facebook, Instagram, TikTok and so on.

We are now equipped to analyze a broad range of queuing systems that play a crucial role in Computer Science and other fields.

Main components of a queuing system

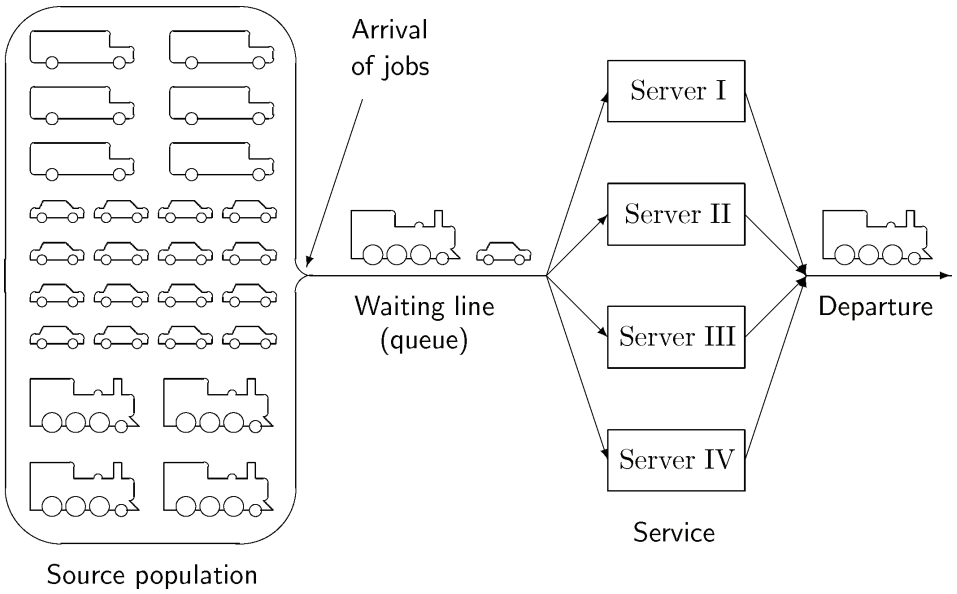


Fig. 3: Main components of a queuing system

The main stages of a queuing system are depicted in Figure 3.

Arrival

Jobs arrive to a queuing system at random times. The **number of arrivals** that occurred by the time t is a counting process $A(t)$. In stationary queuing systems, arrivals occur at **arrival rate**

$$\lambda_A \;=\; \frac{E\big(A(t)\big)}{t},\ \forall\, t>0, \tag{1.1}$$

i.e. the expected number of arrivals per time unit.
Then, the expected time between arrivals, the **mean interarrival time** is

$$\mu_A \;=\; \frac{1}{\lambda_A}. \tag{1.2}$$

Remark 1.3. Usually, arrived jobs are processed in the order of their arrivals, on a "first come-first serve" basis. When a new job arrives, it may find the system in different states.
– If one server is available, it will take the new job.

– If several servers are available, the job may be randomly sent to one of them, or the server may be assigned according to some rules. For example, the fastest server or the least loaded server may be assigned to process the new job.
– If all servers are busy, the new job will join the queue, wait until a previously arrived job is completed (accumulating waiting time) and get routed to the next available server.

Other constraints may take place. For example, a queue may have a *buffer* that limits the number of waiting jobs (like a parking garage or a restaurant). Such a queuing system is a system with **limited capacity**. The total number of jobs in it at any time is bounded by some constant C (capacity). If the capacity is full, a new job cannot enter the system until another job departs.

Also, jobs may leave the queue prematurely, say, after an excessively long waiting time (think people waiting for a table at a busy restaurant).

Servers may also open and close during the day as people need breaks or servers need maintenance.

Complex queuing systems with many extra conditions may be difficult to study analytically; however, Monte Carlo methods can be employed, to find (estimate) characteristics that evaluate the performance of most queuing systems.

Service

Once a server becomes available, it starts processing the next assigned job. Service times are usually random, they depend on the amount of work required by each task and on the efficiency of the server (slow or rapid computer, some people may work faster than others, etc).

The **average service time** is denoted by μ_S and it may vary from one server to another. The **service rate** is defined as the average number of jobs processed by a continuously working server during one unit of time, i.e.

$$\lambda_S \;=\; \frac{1}{\mu_S}. \tag{1.3}$$

Departure

When the service is completed, the job leaves the system.

To summarize, the following parameters and random variables describe the performance of a queuing system.

$$\begin{aligned}
\lambda_A &= \text{arrival rate} \\
\lambda_S &= \text{service rate} \\
\mu_A &= 1/\lambda_A = \text{mean interarrival time} \\
\mu_S &= 1/\lambda_S = \text{mean service time} \\
r &= \lambda_A/\lambda_S = \mu_S/\mu_A = \textbf{utilization} \text{ (arrival-to-service ratio)}
\end{aligned} \tag{1.4}$$

Random variables:

$$\begin{aligned}
X_s(t) &= \text{number of jobs receiving service at time } t \\
X_w(t) &= \text{number of jobs waiting in a queue at time } t \\
X(t) &= X_s(t) + X_w(t) = \text{total number of jobs in the system at time } t \\
S_k &= \text{service time for the } k^{\text{th}} \text{ job} \\
W_k &= \text{waiting time for the } k^{\text{th}} \text{ job} \\
R_k &= S_k + W_k = \textbf{response time} \text{ for the } k^{\text{th}} \text{ job, i.e.} \\
&\quad \text{total time a job spends in the system, from arrival to departure}
\end{aligned} \tag{1.5}$$

A queuing system is **stationary** if the pdf's of S_k, W_k and R_k are independent of k, in which case, we omit the index k in notation.

Utilization r is an important parameter. It shows whether or not a system can function under the current (or higher) rate of arrivals, and whether the system is over- or underloaded.

The main goal in studying queuing systems will be finding the distribution of $X(t)$, the total number of jobs in the system. Then other characteristics can be assessed from that and a comprehensive performance evaluation of a queuing system can be made.

Short Review

So, we have the following parameters and random variables that describe the performance of a stationary queuing system.

$$\begin{aligned}
\lambda_A &= \frac{E\big(A(t)\big)}{t} = \text{arrival rate} \\
\lambda_S &= \text{service rate} \\
\mu_A &= 1/\lambda_A = \text{mean interarrival time} \\
\mu_S &= 1/\lambda_S = \text{mean service time} \\
r &= \lambda_A/\lambda_S = \mu_S/\mu_A = \textbf{utilization} \text{ (arrival-to-service ratio)} \\
X_s(t) &= \text{number of jobs receiving service at time } t \\
X_w(t) &= \text{number of jobs waiting in a queue} \\
X(t) &= X_s(t) + X_w(t) = \text{total number of jobs in the system at time } t \\
S &= \text{service time for a job} \\
W &= \text{waiting time for a job} \\
R &= S + W = \textbf{response time} \text{ for a job} \\
&= \text{total time a job spends in the system, from arrival to departure}
\end{aligned}$$

2 Little's Law

This is one of the most important results in queuing theory. It was first established and used by Philip. M. Morse and other researchers in the 1950's. In 1954, Morse published it, but was not able to prove it, so he challenged his readers to find a situation where it did not hold. **John D. C. Little**, Professor Emeritus at the MIT Sloan School of Management (since 1962), proved it in 1961. Later, in the 1990's and 2000's there were more developments and versions both in theory and in practice.

Little's Law gives a simple relationship between the expected number of jobs, the expected response time, and the arrival rate. It is valid for any stationary queuing system.

Proposition 2.1 (Little's Law).

$$E(X) = \lambda_A E(R). \tag{2.1}$$

Proof. We make a diagram (see Figure 1), with time t on the x-axis and number of arrivals $A(t)$ on

the y-axis. Each job is represented by a rectangle with height 1 and length stretching between its arrival and its departure time.

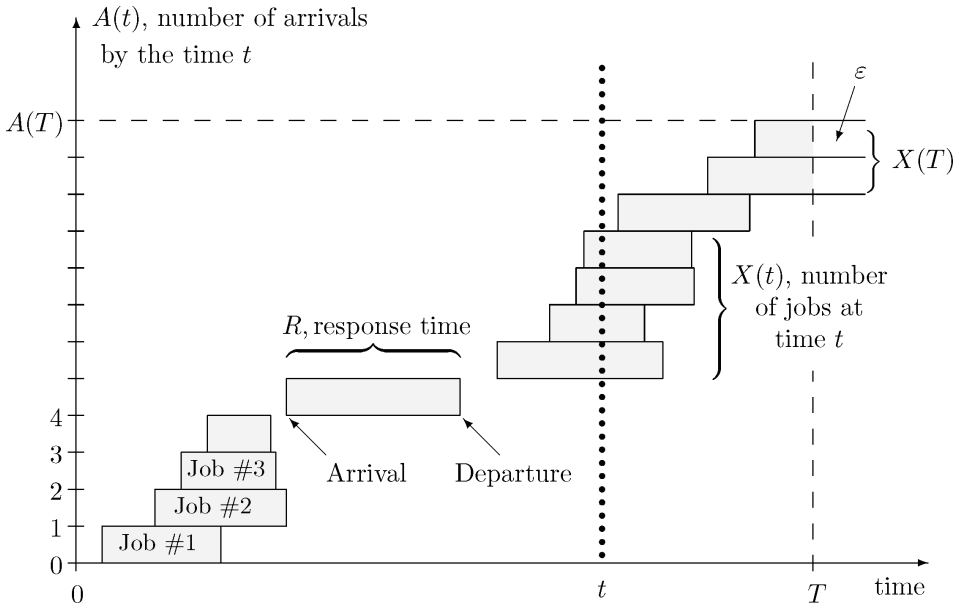


Fig. 1: Illustration of Littles's Law

We will compute the shaded area in two ways, geometrically (with areas) and analytically (with integrals).

Geometrically: add the areas of the rectangles. For job $\#k$, the area of the rectangle is its base length (since its height is 1), so the difference between departure and arrival times, i.e. *response time*:

$$\text{Area (rectangle } k) \;=\; \text{Departure time} \;-\; \text{Arrival time} \;=\; R_k.$$

By the time T ($k = \overline{1, A(T)}$), there are $A(T)$ arrivals. Among them, $X(T)$ jobs remain in the system at time T. *Not all* of these jobs are completed by time T, a portion of them will be completed *after* time T, call that portion ε. Then, the total shaded area is

$$\text{Shaded area} \;=\; \sum_{k=1}^{A(T)} R_k - \varepsilon.$$

Analytically: recall from Calculus that every area can be computed by integration of the *cross-section*. So we let t run from 0 to T and integrate the cross-section of the shaded region at t. As seen on the picture, the length of this cross-section is $X(t)$, the number of jobs in the system at time t. Hence,

$$\text{Shaded area} \;=\; \int_0^T X(t)\,dt.$$

So, we have

$$\int_0^T X(t)\,dt \;=\; \sum_{k=1}^{A(T)} R_k - \varepsilon. \tag{2.2}$$

Take expectations on both side, divide by T and then let $T \to \infty$. We compute separately the LHS (left-hand side) and RHS of (2.2). Recall from Calculus that the *mean value* of a function f on an interval $[a, b]$ is defined as

$$\frac{1}{b-a}\int_a^b f(x)dx.$$

So, in (2.2), we have

$$\begin{aligned}
\text{LHS} &= \lim_{T\to\infty} \frac{1}{T} E\left(\int_0^T X(t)\,dt\right) \\
&= \lim_{T\to\infty} E\left(\frac{1}{T}\int_0^T X(t)\,dt\right) \\
&= \lim_{T\to\infty} E(X) \;=\; E(X).
\end{aligned}$$

On the other side, we have

$$\begin{aligned}
\text{RHS} &= \lim_{T\to\infty} \frac{1}{T} E\left(\sum_{k=1}^{A(T)} R_k - \varepsilon\right) \\
&= \lim_{T\to\infty} \left(\frac{1}{T} E\Big(\sum_{k=1}^{A(T)} R_k\Big) - \frac{\varepsilon}{T}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{E(A(T))} E\big(R_k\big) - 0 \\
&= \lim_{T\to\infty} \frac{E\big(A(T)\big)}{T} E(R) \;=\; \lambda_A E(R),
\end{aligned}
$$

since $\lambda_A = \dfrac{E\big(A(T)\big)}{T}$.
Thus, we have

$$
E(X) \;=\; \lambda_A E(R).
$$

□

Example 2.2. A person walks into a bank at 10:00 a.m. He counts a total of 10 customers in the bank and assumes that this is the typical, average number. He also notices that, on average, new customers walk in every 2 minutes. When should he expect to finish his business and leave the bank?

Solution. The average number of customers in the bank, i.e. the *expected number of jobs* in the system, is

$$
E(X) \;=\; 10.
$$

On average, new customers walk in every 2 minutes, that is the *mean interarrival time*,

$$
\begin{aligned}
\mu_A &= 2 \text{ minutes, so} \\
\lambda_A &= 1/\mu_A \;=\; 1/2 \text{ / minute.}
\end{aligned}
$$

Then the amount of time he is expected to spend in the bank, i.e. the *expected response time* is, by Little's Law,

$$
E(R) \;=\; \frac{1}{\lambda_A} E(X) \;=\; \mu_A E(X) \;=\; 20 \text{ minutes.}
$$

Thus, he should expect to leave at 10:20. ■

Remark 2.3.

1. Little's Law is universal, it applies to any *stationary* queuing system *and* even to the system's components, the queue and the servers.

Thus, we can immediately deduce the equations for the number of waiting jobs,

$$E(X_w) \;=\; \lambda_A E(W),$$

and for the number of jobs currently receiving service,

$$E(X_s) \;=\; \lambda_A E(S).$$

Note that the *same* arrival rate, λ_A, applies to the components, as for the entire queuing system.

2. Looking at the second equation above, $E(S)$ is the expected or the *mean* service time, i.e. μ_S. So, we have

$$E(X_s) \;=\; \lambda_A \cdot \mu_S \;=\; \frac{\lambda_A}{\lambda_S} \;=\; r,$$

so we just obtained another important definition of *utilization*, which also justifies its name.

Definition 2.4. ***Utilization*** r *is the expected number of jobs receiving service at any given time.*

Little's Law only relates *expectations* of the number of jobs and their response time. In the remaining sections of this chapter, we evaluate the *entire distribution* of $X(t)$, which will help us compute various probabilities and expectations of interest. These quantities will describe and predict the performance of a queuing system.

Definition 2.5. *The number of jobs in a queuing system,* $X(t)$*, is called a* ***queuing process****.*

Since $X(t)$ is the *number* of jobs in the system, it is clearly a *discrete-state* stochastic process. The time set may be discrete or continuous and we will look at both cases.

In general, a queuing process is *not* a counting process because jobs arrive and depart, therefore, their number may increase and decrease, whereas any counting process is nondecreasing. However, we will use counting processes to model arrivals and service of jobs.

Another aspect is the number of servers in a queuing system, one or more. Again, we will consider both situations (in the end, even considering the case where the number of servers goes to infinity).

3 Bernoulli Single-Server Queuing Process

Definition 3.1. *A* ***Bernoulli single-server queuing process (B1SQP)*** *is a discrete-time queuing process with the following characteristics:*

- *one server;*
- *unlimited capacity;*
- *arrivals occur according to a Binomial process, and the probability of a new arrival during each frame is p_A;*
- *the probability of a service completion (and thus, a departure) during each frame is p_S, provided that there is at least one job in the system at the beginning of the frame;*
- *service times and interarrival times are independent;*
- *jobs are being serviced in the order of their arrival.*

Examples of processes modeled by a B1SQS include: customers waiting at an ATM, cars coming to a car wash or a gas station (with only one service station), documents arriving to a printer, clients calling a customer service representative, etc.

Everything we know about Binomial counting processes applies to job arrivals and to service completions, as long as there is at least one job in the system. So, we know that:
– the number of arrivals by time t, $A(t)$, is a Binomial counting process with probability p_A;
– the number of jobs being serviced at time t, $X_s(t)$, is a Binomial counting process with probability p_S (when there is at least one job in the system);
– there is a Shifted Geometric(p_A) number of frames between successive arrivals;
– there is a Shifted Geometric(p_S) number of frames between successive service completions (i.e. each service takes a $SGeo(p_S)$ number of frames);
– $p_A = \lambda_A \Delta$;
– $p_S = \lambda_S \Delta$.

Markov property

Obviously, a B1SQS is a Markov chain. Since the probabilities p_A and p_S never change, it is also a *homogeneous* Markov chain. The number of jobs in the system increases by 1 with every arrival and decreases by 1 with each departure. Conditions of a Binomial process guarantee that *at most one arrival* and *at most one departure* may occur during each frame.

The states are $\{0, 1, \ldots\}$ (number of jobs in the system). Let us find the transition probabilities.

$$\begin{aligned} p_{00} &= P\big(\text{ 0 arrivals }\big) = 1 - p_A \\ p_{01} &= P\big(\text{ 1 arrival }\big) = p_A. \end{aligned}$$

In general, for $i \geq 1$,

$$\begin{aligned}
p_{i,i-1} &= P\big(\text{ 0 arrivals and 1 departure }\big) = P\big(\{0 \text{ arrivals}\} \cap \{1 \text{ departure}\}\big) = (1-p_A)p_S \\
p_{i,i} &= P\Big(\big(\{0 \text{ arrivals}\} \cap \{0 \text{ departures}\}\big) \cup \big(\{1 \text{ arrival}\} \cap \{1 \text{ departure}\}\big)\Big) \\
&= P\big(\{0 \text{ arrivals}\} \cap \{0 \text{ departures}\}\big) + P\big(\{1 \text{ arrival}\} \cap \{1 \text{ departure}\}\big) \\
&= (1-p_A)(1-p_S) + p_A p_S \\
p_{i,i+1} &= P\big(\{1 \text{ arrival}\} \cap \{0 \text{ departures}\}\big) = p_A(1-p_S)
\end{aligned}$$

All the other transition probabilities are 0, since the number of jobs cannot change by more than 1 in any single frame.
So, the transition probability matrix is

$$P = \begin{bmatrix}
1-p_A & p_A & 0 & \dots & 0 & \dots \\
(1-p_A)p_S & \begin{matrix}(1-p_A)(1-p_S)\\ +p_A p_S\end{matrix} & p_A(1-p_S) & \dots & 0 & \dots \\
0 & (1-p_A)p_S & \begin{matrix}(1-p_A)(1-p_S)\\ +p_A p_S\end{matrix} & \dots & 0 & \dots \\
0 & 0 & (1-p_A)p_S & \dots & 0 & \dots \\
\vdots & \vdots & \vdots & & \ddots &
\end{bmatrix}, \quad (3.1)$$

an $\infty \times \infty$ tridiagonal matrix. Below, see the transition diagram.

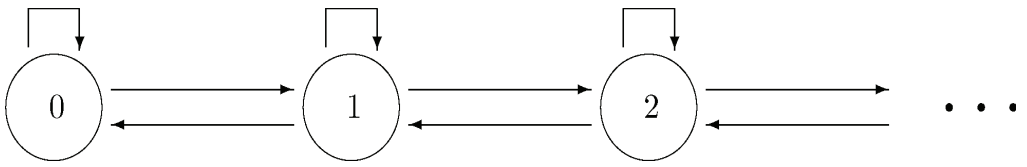


Fig. 2: Transition diagram for a B1SQS

The transition probability matrix may be used, for example, to simulate this queuing system and study its performance, as we did with general Markov chains. One can also compute k-step

transition probabilities and predict the load of a server or the length of a queue at any time in the future.

Example 3.2. Jobs (documents) are sent to a printer at the rate of 20 per hour. It takes an average of 40 seconds to print a document. Currently, the printer is printing a job, and there is another job stored in a queue. Assume a B1SQS with 20-second frames is modeling this printer.
a) Compute the probability that the printer will be idle in 2 minutes.
b) Find the expected total number of jobs in the system in 2 minutes.
c) What is the expected length of the queue in 2 minutes?
d) What is the expected waiting time for a document in 2 minutes?
e) On average, how long does it take to get the printout of a document in 2 minutes?

Solution.
First off, let us note that any printer represents a single-server queuing system, because it can process only one job at a time while other jobs are waiting in a queue.

Now, parameters are given in hours, in minutes and in seconds, so let us choose the "middle" one, i.e., express everything in minutes. We are given:

$$\begin{aligned}
\lambda_A &= 20 \text{ / hour } = 1/3 \text{ / minute, so} \\
\mu_A &= 3 \text{ minutes,} \\
\mu_S &= 40 \text{ seconds } = 2/3 \text{ minutes, so} \\
\lambda_S &= 1/\mu_S = 3/2 \text{ / minute,} \\
\Delta &= 20 \text{ seconds } = 1/3 \text{ minutes.}
\end{aligned}$$

Then

$$\begin{aligned}
p_A &= \lambda_A \Delta = 1/9, \ 1 - p_A = 8/9, \\
p_S &= \lambda_S \Delta = 1/2, \ 1 - p_S = 1/2.
\end{aligned}$$

The transition probabilities are

$$\begin{aligned}
p_{00} &= 1 - p_A = 8/9, \\
p_{01} &= p_A = 1/9, \\
p_{i,i-1} &= (1-p_A)p_S = 8/9 \cdot 1/2 = 4/9, \\
p_{i,i} &= (1-p_A)(1-p_S) + p_A p_S = 8/9 \cdot 1/2 + 1/9 \cdot 1/2 = 1/2, \\
p_{i,i+1} &= p_A(1-p_S) = 1/9 \cdot 1/2 = 1/18.
\end{aligned}$$

Hence,

$$P = \begin{bmatrix}
8/9 & 1/9 & 0 & 0 & \dots \\
4/9 & 1/2 & 1/18 & 0 & \dots \\
0 & 4/9 & 1/2 & 1/18 & \dots \\
0 & 0 & 4/9 & 1/2 & \dots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.$$

Now, in $t = 2$ minutes, there are $n = \dfrac{t}{\Delta} = 6$ frames. The distribution of X after 6 frames is

$$P_6 = P_0 \cdot P^6.$$

The initial distribution (2 jobs in the system) is

$$P_0 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; \dots].$$

Here is an interesting problem. How do we deal with matrix P that has *infinitely* many rows and columns? Fortunately, we only need a small portion of this matrix. In the course of 6 frames, the number of jobs in the system, $X(t)$, can change by 6 *at most* (see Figure 2), i.e. it can reach a *maximum* of 8. Thus, it is sufficient to consider the first 9 rows and 9 columns of P *only*, corresponding to states $\{0, 1, \dots, 8\}$.

So, we consider P_0 (and P_6) as having length 9,

$$P_0 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]$$

and P a 9×9 matrix,

$$P \;=\; \left[\begin{array}{ccccccccc} 8/9 & 1/9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4/9 & 1/2 & 1/18 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4/9 & 1/2 & 1/18 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4/9 & 1/2 & 1/18 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4/9 & 1/2 & 1/18 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4/9 & 1/2 & 1/18 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/9 & 1/2 & 1/18 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4/9 & 1/2 & 1/18 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/9 & 1/2 \end{array}\right].$$

In 2 minutes (6 frames), the distribution will be

$$P_6 \;=\; P_0 \cdot P^6 \;=\; [\,0.6436 \quad 0.25 \quad 0.0799 \quad 0.0218 \quad 0.0041 \quad 0.0005 \quad 0 \quad 0 \quad 0\,].$$

Now we can answer all the questions.

a) The probability that the printer is idle after 2 minutes is the probability of 0 jobs in the system at that time, i.e.

$$P_6(0) \;=\; 0.6436.$$

b) In 2 minutes, the total number of jobs in the system, $X(2)$, has pdf

$$X\begin{pmatrix} 0 & \dots & 8 \\ & P_6 & \end{pmatrix} \;=\; \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 0.6436 & 0.25 & 0.0799 & 0.0218 & 0.0041 & 0.0005 & 0 & 0 & 0 \end{pmatrix},$$

so

$$E(X) \;=\; \sum_{k=0}^{8} kP_6(k) \;=\; 0.4944 \text{ jobs.}$$

c) Out of the X jobs in the system above, X_w jobs are waiting in a queue and X_s are being serviced. The expected length of the queue is the expected value

$$\begin{aligned} E(X_w) &= E(X - X_s) \\ &= E(X) - E(X_s). \end{aligned}$$

We have found $E(X)$, so let us turn our attention to X_s.

Since the server (printer) can process *at most* 1 job at a time, X_s is either 0 or 1, i.e. it has a Bernoulli distribution. With what parameter p? The parameter is the probability of "success", in this case, the probability that the system is working, so, *not* idle:

$$p \;=\; P(\text{printer is busy}) \;=\; 1 - P(\text{printer is idle}) \;=\; 1 - 0.6436 \;=\; 0.3564.$$

So the pdf of X_s is

$$X_s \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

and its expected value

$$E(X_s) \;=\; p \;=\; 0.3564.$$

Then the expected queue length is

$$E(X_w) \;=\; E(X) - E(X_s) \;=\; 0.4944 - 0.3564 \;=\; 0.138 \text{ jobs.}$$

d) The expected waiting time for a document is $E(W)$. By Little's Law, we have

$$\begin{aligned} E(W) \;&=\; \frac{1}{\lambda_A} E(X_w) \;=\; \mu_A E(X_w) \\ &=\; 3 \cdot 0.138 \text{ minutes} \;=\; 0.414 \text{ minutes} \\ &=\; 24.84 \text{ seconds.} \end{aligned}$$

e) This is the expected total time the job spends in the system, i.e., the expected *response time* of a job, in 2 minutes. Again, by Little's Law, that number is

$$E(R) \;=\; \frac{1}{\lambda_A} E(X) \;=\; 3 \cdot 0.4944 \text{ minutes} \;=\; 1.4832 \text{ minutes.}$$

■

Remark 3.3.

1. A B1SQS is an *irregular* Markov chain. *Any* k-step transition probability matrix contains zeros because a k-step transition from 0 to $k+1$ is *impossible*. It requires at least $k+1$ arrivals, and this cannot happen by the conditions of the Binomial process of arrivals.

2. However, without the Binomial counting process restrictions, it can be shown that any system whose service rate exceeds the arrival rate (i.e., jobs can be served *faster* than they arrive, so there

is no overload),

$$\lambda_S > \lambda_A,$$

does have a steady-state distribution. Its computation is possible, despite the infinite dimension of P, but a little complicated. Instead, we will compute the steady-state distribution of a *continuous* queuing process, obtained by letting the frame size $\Delta \to 0$.

Systems with limited capacity

As we have seen, the number of jobs in a B1SQS may potentially reach any value. However, in practice, many systems have limited resources for storing jobs. Then, there is a maximum number of jobs C that can possibly be in the system simultaneously. This number is called **capacity**. As examples, consider people going to a restaurant, cars entering a parking lot, customers going into a bank, etc.

How does the situation change for a queuing system with a limited capacity $C < \infty$? Not much, but it *does* make a difference. Up until the capacity C is reached, the system operates as before. Things change when $X = C$. At this time, the system is full, so it can accept new jobs into its queue *only* if some job departs. We have

$$\begin{aligned}
p_{C,C-1} &= P\big(\text{ 0 arrivals } \cap \text{ 1 departure }\big) = (1-p_A)p_S \text{ (as before)},\\
p_{C,C} &= P\big(\big(\text{ 0 arrivals } \cap \text{ 0 departures }\big) \cup \big(\text{ 1 arrival } \cap \text{ 1 departure }\big)\\
&\quad \cup \big(\text{ 1 arrival } \cap \text{ 0 departures }\big)\big)\\
&= (1-p_A)(1-p_S) + p_A p_S + p_A(1-p_S)\\
&= 1-(1-p_A)p_S.
\end{aligned}$$

This Markov chain has states $0, 1, ..., C$, its transition probability matrix is finite, and it is *regular* (any state can be reached in C steps). The transition diagram for a system with limited capacity is given in Figure 3.

Example 3.4. A customer service representative has a telephone with 2 lines, so she can talk to a customer while having another one "on hold". Suppose the representative gets an average of 10 calls per hour and the average phone conversation lasts 4 minutes. Assuming a B1SQS with 1-minute frames find the steady-state distribution and interpret it.

Solution.

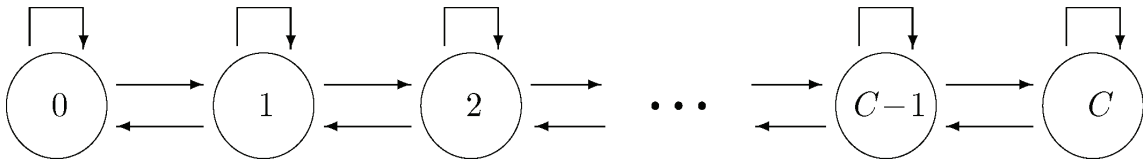


Fig. 3: Transition diagram for a B1SQS with limited capacity C

Obviously, this is a system with limited capacity $C = 2$. When the capacity is reached and someone tries to call, (s)he will get a busy signal or voice mail.
This Markov chain $X(t)$ has 3 states, $0, 1, 2$ and we have:

$$\begin{aligned} \lambda_A &= 10 \text{ / hour } = 1/6 \text{ / minute,} \\ \mu_S &= 4 \text{ minutes, so} \\ \lambda_S &= 1/\mu_S = 1/4 \text{ / minute,} \\ \Delta &= 1 \text{ minute,} \end{aligned}$$

so,

$$\begin{aligned} p_A &= \lambda_A \Delta = 1/6, \ 1 - p_A = 5/6, \\ p_S &= \lambda_S \Delta = 1/4, \ 1 - p_S = 3/4. \end{aligned}$$

The transition probability matrix, of dimensions 3×3, is

$$P = \begin{bmatrix} 1 - p_A & p_A & 0 \\ (1 - p_A)p_S & (1 - p_A)(1 - p_S) + p_A p_S & p_A(1 - p_S) \\ 0 & (1 - p_A)p_S & 1 - (1 - p_A)p_S \end{bmatrix} = \begin{bmatrix} 5/6 & 1/6 & 0 \\ 5/24 & 2/3 & 1/8 \\ 0 & 5/24 & 19/24 \end{bmatrix}.$$

The steady-state distribution is found, as usually, from the system

$$\begin{cases} \pi P &= \pi \\ \sum_{k=0}^{2} \pi_k &= 1, \end{cases}$$

which leads to

$$\begin{cases} \pi_0 - \dfrac{5}{4}\pi_1 & = & 0 \\ \dfrac{3}{5}\pi_1 - \pi_2 & = & 0 \\ \pi_0 + \pi_1 + \pi_2 & = & 1, \end{cases}$$

with solution

$$\begin{aligned} \pi_0 &= \frac{25}{57} \approx 0.439, \\ \pi_1 &= \frac{20}{57} \approx 0.351, \\ \pi_2 &= \frac{12}{57} \approx 0.21. \end{aligned}$$

Interpretation: 43.9% of the time the representative is not talking on the phone (and, implicitly, there is no one on hold), 35.1% of the time she talks to a customer, but the second line is open, and 21% of the time both lines are busy (one talking, one holding) and no new calls can get through.

■

4 M / M / 1 Queuing Systems

We discuss now continuous-time queuing systems with the usual approach: consider a discrete-time queuing system and let the frame size $\Delta \to 0$.
First, let us explain the notation:
Notation A queuing system is denoted by **A / S / k / C / P**, where
− **A** denotes the distribution of **interarrival times**;
− **S** denotes the distribution of **service times**;
− **k** denotes the number of **servers**;
− **C** denotes the **capacity**;
− **P (or K)** denotes the size of the **source population**.
Usually, the default values for the last two are $C = P = \infty$ and they are dropped from the notation. When the Exponential distribution is considered for A or S, then it is denoted by M, because it is *memoryless* and the resulting process is *Markov*. You may see other notations, like G for "general" (any distribution), D for "deterministic" (fixed interarrival time), etc.

Definition 4.1. *An **M / M / 1 queuing process** is a continuous-time Markov queuing process with the following characteristics:*

- *one server;*
- *unlimited capacity;*
- *Exponential interarrival times with arrival rate λ_A;*
- *Exponential service times with service rate λ_S;*
- *service times and interarrival times are independent.*

Remark 4.2. Let us recall that Exponential interarrival times imply a Poisson process of arrivals with parameter λ_A. This is a very popular model for telephone calls and many other types of arriving jobs.

We study M / M / 1 systems by starting with a B1SQS and letting its frame size Δ go to zero. We want to derive the steady-state distribution and other quantities of interest that measure the system's performance.
Recall that

$$\begin{aligned} p_A &= \lambda_A \Delta, \\ p_S &= \lambda_S \Delta \end{aligned}$$

and as Δ gets small, Δ^2 becomes practically negligible. Then the transition probabilities are

$$\begin{aligned}
p_{00} &= 1 - p_A = 1 - \lambda_A\Delta \\
p_{01} &= p_A = \lambda_A\Delta \\
p_{i,i-1} &= (1-p_A)p_S = (1-\lambda_A\Delta)\lambda_S\Delta \approx \lambda_S\Delta \\
p_{i,i} &= (1-p_A)(1-p_S) + p_A p_S \approx 1 - \lambda_A\Delta - \lambda_S\Delta \\
p_{i,i+1} &= p_A(1-p_S) \approx \lambda_A\Delta,
\end{aligned}$$

for $i = 1, 2, \ldots$. The transition probability matrix becomes

$$P \approx \begin{bmatrix}
1-\lambda_A\Delta & \lambda_A\Delta & 0 & \ldots & 0 & \ldots \\
\lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \lambda_A\Delta & \ldots & 0 & \ldots \\
0 & \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \ldots & 0 & \ldots \\
0 & 0 & \lambda_S\Delta & \ldots & 0 & \ldots \\
\vdots & \vdots & \vdots & & \ddots &
\end{bmatrix}. \tag{4.1}$$

Let us find the steady-state distribution from

$$\begin{cases}
\pi P &= \pi \\
\sum_{i=0}^{\infty} \pi_i &= 1,
\end{cases}$$

a system of infinitely many equations with infinitely many unknowns.

The first equation is

$$\begin{aligned}
[\pi_0 \ \pi_1 \ \pi_2 \ \ldots] \cdot \begin{bmatrix} 1-\lambda_A\Delta \\ \lambda_S\Delta \\ 0 \\ \vdots \end{bmatrix} &= \pi_0, \text{ i.e.} \\
(1-\lambda_A\Delta)\pi_0 + \lambda_S\Delta\pi_1 &= \pi_0, \text{ i.e.} \\
-\lambda_A\Delta\pi_0 + \lambda_S\Delta\pi_1 &= 0, \text{ i.e.} \\
\lambda_S\pi_1 &= \lambda_A\pi_0.
\end{aligned}$$

This is called the *first balance equation*. From here, we get

$$\pi_1 \;=\; \frac{\lambda_A}{\lambda_S}\pi_0 \;=\; r\pi_0. \tag{4.2}$$

The second equation is

$$\begin{aligned}
[\pi_0\ \pi_1\ \pi_2\ \ldots] \cdot \begin{bmatrix} \lambda_A\Delta \\ 1-\lambda_A\Delta-\lambda_S\Delta \\ \lambda_S\Delta \\ 0 \\ \vdots \end{bmatrix} &= \pi_1, \text{ i.e.} \\
\lambda_A\Delta\pi_0 + (1-\lambda_A\Delta-\lambda_S\Delta)\pi_1 + \lambda_S\Delta\pi_2 &= \pi_1, \text{ i.e.} \\
\lambda_A\Delta\pi_0 - \lambda_A\Delta\pi_1 - \lambda_S\Delta\pi_1 + \lambda_S\Delta\pi_2 &= 0, \text{ i.e. (since } \lambda_A\pi_0 = \lambda_S\pi_1) \\
-\lambda_A\Delta\pi_1 + \lambda_S\Delta\pi_2 &= 0, \text{ i.e.} \\
\lambda_S\pi_2 &= \lambda_A\pi_1.
\end{aligned}$$

Thus, we obtained the *second balance equation*, from which

$$\pi_2 \;=\; \frac{\lambda_A}{\lambda_S}\pi_1 \;=\; r\pi_1 \;=\; r^2\pi_0. \tag{4.3}$$

This trend of balance equations will continue exactly the same way, because every next column of matrix P is just the same as the previous column, only shifted down by 1 position. Thus, the general balance equation looks like

$$\lambda_S\pi_i \;=\; \lambda_A\pi_{i-1},$$

or

$$\pi_i \;=\; r\pi_{i-1}.$$

Combining it with the previous equations, we have

$$\pi_i \;=\; r\pi_{i-1} \;=\; r^2\pi_{i-2} \;=\; \ldots \;=\; r^i\pi_0,\ i = 1, 2, \ldots. \tag{4.4}$$

Finally, in the *normalizing equation* $\sum_{i=0}^{\infty} \pi_i = 1$, we get

$$\sum_{i=0}^{\infty} \pi_i \;=\; \sum_{i=0}^{\infty} r^i \pi_0 \;=\; 1. \tag{4.5}$$

Now, the Geometric series $\pi_0 \sum_{i=0}^{\infty} r^i$ is convergent if its ratio $r < 1$, in which case the series is equal to $\frac{\pi_0}{1-r}$. So, assuming the utilization r is less than 1, the last equation becomes

$$\begin{aligned} \frac{\pi_0}{1-r} &= 1, \text{ i.e.} \\ \pi_0 &= 1 - r. \end{aligned}$$

Then the steady-state distribution of this queuing process is

$$\pi_i \;=\; r^i(1-r),\ i = 0, 1, \ldots \tag{4.6}$$

So the pdf of $X(t)$ (the total number of jobs in the system at time t) is

$$X(t) \begin{pmatrix} i \\ (1-r)r^i \end{pmatrix}_{i=0,1,\ldots}, \tag{4.7}$$

a $Geo(p)$ distribution with parameter $p = 1 - r$. Then, we know

$$\begin{aligned} E(X) &= \frac{q}{p} = \frac{r}{1-r}, \\ V(X) &= \frac{q}{p^2} = \frac{r}{(1-r)^2}. \end{aligned} \tag{4.8}$$

Evaluation of the system's performance

Now we can analyze the main parameters and distributions that characterize the queuing system, directly from the distribution (4.7).

Utilization

We know $r = \dfrac{\lambda_A}{\lambda_S}$. Now we also have $r = 1 - \pi_0$. What does that mean?

$$\pi_0 \;=\; P(X = 0) \;=\; P(\text{there are } \textit{no} \text{ jobs in the system}) \;=\; P(\text{the system is } \textit{idle}),$$

so

$$r \;=\; P(X > 0) \;=\; 1 - \pi_0 \;=\; 1 - P(\text{the system is idle}) \;=\; P(\text{the system is busy}). \quad (4.9)$$

So, we can say that r is the proportion of time when the system is put to work or *utilized*, hence the name **utilization**.

Obviously, the system is functional only if $r < 1$ (we used this for the convergence of the Geometric series). If $r \geq 1$, the system gets *overloaded*. Arrivals are too frequent compared to the service rate and the system cannot manage the incoming flow of jobs. The number of jobs will accumulate (unless it has a limited capacity) until the system will no longer function.

Waiting time

When a job arrives, it finds the system with X jobs in it. The new job waits in a queue, while those X jobs are being serviced. Thus, its waiting time is the sum of service times of X jobs

$$W \;=\; S_1 + S_2 + \ldots + S_X.$$

Recall that service times are Exponential and this distribution has the *memoryless* property (i.e. $P(S > x + y \mid S > x) = P(S > y)$). So, even if the first job has already started service, its *remaining* service time still has $Exp(\lambda_S)$ distribution, regardless of how long it has already been served or *when* its service time began. Then, the expected waiting time is

$$\begin{aligned} E(W) \;&=\; E(S_1 + S_2 + \ldots + S_X) \;=\; \sum_{i=1}^{X} E(S_i) \\ &=\; E(S \cdot X) \;=\; E(S)E(X) \\ &=\; \mu_S \cdot \frac{r}{1-r} \;=\; \frac{r}{\lambda_S(1-r)}. \end{aligned} \quad (4.10)$$

Remark 4.3.

1. At the step $E(S \cdot X) \;=\; E(S)E(X)$, we actually used the fact that service times are *independent*

of the number of jobs in the system at that time.
2. The random variable W, the waiting time, is a rare example of a variable whose distribution is neither discrete nor continuous. Notice that it has a probability *distribution (mass)* function at 0, because

$$P(W = 0) \;=\; P(\text{the system is } \textit{idle}) \;=\; 1 - r$$

is the probability that the server is idle and available and there is no waiting time for a new job. On the other hand, for all $x > 0$, it has a probability *density* function. Given any positive number of jobs $X = n$, the waiting time is the sum of n independent $Exp(\lambda_S)$ times, which is a $Gamma(n, 1/\lambda_S)$ random variable, so continuous. Such a distribution is called *mixed*.

Response time

Response time is the time a job spends in the system, from its arrival to its departure. It consists of waiting time (if any) and service time. So, the expected response time is then

$$\begin{aligned} E(R) &= E(W) + E(S) \\ &= \mu_S \cdot \frac{r}{1-r} + \mu_S \;=\; \mu_S\Big(\frac{r}{1-r} + 1\Big) \\ &= \frac{\mu_S}{1-r} \;=\; \frac{1}{\lambda_S(1-r)}. \end{aligned} \tag{4.11}$$

Queue

The length of the queue is the number of waiting jobs

$$X_w \;=\; X - X_s.$$

As we have discussed in Example 3.2. (Lecture 8), the number of jobs being serviced, X_s, at any time is either 0 or 1 (because there is only one server), so it has a Bernoulli distribution with parameter

$$P(\text{the system/server is busy}) \;=\; r$$

and, hence,

$$E(X_s) \;=\; 0 \cdot (1-r) + 1 \cdot r \;=\; r.$$

Then, the expected queue length is

$$\begin{aligned} E(X_w) &= E(X) - E(X_s) \\ &= \frac{r}{1-r} - r \;=\; r\Big(\frac{1}{1-r} - 1\Big) \\ &= \frac{r^2}{1-r}. \end{aligned} \tag{4.12}$$

So, to summarize:

Main performance characteristics of an M / M / 1 queuing system

- **Expected number of jobs in the system**

$$E(X) \;=\; \frac{r}{1-r},$$

- **Expected queue length**

$$E(X_w) \;=\; \frac{r^2}{1-r},$$

- **Expected number of jobs being serviced**

$$E(X_s) \;=\; r,$$

- **Expected response time**

$$E(R) \;=\; \frac{\mu_S}{1-r} \;=\; \frac{1}{\lambda_S(1-r)},$$

- **Expected waiting time**

$$E(W) \;=\; \frac{\mu_S r}{1-r} \;=\; \frac{r}{\lambda_S(1-r)},$$

- **Expected service time**

$$E(S) \;=\; \mu_S,$$

- **Utilization**

$$\begin{aligned} r &= P(X > 0) \;=\; 1 - \pi_0 \;=\; P(\text{system is busy}), \\ 1-r &= P(X = 0) \;=\; \pi_0 \;=\; P(\text{system is idle}). \end{aligned}$$

Remark 4.4. Little's Law applies to M / M / 1 queuing systems and their components, the queue and the server. Assuming the system is functional ($r < 1$), all jobs go through the entire system, and thus, each component is subject to the same arrival rate λ_A. Notice that, indeed, we have

$$\begin{aligned}
\lambda_A E(R) &= \lambda_A \cdot \frac{1}{\lambda_S(1-r)} = \frac{r}{1-r} = E(X) \\
\lambda_A E(W) &= \lambda_A \cdot \frac{r}{\lambda_S(1-r)} = \frac{r^2}{1-r} = E(X_w) \\
\lambda_A E(S) &= \lambda_A \cdot \mu_S = \frac{\lambda_A}{\lambda_S} = r = E(X_s).
\end{aligned}$$

Example 4.5. Messages arrive to a communication center at random times according to a Poisson process, with an average of 5 messages per minute. They are transmitted through a single channel in the order they were received. On average, it takes 10 seconds to transmit a message. Compute the main performance characteristics for this center.

Solution. Recall that a Poisson process of arrivals *implies* Exponential interarrival times (and the other way around). Since messages are transmitted (i.e. jobs are being serviced) in the order they arrive, we also have Exponential service times. Thus, conditions of an M / M / 1 queuing system are satisfied.
We have

$$\begin{aligned}
\lambda_A &= 5 \text{ / minute}, \\
\mu_S &= 10 \text{ seconds} = \frac{1}{6} \text{ minutes}, \\
\lambda_S &= 6 \text{ / minute}, \\
r &= \frac{5}{6} = 0.833 < 1.
\end{aligned}$$

This is also the proportion of time, 83.3%, when the channel is busy and the probability of a non-zero waiting time. Then, we have:
Average number of messages stored in the system at any time

$$E(X) = \frac{r}{1-r} = 5.$$

Out of these, average number of messages waiting to be transmitted

$$E(X_w) \;=\; \frac{r^2}{1-r} \;=\; \frac{25}{6} \;\approx\; 4.17.$$

Average number of messages being transmitted

$$E(X_s) \;=\; r \;=\; \frac{5}{6} \;\approx\; 0.83.$$

When a message arrives to the center, its average waiting time until transmission is

$$E(W) \;=\; \frac{\mu_S r}{1-r} \;=\; \frac{r}{\lambda_S(1-r)} \;=\; \frac{5}{6}\,\text{minutes} \;=\; 50\,\text{seconds}.$$

The total time from arrival until the end of transmission has an average of

$$E(R) \;=\; \frac{\mu_S}{1-r} \;=\; \frac{1}{\lambda_S(1-r)} \;=\; 1\,\text{minute} \;=\; 60\,\text{seconds}.$$

■

Notice that the utilization was less than 1, but not by much. Let us try a little bit of forecasting for this system and see what happens when the arrival rate is *slightly* increased, keeping the service rate the same.

Example 4.6. Suppose that next year the customer base of this transmission center is projected to increase by 10%, and thus, its incoming traffic rate, λ_A, increases by 10%, also. How will this affect the center's performance?

Solution. So, with that increase, we now have

$$\begin{aligned} \lambda_A \;&=\; 5 + 0.1 \cdot 5 \;=\; 5.5 \;=\; \frac{11}{2}/\ \text{minute}, \\ r \;&=\; \frac{11}{2} \cdot \frac{1}{6} \;=\; \frac{11}{12} \;<\; 1. \end{aligned}$$

The new system's performance parameters are

$$\begin{aligned}
E(X) &= \frac{r}{1-r} = 11 \text{ (compared to } 5 \text{ before)},\\
E(X_w) &= \frac{r^2}{1-r} = 10.08 \text{ (compared to } 4.17 \text{ before)},\\
E(X_s) &= r = 0.92 \text{ (compared to } 0.83 \text{ before)},\\
E(W) &= \frac{\mu_S}{1-r} = 110 \text{ seconds (compared to } 50 \text{ before)},\\
E(R) &= \frac{\mu_S}{1-r} = 120 \text{ seconds (compared to } 60 \text{ before)}.
\end{aligned}$$

■

Notice that the response time, the waiting time, the average number of stored messages (and hence, the average required amount of memory) more than doubled when the number of customers increased by a mere 10%. The utilization r is still less than 1, but *dangerously close* to 1, when the system gets overloaded. For high values of r, various parameters of the system increase rapidly.

We could forecast the two-year future of the system, assuming a 10% increase of a customer base each year. It appears that during the second year the utilization will exceed 1, making the system unable to function. What solutions are there? Either increase the service rate (by using better equipment, higher internet speed, etc) **or** add more channels (servers) to help handle all the arriving messages, so have a *multiserver* queuing system. The new system will then have more than one channel-server, and it will be able to process more arriving jobs

5 Multiserver Queuing Systems

We now consider queuing systems with several servers. We assume that each server can perform the same range of services; however, in general, some servers may be faster than others. Thus, the service times for different servers may potentially have different distributions.

When a job arrives, it either finds all servers busy serving jobs, or it finds one or several available servers. In the first case, the job will wait in a queue for its turn, whereas in the second case, it will be routed to one of the idle servers. A mechanism assigning jobs to available servers may be random, or it may be based on some rule.

The number of servers may be finite or infinite. A system with infinitely many servers can afford an unlimited number of concurrent users (e.g. any number of people can watch a TV channel simultaneously), so there is no queue, no waiting time.

As before (the single server case), we start with a discrete-time k-server queuing process (de-

scribed in terms of Bernoulli trials), verify that the number of jobs in the system at time t is a Markov process, find its transition probability matrix, then get a continuous-time process by letting the frame size $\Delta \to 0$, compute its steady-state distribution π and finally use it to evaluate the system's long-term performance characteristics.

We treat a few common and analytically simple cases in detail. Sure enough, advanced theory goes further, but it is beyond the scope of this course. However, as mentioned previously, more complex and non-Markov queuing systems can be analyzed by Monte Carlo methods.

Remark 5.1. The utilization r no longer has to be less than 1. A system with k servers can handle k times the traffic of a single-server system; therefore, it will function with any $r < k$.

5.1 Bernoulli k-Server Queuing Process

Definition 5.2. *A **Bernoulli k-server queuing process (BkSQP)** is a discrete-time queuing process with the following characteristics:*

- *k servers;*
- *unlimited capacity;*
- *arrivals occur according to a Binomial process with probability of a new arrival during each frame p_A;*
- *during each frame, each busy server completes its job with probability p_S, independently of the other servers and independently of the process of arrivals.*

So, all interarrival times and all service times are independent Shifted Geometric random variables (multiplied by the frame length Δ) with parameters p_A and p_S, respectively. Therefore, since Shifted Geometric variables have a memoryless property, again this process is Markov. The novelty is that now several jobs may finish during the same frame.

Suppose that $X_s = n$ jobs are currently getting service. During the next frame, each of them may finish and depart, independently of the other jobs. Then the number of departures, X_d, is the number of successes in n independent Bernoulli trials (with "success" meaning that a job's service is finished), and thus, has $Bino(n, p_S)$ distribution. Let us recall the pdf

$$X_d \begin{pmatrix} l \\ C_n^l p_S^l (1-p_S)^{n-l} \end{pmatrix}_{l=\overline{0,n}}.$$

This will help us compute the transition probability matrix.

Transition probability matrix

Suppose there are i jobs in the k-server system. Then, the number of busy servers, n, is the smaller of the number of jobs i and the total number of servers k,

$$n = \min\{i, k\}.$$

Indeed,
– for $i \le k$, the number of servers is sufficient for the current jobs, all jobs are getting service, and the number of departures X_d during the next frame is $Bino(i, p_S)$;
– for $i > k$, there are more jobs than servers. Then all k servers are busy, and the number of departures X_d during the next frame is $Bino(k, p_S)$.

Again, at most 1 job can arrive during each frame and that happens with probability p_A. Let us compute the transition probabilities

$$p_{ij} \;=\; P\big(X(t+\Delta) = j \mid X(t) = i\big).$$

We have

$$\begin{aligned}
p_{00} &= P\big(\text{ 0 arrivals }\big) \;=\; 1 - p_A, \\
p_{01} &= P\big(\text{ 1 arrival }\big) \;=\; p_A, \\
p_{i,i+1} &= P\big(\text{ 1 arrival } \cap \text{ 0 departures }\big) \;=\; p_A(1-p_S)^n, \\
p_{i,i+j} &= 0, \; \forall j > 1, \\
p_{i,i} &= P\big((\text{ 1 arrival } \cap \text{ 1 departure }) \cup (\text{ 0 arrivals } \cap \text{ 0 departures })\big) \\
&= p_A \, C_n^1 \, p_S(1-p_S)^{n-1} + (1-p_A)(1-p_S)^n, \\
p_{i,i-1} &= P\big((\text{ 1 arrival } \cap \text{ 2 departures }) \cup (\text{ 0 arrivals } \cap \text{ 1 departure })\big) \\
&= p_A \, C_n^2 \, p_S^2(1-p_S)^{n-2} + (1-p_A) \, C_n^1 \, p_S(1-p_S)^{n-1}, \\
p_{i,i-2} &= P\big((\text{ 1 arrival } \cap \text{ 3 departures }) \cup (\text{ 0 arrivals } \cap \text{ 2 departures })\big) \\
&= p_A \, C_n^3 \, p_S^3(1-p_S)^{n-3} + (1-p_A) \, C_n^2 \, p_S^2(1-p_S)^{n-2}, \\
&\dots \\
p_{i,i-n} &= P\big(\text{ 0 arrivals } \cap \; n \text{ departures }\big) \;=\; (1-p_A)p_S^n, \\
p_{i,i-j} &= 0, \; \forall j > n.
\end{aligned}$$

A transition diagram for a 2-server system is shown in Figure 1. The number of concurrent jobs can

make transitions from i to $i-2, i-1, i$ and $i+1$.

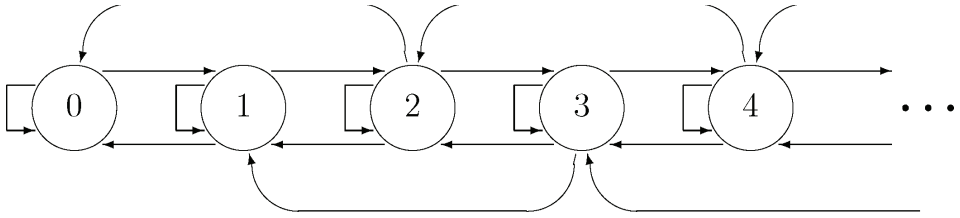


Fig. 1: Transition diagram for a B2SQS

For systems with a limited capacity $C < \infty$, the last probability changes

$$\begin{aligned} p_{C,C} &= P\big(\text{(1 arrival } \cap \text{ 1 departure) } \cup \text{ (0 arrivals } \cap \text{ 0 departures)} \\ &\quad \cup \text{ (1 arrival } \cap \text{ 0 departures)}\big) \\ &= p_A\, C_n^1\, p_S(1-p_S)^{n-1} + (1-p_A)(1-p_S)^n + p_A(1-p_S)^n \\ &= np_Ap_S(1-p_S)^{n-1} + (1-p_S)^n. \end{aligned}$$

Example 5.3. There are two customer service representatives on duty answering customers' calls. When both of them are busy, two more customers may be "on hold", but other callers will receive a busy signal. Customers call at the rate of 1 call every 5 minutes and the average service takes 8 minutes. Assuming a B2SQS with limited capacity and 1-minute frames, find
a) the steady-state distribution of the number of concurrent jobs in the system;
b) the proportion of callers who get a busy signal;
c) the percentage of time each representative is busy, if each of them takes 50% of all calls.

Solution.
a) We have $k = 2$ servers, capacity $C = 4$ and parameters

$$\begin{aligned} \lambda_A &= 1/5 \text{ / minute } = 0.2/ \text{ minute}, \\ \lambda_S &= 1/8 \text{ / minute } = 0.125/ \text{ minute}, \\ \Delta &= 1 \text{ minute}. \end{aligned}$$

So,

$$\begin{aligned} p_A &= \lambda_A \Delta = 0.2, \quad 1 - p_A = 0.8, \\ p_S &= \lambda_S \Delta = 0.125, \quad 1 - p_S = 0.875. \end{aligned}$$

There are 5 states, $\{0, 1, 2, 3, 4\}$. The transition probability matrix is

$$P = \begin{bmatrix} 0.8000 & 0.2000 & 0 & 0 & 0 \\ 0.1000 & 0.7250 & 0.1750 & 0 & 0 \\ 0.0125 & 0.1781 & 0.6562 & 0.1531 & 0 \\ 0 & 0.0125 & 0.1781 & 0.6562 & 0.1531 \\ 0 & 0 & 0.0125 & 0.1781 & 0.8094 \end{bmatrix}.$$

The steady-state distribution is

$$\pi = [\pi_0 \ \pi_1 \ \pi_2 \ \pi_3 \ \pi_4] = [0.1527 \quad 0.2753 \quad 0.2407 \quad 0.1837 \quad 0.1476].$$

b) Callers hear a busy signal when the system is full, i.e. $X = C = 4$. So that probability is

$$P(X = C) = \pi_4 = 0.1476.$$

c) Each representative is busy when there are $2, 3$ or 4 jobs in the system, plus a half of the time when there is 1 job (because there is a 50% chance that the other representative handles this job). This totals

$$\pi_2 + \pi_3 + \pi_4 + 0.5\pi_1 = 0.709 \text{ or } 70.9\% \text{ of the time.}$$

■

5.2 $M/M/k$ Queuing Systems

An M / M / k queuing system is a multiserver extension of an M / M / 1 system.

Definition 5.1. *An **M / M / k queuing process** is a continuous-time Markov queuing process with the following characteristics:*

- k *servers;*
- *unlimited capacity;*
- *Exponential interarrival times with arrival rate* λ_A*;*
- *Exponential service times with service rate* λ_S*;*
- *independent service and arrival times, independent service times of all servers.*

Once again, we use the same approach as before, move from the discrete-time BkSQP to the continuous-time M / M / k process by letting the frame size $\Delta \to 0$. Recall that

$$\begin{aligned} p_A &= \lambda_A \Delta, \\ p_S &= \lambda_S \Delta. \end{aligned}$$

For very small Δ, we neglect terms of the form Δ^l, for $l \geq 2$, so the transition probabilities for a BkSQP become:

$$\begin{aligned} p_{i,i+1} &= p_A(1-p_S)^n = \lambda_A\Delta(1-\lambda_S\Delta)^n \approx \lambda_A\Delta = \underline{p_A} \\ p_{i,i} &= p_A\, C_n^1\, p_S(1-p_S)^{n-1} + (1-p_A)(1-p_S)^n \\ &= \lambda_A\Delta\, n\, \lambda_S\Delta(1-\lambda_S\Delta)^{n-1} + (1-\lambda_A\Delta)(1-\lambda_S\Delta)^n \\ &\approx (1-\lambda_A\Delta)\left(1 - C_n^1\lambda_S\Delta + \ldots\right) \\ &\approx 1-\lambda_A\Delta - n\,\lambda_S\Delta = \underline{1-p_A-np_S} \\ p_{i,i-1} &= p_A\, C_n^2\, p_S^2(1-p_S)^{n-2} + (1-p_A)\, C_n^1\, p_S(1-p_S)^{n-1} \\ &= \lambda_A\Delta\frac{n(n-1)}{2}\,(\lambda_S\Delta)^2\,(1-\lambda_S\Delta)^{n-2} \\ &+ (1-\lambda_A\Delta)\, n\, \lambda_S\Delta(1-\lambda_S\Delta)^{n-1} \\ &\approx n\,\lambda_S\Delta = \underline{np_S} \\ p_{i,j} &= 0,\ \forall j \neq i-1, i, i+1. \end{aligned}$$

Recall that $n = \min\{i, k\}$ is the number of jobs receiving service among the total of i jobs in the system. Also, since Δ is very small, we ignored terms proportional to Δ^2, Δ^3, etc. Then, no more than one event, arrival or departure, may occur during each frame. Probability of more than one event is of the order $O(\Delta^2)$. Changing the number of jobs by 2 requires at least 2 events, and thus, such changes cannot occur during one frame. At the same time, transition from i to $i-1$ may be caused by a departure of any one of the n currently served jobs. This is why we have the departure probability p_S multiplied by n.

So, the transition probability matrix is

$$P \approx \begin{bmatrix} 1-p_A & p_A & 0 & 0 & \dots & 0 & \dots \\ p_S & 1-p_A-p_S & p_A & 0 & \dots & 0 & \dots \\ 0 & 2p_S & 1-p_A-2p_S & p_A & \dots & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 & \dots \\ 0 & 0 & \dots & kp_S & 1-p_A-kp_S & 0 & \dots \\ 0 & 0 & 0 & 0 & kp_S & 1-p_A-kp_S & \dots \\ \vdots & \vdots & \vdots & \vdots & & \ddots & \end{bmatrix} \tag{5.1}$$

For example, for $k = 3$ servers, the transition probability matrix is

$$P \approx \begin{bmatrix} 1-p_A & p_A & 0 & 0 & 0 & 0 & \dots \\ p_S & 1-p_A-p_S & p_A & 0 & 0 & 0 & \dots \\ 0 & 2p_S & 1-p_A-2p_S & p_A & 0 & 0 & \dots \\ 0 & 0 & 3p_S & 1-p_A-3p_S & p_A & 0 & \dots \\ 0 & 0 & 0 & 3p_S & 1-p_A-3p_S & p_A & \dots \\ 0 & 0 & 0 & 0 & 3p_S & 1-p_A-3p_S & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Next we find the steady-state distribution, as usually, from

$$\begin{cases} \pi P &= \pi \\ \sum_{i=0}^{\infty} \pi_i &= 1, \end{cases}$$

again, a system of infinitely many equations with infinitely many unknowns.
The first balance equation is

$$\begin{aligned} [\pi_0 \ \pi_1 \ \pi_2 \ \ldots] \cdot \begin{bmatrix} 1-p_A \\ p_S \\ 0 \\ \vdots \end{bmatrix} &= \pi_0, \text{ i.e.} \\ (1-p_A)\pi_0 + p_S\pi_1 &= \pi_0, \text{ i.e.} \\ -p_A\pi_0 + p_S\pi_1 &= 0, \text{ i.e.} \\ p_S\pi_1 &= p_A\pi_0, \text{ i.e.} \\ \lambda_S\pi_1 &= \lambda_A\pi_0. \end{aligned}$$

So,

$$\pi_1 = \frac{\lambda_A}{\lambda_S}\pi_0 = r\pi_0. \tag{5.2}$$

The second balance equation is

$$\begin{aligned} [\pi_0 \ \pi_1 \ \pi_2 \ \ldots] \cdot \begin{bmatrix} p_A \\ 1-p_A-p_S \\ 2p_S \\ 0 \\ \vdots \end{bmatrix} &= \pi_1, \text{ i.e.} \\ p_A\pi_0 + (1-p_A-p_S)\pi_1 + 2p_S\pi_2 &= \pi_1, \text{ i.e.} \\ p_A\pi_0 - p_A\pi_1 - p_S\pi_1 + 2p_S\pi_2 &= 0, \text{ i.e. (since } p_A\pi_0 = p_S\pi_1) \\ 2p_S\pi_2 &= p_A\pi_1, \text{ i.e.} \\ 2\lambda_S\pi_2 &= \lambda_A\pi_1. \end{aligned}$$

Thus, we get

$$\pi_2 \;=\; \frac{1}{2}\frac{\lambda_A}{\lambda_S}\pi_1 \;=\; \frac{1}{2}r\pi_1 \;=\; \frac{1}{2}r^2\pi_0. \tag{5.3}$$

The third balance equation is

$$\begin{aligned}
[\pi_0\ \ \pi_1\ \ \pi_2\ \ \pi_3\ \ \ldots] \cdot \begin{bmatrix} 0 \\ p_A \\ 1-p_A-2p_S \\ 3p_S \\ 0 \\ \vdots \end{bmatrix} &= \pi_2, \text{ i.e.} \\
p_A\pi_1 + (1-p_A-2p_S)\pi_2 + 3p_S\pi_3 &= \pi_2, \text{ i.e.} \\
p_A\pi_1 - p_A\pi_2 - 2p_S\pi_2 + 3p_S\pi_3 &= 0, \text{ i.e. (since } p_A\pi_1 = 2p_S\pi_2) \\
3p_S\pi_3 &= p_A\pi_2, \text{ i.e.} \\
3\lambda_S\pi_3 &= \lambda_A\pi_2.
\end{aligned}$$

So,

$$\pi_3 \;=\; \frac{1}{3}\frac{\lambda_A}{\lambda_S}\pi_2 \;=\; \frac{1}{3}r\pi_2 \;=\; \frac{1}{2\cdot 3}r^3\pi_0 \;=\; \frac{1}{3!}r^3\pi_0. \tag{5.4}$$

We see a pattern forming. The k^{th} balance equation will yield

$$\pi_k \;=\; \frac{1}{k}r\pi_{k-1} \;=\; \frac{1}{k!}r^k\pi_0. \tag{5.5}$$

Then things change. Let us see the the $(k+1)^{\text{st}}$ equation.

$$
\begin{aligned}
[\ldots\ \ \pi_{k-1}\ \ \pi_k\ \ \pi_{k+1}\ \ \ldots] \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ p_A \\ 1 - p_A - kp_S \\ kp_S \\ 0 \\ \vdots \end{bmatrix} &= \pi_k, \text{ i.e.} \\
p_A\pi_{k-1} + (1 - p_A - kp_S)\pi_k + kp_S\pi_{k+1} &= \pi_k, \text{ i.e.} \\
p_A\pi_{k-1} - p_A\pi_k - kp_S\pi_k + kp_S\pi_{k+1} &= 0, \text{ i.e. (since } p_A\pi_{k-1} = kp_S\pi_k) \\
kp_S\pi_{k+1} &= p_A\pi_k, \text{ i.e.} \\
k\lambda_S\pi_{k+1} &= \lambda_A\pi_k,
\end{aligned}
$$

which yields

$$
\pi_{k+1} = \frac{1}{k} r\pi_k = \left(\frac{r}{k}\right) \frac{r^k}{k!}\pi_0. \tag{5.6}
$$

All the rest of the equations will be of the same form

$$
\begin{aligned}
\pi_{k+2} &= \frac{1}{k} r\pi_{k+1} = \left(\frac{r}{k}\right)^2 \frac{r^k}{k!}\pi_0 \\
&\ldots
\end{aligned} \tag{5.7}
$$

Now we substitute them all in the normalizing equation $\sum_{i=0}^{\infty} \pi_i = 1$. We get

$$
\begin{aligned}
1 &= \pi_0 + \pi_1 + \dots \\
&= \left(\pi_0 + \pi_1 + \dots + \pi_{k-1}\right) + \left(\pi_k + \pi_{k+1} + \dots\right) \\
&= \pi_0 \left[\left(1 + r + \frac{r^2}{2!} + \dots + \frac{r^{k-1}}{(k-1)!}\right) + \frac{r^k}{k!}\left(1 + \frac{r}{k} + \left(\frac{r}{k}\right)^2 + \dots\right)\right] \\
&= \pi_0 \left(\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{k!} \cdot \frac{1}{1 - r/k}\right) \\
&= \pi_0 \left(\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{k!(1 - r/k)}\right),
\end{aligned}
$$

where, in the last part, the Geometric series $\displaystyle\sum_{i=0}^{\infty} \left(\frac{r}{k}\right)^i$ is convergent and equal to $\dfrac{1}{1 - r/k}$, if the ratio $r/k < 1$, i.e. $r < k$. So, the M / M / k steady-state distribution of number of jobs has pdf

$$
\begin{aligned}
\pi_0 &= P(X = 0) = \frac{1}{\displaystyle\sum_{i=0}^{k-1} \frac{r^i}{i!} + \frac{r^k}{k!(1 - r/k)}}, \\
\pi_x &= P(X = x) = \begin{cases} \dfrac{r^x}{x!}\pi_0, & \text{for } x < k \\ \dfrac{r^k}{k!}\pi_0 \left(\dfrac{r}{k}\right)^{x-k}, & \text{for } x \geq k \end{cases},
\end{aligned}
\tag{5.8}
$$

provided that

$$
r = \frac{\lambda_A}{\lambda_S} < k.
$$

Example 5.2. Consider again Example 4.5 (and 4.6) from Lecture 9 about the message transmission center (*Messages arrive to a communication center at random times according to a Poisson process, with an average of* 5 *messages per minute. They are transmitted through a single channel in the order they were received. On average, it takes* 10 *seconds to transmit a message*). Recall that when the number of customers increased by 10%, all the parameters of the system increased significantly, some of them even by more than 100%. Suppose now that the arrival rate has doubled to 10 messages per minute. On average, it still takes 10 seconds to transmit a message, but assume

that 2 additional channels are built with the same parameters as the first channel. Evaluate the new system's performance. What percentage of messages will be sent immediately, with no waiting time?

Solution. This is now an M / M / 3 system with

$$\begin{aligned} \lambda_A &= 10 \text{ / minute} = 1/6 \text{ / second}, \\ \mu_S &= 10 \text{ seconds} = \frac{1}{6} \text{ minutes}, \\ \lambda_S &= 6 \text{ / minute}, \\ r &= \frac{10}{6} = \frac{5}{3} = 1.667 > 1, \text{ but } r < 3. \end{aligned}$$

Before we proceed with the computation of $E(X)$, let us recall a few formulas related to the Geometric series:

- the Geometric series $\sum_{i=0}^{\infty} a_0 q^i$ is convergent if the ratio $|q| < 1$ and its sum is equal to

$$\sum_{i=0}^{\infty} a_0 q^i = \frac{a_0}{1-q};$$

- under the same conditions (differentiating the equation above with respect to q), the following series is also convergent

$$\sum_{i=0}^{\infty} a_0 i\, q^i = \frac{a_0 q}{(1-q)^2}.$$

The steady-state distribution, by (5.8) is given by

$$\pi_0 = \frac{1}{\sum_{i=0}^{2} \frac{r^i}{i!} + \frac{r^3}{3!(1-r/3)}} = \frac{1}{1 + r + \frac{r^2}{2!} + \frac{r^3}{3!(1-r/3)}} = 0.1727$$

and

$$\pi_x \;=\; \begin{cases} \dfrac{r^x}{x!}\pi_0, & \text{for } x = 1, 2 \\ \\ \dfrac{r^3}{3!}\pi_0 \left(\dfrac{r}{3}\right)^{x-3}, & \text{for } x = 3, 4, \ldots \end{cases} .$$

Then

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\pi_x \;=\; \sum_{x=0}^{2} x\pi_x + \sum_{x=3}^{\infty} x\pi_x \\ &= \pi_0 \left(0 + 1 \cdot r + 2 \cdot \frac{r^2}{2}\right) + \pi_0 \frac{r^3}{3!} \sum_{x=3}^{\infty} x \left(\frac{r}{3}\right)^{x-3} \\ &= \pi_0 \left(r + r^2\right) + \pi_0 \frac{r^3}{3!} \sum_{j=0}^{\infty} (j+3) \left(\frac{r}{3}\right)^{j} \\ &= \pi_0 \left(r + r^2\right) + \pi_0 \frac{r^3}{3!} \left[\sum_{j=0}^{\infty} j \left(\frac{r}{3}\right)^{j} + 3 \sum_{j=0}^{\infty} \left(\frac{r}{3}\right)^{j}\right] \\ &= \pi_0 \left(r + r^2\right) + \pi_0 \frac{r^3}{6} \left[\frac{r/3}{(1 - r/3)^2} + 3 \frac{1}{1 - r/3}\right] \\ &= \pi_0 \left(r + r^2\right) + \frac{\pi_0 r^3 (9 - 2r)}{2(3 - r)^2} \\ &= \pi_0 \left(r + r^2 + \frac{r^3 (9 - 2r)}{2(3 - r)^2}\right) \\ &= 2.0418. \end{aligned}$$

Thus, the average number of messages stored in the system at any time is

$$E(X) \;=\; 2.0418.$$

By Little's law, the total time from arrival until the end of transmission has an average of

$$E(R) \;=\; E(X)/\lambda_A \;=\; 0.20418 \,\text{minutes} \;=\; 12.2508 \,\text{seconds}.$$

When a message arrives to the center, its average waiting time until transmission is

$$E(W) \;=\; E(R) - E(S) \;=\; E(R) - \mu_S \;=\; 12.2508 - 10 \;=\; 2.2508 \,\text{seconds}.$$

Then, using Little's law again, the average number of messages waiting to be transmitted is

$$E(X_w) \;=\; \lambda_A E(W) \;=\; 1/6 \cdot 2.2508 \;=\; 0.3751.$$

Finally, the average number of messages being transmitted is

$$E(X_s) \;=\; E(X) - E(X_w) \;=\; 2.0418 - 0.3751 \;=\; 1.6667 \;=\; r.$$

Alternatively, for the last one, by Little's law,

$$E(X_s) \;=\; \lambda_A E(S) \;=\; \lambda_A \mu_S \;=\; \frac{\lambda_A}{\lambda_S} \;=\; r,$$

just like in the case of an M / M / 1 system.

To answer the last question, a message does not wait at all if there is an idle server (channel) to service (transmit) it. That happens when the number of jobs in the system is *less* than the number of servers $k = 3$. So,

$$\begin{aligned} P(W=0) &= P(X<3) \;=\; P(X=0 \text{ or } X=1 \text{ or } X=2) \\ &= \pi_0 + \pi_1 + \pi_2 \\ &= \pi_0 \left(1 + r + \frac{r^2}{2!}\right) \\ &= \frac{73}{18}\pi_0 \;=\; 0.7004. \end{aligned}$$

Or, we can directly compute

$$\begin{aligned} \pi_1 &= \frac{r}{1!}\pi_0 \;=\; 0.2878, \\ \pi_2 &= \frac{r^2}{2!}\pi_0 \;=\; 0.2398, \\ P(W=0) &= \pi_0 + \pi_1 + \pi_2 \;=\; 0.7004. \end{aligned}$$

That means that 70% of the messages are transmitted immediately, with no waiting time.

■

6 $M/M/\infty$ Queuing Systems

Let us now consider an unlimited number of servers $k = \infty$. That *completely* eliminates the waiting time. Whenever a job arrives, there will always be servers available to handle it and thus,

$$\begin{aligned} X &= X_s, \text{ the number of jobs in the system is the number of jobs receiving service,} \\ R &= S, \text{ response time consists of service time only,} \\ X_w &= 0, \text{ no jobs waiting in queue,} \\ W &= 0, \text{ no waiting time.} \end{aligned}$$

All the formulas we derived for M / M / k systems apply to M / M / ∞ systems, by letting the number of servers $k \to \infty$. Let us see what we get.
First off, the number of jobs will always be less than the number of servers ($i < k$), so we always have $n = i$. That is, with i jobs in the system, exactly i servers are busy.
The transition probability matrix for the number of jobs in the system, X, is given by

$$P = \begin{bmatrix} 1-p_A & p_A & 0 & 0 & 0 & 0 & \dots \\ p_S & 1-p_A-p_S & p_A & 0 & 0 & 0 & \dots \\ 0 & 2p_S & 1-p_A-2p_S & p_A & 0 & 0 & \dots \\ 0 & 0 & 3p_S & 1-p_A-3p_S & p_A & 0 & \dots \\ 0 & 0 & 0 & 4p_S & 1-p_A-4p_S & p_A & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Let us see what becomes the steady-state distribution. The first component, π_0, becomes

$$\pi_0 = P(X=0) = \frac{1}{\sum_{i=0}^{\infty} \frac{r^i}{i!} + \lim_{k\to\infty} \left(\frac{r^k}{k!} \cdot \frac{1}{1-r/k} \right)}.$$

Now,

$$\lim_{k\to\infty}\left(\frac{r^k}{k!}\cdot\frac{1}{1-r/k}\right) \;=\; \lim_{k\to\infty}\frac{r^k}{k!} \;=\; 0,$$

because the factorial converges faster to ∞ than the exponential function. The other term is the Taylor series of the function e^r, so the steady-state distribution is

$$\begin{aligned}\pi_0 &= e^{-r},\\ \pi_i &= \frac{r^i}{i!}e^{-r},\ \ \forall i\geq 1.\end{aligned} \tag{6.1}$$

So the pdf of $X(t)$, the number of concurrent jobs in an M / M / ∞ system at time t, is

$$X(t)\ \begin{pmatrix} i \\ \dfrac{r^i}{i!}e^{-r}\end{pmatrix}_{i=0,1,\ldots}, \tag{6.2}$$

a Poisson distribution with parameter $r = \dfrac{\lambda_A}{\lambda_S}$ (which can be arbitrarily large), with mean and variance

$$E(X) \;=\; V(X) \;=\; r. \tag{6.3}$$

Remark 6.1. Clearly, nobody can physically build an *infinite* number of devices. In practice, having an unlimited number of servers simply means that any number of concurrent jobs can be served simultaneously. Example: internet service providers or telephone companies (which allow virtually any number of concurrent connections), an unlimited number of people can watch a TV channel or listen to a radio station, etc. A model with infinitely many servers is a reasonable approximation for a system where jobs typically don't wait and get their service immediately. This may be appropriate for a computer server, a grocery store, Facebook, etc.

Example 6.2. A certain powerful server can afford practically any number of concurrent users. Users connect to the server at random times, every 3 minutes, on the average, according to a Poisson counting process. Each user spends an Exponential amount of time on the server with an average of 1 hour and disconnects from it, independently of other users. Find

a) the fraction of time when no users are connected to the server;

b) the expected number of concurrent users at any time;

c) if a message is sent to all users, the probability that 15 or more users will receive this message immediately.

Solution. This fits the description of an M / M / ∞ system with

$$\begin{aligned}
\mu_A &= 3 \text{ minutes, so} \\
\lambda_A &= 1/\mu_A = 1/3 \text{ / minute,} \\
\mu_S &= 1 \text{ hour} = 60 \text{ minutes, so} \\
\lambda_S &= 1/\mu_S = 1/60 \text{ / minute,} \\
r &= \frac{\lambda_A}{\lambda_S} = \frac{1/3}{1/60} = 20.
\end{aligned}$$

The number of concurrent users has $Poiss(20)$ distribution.

a)

$$P(X = 0) = \pi_0 = e^{-20} = 2.06 \cdot 10^{-9} = 0.$$

This server is practically *never* idle.

b) The expected number of concurrent users is

$$E(X) = r = 20 \text{ users.}$$

Also, if an urgent message is sent to all the users, then 20 users, on the average, will see it immediately.

c) Fifteen or more users will receive a message immediately if 15 or more users are connected, so, with probability

$$\begin{aligned}
P(X \geq 15) &= 1 - P(X < 15) = 1 - P(X \leq 14) \\
&= 1 - poisscdf(14, 20) = 0.8951.
\end{aligned}$$

■

7 Simulation of Queuing Systems

We developed a theory and understood how to analyze and evaluate rather basic queuing systems: Bernoulli and M / M / k. Most of the results were obtained from the Markov property of the considered queuing processes. For these systems, we derived a steady-state distribution of the number of concurrent jobs and computed the vital performance characteristics from it.

In practice, however, many queuing systems have a rather complex structure. Jobs may arrive according to a non-Poisson process, often the rate of arrivals changes during the day (there is a rush hour on highways or on the internet, etc.). Service times may have different distributions and they are not always memoryless, thus the Markov property may not be satisfied. The number of servers may also change during the day (additional servers may turn on during rush hours). Some customers may get dissatisfied with a long waiting time and quit in the middle of their queue. And so on. Queuing theory does not cover all the possible situations. On the other hand, we can simulate the behavior of almost any queuing system and study its properties by Monte Carlo methods.

A queuing system is Markov only when its interarrival and service times are memoryless. Then the future can be predicted from the present without relying on the past. It can be simulated using the algorithm given for Markov chains (Algorithm 2.13 in Lecture 5). To study long-term characteristics of a queuing system, the initial distribution of X_0 typically does not matter, so we may start this algorithm with 0 jobs in the system and then "switch on" the servers.
Even when the system is Markov, some interesting characteristics do not follow from its steady-state distribution directly, but they can be estimated from a Monte Carlo study.

Performance of more complicated and advanced queuing systems can be evaluated by Monte Carlo methods. One needs to simulate arrivals of jobs, assignment of servers and service times and to keep track of all variables of interest. Monte Carlo methods of Chapter 2 let us simulate and evaluate rather complex queuing systems far beyond Bernoulli and M / M / k. As long as we know the distributions of interarrival and service times, we can generate the processes of arrivals and services. To assign jobs to servers, we keep track of servers that are available each time when a new job arrives. When all the servers are busy, the new job will enter a queue. As we simulate the work of a queuing system, we keep records of events and variables that are of interest to us. After a large number of Monte Carlo runs, we average our records in order to estimate probabilities by long-run proportions and expected values by long-run averages.