582692 Randomized Algorithms II

Spring 2013, period IV
Jyrki Kivinen

Position of the course in the studies

- 4 credits
- advanced course (syventävät opinnot) in algorithms and machine learning
- prerequisites: basic understanding of probabilities and design and analysis of algorithms
- covers application of probabilities in designing and analysing algorithms
- continuation from *Randomized algorithms I* which however is not a prerequisite
- applications of probability theory figure prominently also on a number of courses about machine learning
- theory of probability is the topic for many courses in mathematics
- this course is mainly theoretical from a computer science point of view, fairly application-oriented from maths point of view

Passing the course, grading

Maximum score 60 points:

- course exam 48 points
- homework 12 points

Minimum passing score is about 30 points, requirement for best grade about 50 points.

Homework sessions begin on the second week of lectures. Solutions to homework problems are turned in **in writing** before the session. Details and deadlines will be announced on the course web page.

Each problem is graded from 0 to 3:

1 a reasonable attempt
2 work in the right direction, largely successful
3 seems to be more or less correct.

The homework points will be scaled to course points as follows:

- 0 % of the maximum gives 0 points
- 80 % or more of the maximum gives 12 points
- linear interpolation in between.

Material

The course is based on the textbook

M. Mitzenmacher, E. Upfal: Probability and Computing

to which the students are expected to have access. We will cover Chapters 7, 8 and 10, and possibly parts of 11 and 12.

(Chapters 1–6 were covered in *Randomized Algorithms I*. The topic of Chapter 10, which we skip, is covered in the course *Information-Theoretic Modeling*.)

References to the textbook in these notes are in style [M&U Thm 3.2].

The lecture notes will appear on the course home page but are not intended to cover the material in full.

Also the homework will be based on problems from the textbook.

Contents of the course

Main topics covered in the preceding course *Randomized algorithms I* included

1. theory of probability (quick refresher)
2. discrete random variables (quick refresher)
3. moments of a random variable
4. Chernoff bounds
5. balls and bins
6. "the probabilistic method"

Here in *Randomized algorithms II* we will continue with

1. Markov chains
2. continuous random variables, Poisson processes
3. Monte Carlo methods
4. (martingales, if there's time).

1. Markov Chains

Markov chains are stochastic processes with various uses:

1. many random phenomena, such as queueing, are naturally modelled as Markov chains

2. some basic techniques of randomized algorithms, such as randomized local search and simulated annealing, can be analysed using Markov chains

3. Markov chains can be used to generate random samples from complicated distributions.

In modelling tasks mentioned in part 1, we are often mainly interested in the stationary distribution towards which the process converges.

In parts 2 and 3 we also need to consider how long a "burn-in" we need to get reasonably close to the stationary distibution.

Generalizations such as partially observed Markov decision processes (POMDPs) are important in reinforcement learning.

Fundamental concepts [M&U Section 7.1]

A stochastic process is a sequence of random variables $\mathbf{X} = (X(t))_{t \in T}$. Often we denote $X(t)$ by X_t. Often t is interpreted to be a point of time. The value of X_t is then called the state of the process at time t.

If for all t the range of the random variable X_t is countable, then $\mathbf{X}$ is a discrete space process. If T is countable, then $\mathbf{X}$ is a discrete time process.

A discrete time process (X_t) is a Markov chain if

$$\Pr(X_t = a_t \mid X_0 = a_0, \ldots, X_{t-1} = a_{t-1}) = \Pr(X_t = a_t \mid X_{t-1} = a_{t-1})$$

for all t and (a_i). That is, to predict the next state of the process, knowing the full history of the process gives no extra information compared to just knowing the present state. This does **not** mean that X_t would be independent of $X_0, \ldots, X_{t-2}$. (It is conditionally independent given X_{t-1}).

If additionally the transition probabilities are same at all times, that is,

$$\Pr(X_t = a \mid X_{t-1} = b) = \Pr(X_{t'} = a \mid X_{t'-1} = b)$$

for all t, t', a, b, the chain is homogenous. Here we always assume our chains to be homogenous.

For simplicity, we assume that the state space of a discrete space chain is $\{0, \ldots, n\}$ for some n, or $\mathbb{N}$ if it is infinite. We will later consider continuous time chains, but for now we assume that time is discrete and $T = \mathbb{N}$.

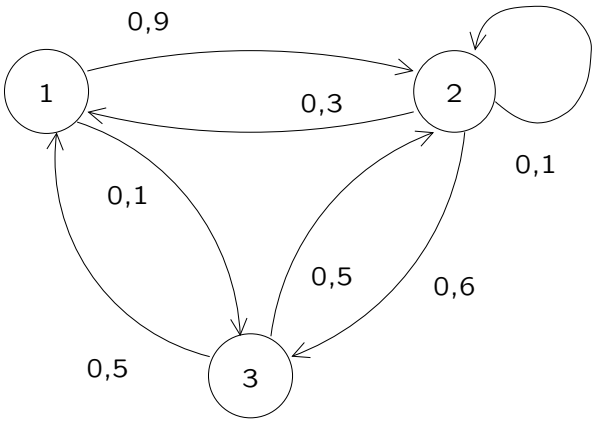
The chain can be defined by giving the distribution of the initial state X_0 and the transition matrix $\mathbf{P}$ where

$$P_{i,j} = \Pr(X_t = j \mid X_{t-1} = i).$$

This implies that each row of the transition matrix sums to 1.

It may be useful to visualize the chain as a directed graph where edge (i, j) has weight $P_{i,j}$. Edges with weight zero can be left out of the picture.

Example 1.1: A three-state Markov chain as a graph and a matrix:



$$P = \begin{pmatrix} 0 & 9/10 & 1/10 \\ 3/10 & 1/10 & 6/10 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

A complete description should also include the initial distribution, but we are usually interested in properties that do not depend on the initial distribution.

We write

$$P_{i,j}^m = \Pr(X_{t+m} = j \mid X_t = i).$$

That is,

$$P_{i,j}^m = \sum_k P_{i,k} P_{k,j}^{m-1}$$

from which induction yields

$$P_{i,j}^m = (\mathbf{P}^m)_{i,j}$$

where $\mathbf{P}^m$ is the m-fold matrix product of P with itself.

Defining a vector $\boldsymbol{p}(t)$ as $p_i(t) = \Pr(X_t = i)$ we then get

$$\boldsymbol{p}(t+m) = \boldsymbol{p}(t)\mathbf{P}^m.$$

Example: 2-SAT [M&U Section 7.1.1]

We consider a randomized algorithm for satisfiability of Boolean 2-CNF formulas (conjunctions of clauses where each clause has exactly two literals). The problem is well known to be solvable in deterministic polynomial time, whereas k-SAT for $k \geq 3$ is NP-complete.

Let n be the number of variables. Hence, there can be $O(n^2)$ clauses. We consider the following algorithm, where the parameter m regulates the success probility.

1. Assign random values to the variables.
2. Repeat $2mn^2$ times or until the formula is satisfied:
   (a) Choose a random clause that is not satisfied.
   (b) Choose randomly one literal from the clause
   and change its value (thus satisfying the clause).
3. If the formula is satisfied, return the value assignment.
   Otherwise return "not satisfiable."

One iteration in part 2 can clearly be done in polynomial time. We now consider the number of iterations needed to find a satisfying assignment.

Assume that the formula is satisfiable, and S is a satisfying assignment for the variables. Let A_i be the assignment of the algorithm after iteration i, and let X_i be the number of variables on which S and A_i agree.

Hence, a sufficient condition for the algorithm giving a correct answer at iteration i is that $X_i = n$. To simplify notation, in this case we also define $X_j = n$ for $j > i$.

Clearly $\Pr(X_{i+1} = 1 \mid X_i = 0) = 1$. Since S satisfied the clause chosen in 2(a) in iteration i, but A_i does not, the swap in 2(b) has probability at least $1/2$ of changing one variable to the value that agrees with S. For $i \leq j < n$ we have therefore

$$\begin{aligned} \Pr(X_{i+1} = j+1 \mid X_i = j) &\geq 1/2 \\ \Pr(X_{i+1} = j-1 \mid X_i = j) &\leq 1/2. \end{aligned}$$

The random variables X_i do not constitute a Markov chain, because X_{i+1} depends on which particular clauses are unsatisfied, and this in turn depends on all preceding choices, not just the number X_i.

We define a Markov chain (Y_t) where $Y_0 = X_0$ and

$$\begin{aligned} \Pr(Y_{i+1} = 1 \mid Y_i = 0) &= 1 \\ \Pr(Y_{i+1} = j+1 \mid Y_i = j) &= 1/2 \quad \text{if } j \geq 1 \\ \Pr(Y_{i+1} = j-1 \mid Y_i = j) &= 1/2 \quad \text{if } j \geq 1. \end{aligned}$$

Intuitively it seems clear that X_i grows at least as fast as Y_i. To make this more precise, consider defining $Y_i = X_i - R_i$, where

- We initialize $R_0 = 0$.
- If $R_i = X_i \geq 1$, then $R_{i+1} = X_{i+1} - 1$.
- Otherwise if $X_i = n$, then $R_{i+1} = R_i + 1$ with probability $1/2$ and $R_{i+1} = R_i - 1$ with probability $1/2$.
- Otherwise if $X_i \geq 1$ and both literals of the clause chosen in iteration $i+1$ get different values in assignments A_i ja S, then $R_{i+1} = R_i$ with probability $1/2$ and $R_{i+1} = R_i + 2$ with probability $1/2$.
- Otherwise $R_{i+1} = R_i$.

The rules for updating R_i are chosen such that $Y_i = X_i - R_i$ is a Markov chain with the desired transition probabilities. Additionally, if $Y_i \geq n$, then $X_i = n$ and a solution has been found.

We want to know whether $X_i = n$ holds for some $i \leq 2mn^2$. A sufficient condition for this is $Y_i \geq n$ for some $i \leq 2mn^2$.

Fix some $0 \leq j \leq n$ and time t. The value t must be large enough that Y_t may have value j; otherwise the precise value is unimportant. Let Z_j be the number of iterations for the algorithm to terminate starting from time t with $Y_t = j$, and $h_j = \mathbf{E}[Z_j]$. Then

$$h_j = \mathbf{E}[\min\{\, i \mid Y_{t+i} \geq n \,\} \mid Y_t = j].$$

Clearly $h_n = 0$. On the other hand, $h_0 = h_1 + 1$, because starting from $Y_t = 0$ we always get $Y_{t+1} = 1$. For all $1 \leq j < n$ we have

$$\mathbf{E}[Z_j] = \mathbf{E}\left[\tfrac{1}{2}(1 + Z_{j-1}) + \tfrac{1}{2}(1 + Z_{j+1})\right].$$

By linearity of expectation, this implies

$$h_j = \tfrac{1}{2}h_{j-1} + \tfrac{1}{2}h_{j+1} + 1.$$

We have a system of $n+1$ linear equations:

$$\begin{aligned} h_n &= 0 \\ h_j &= \tfrac{1}{2}h_{j-1} + \tfrac{1}{2}h_{j+1} + 1, \quad 1 \le j \le n-1 \\ h_0 &= h_1 + 1. \end{aligned}$$

Write $\Delta_j = h_j - h_{j-1}$, so we get

$$\Delta_j = \Delta_{j+1} + 2, \quad 1 \le j \le n-1$$

and $\Delta_1 = -1$. Therefore,

$$\Delta_j = 1 - 2j.$$

Thus

$$0 = h_n = h_0 + \sum_{j=1}^{n} \Delta_j = h_0 + n - 2\frac{n(n+1)}{2} = h_0 - n^2,$$

so $h_0 = n^2$. Clearly $h_j \le n^2$ for all $j \ge 0$.

We have proved the following.

Lemma 1.2 [M&U Lemma 7.1]**:** If a 2-CNF formula is satisfiable, then the algorithm given above makes in expectation at most n^2 iterations before finding a satisfying assignment. □

This implies the actual result.

Theorem 1.3 [M&U Thm 7.2]**:** If a 2-CNF formula is not satisfiable, the algorithm always give the correct answer. If the formula is satisfiable, the algorithm produces a satisfying assignment with probability at least $1 - 2^{-m}$.

Proof: The claim for non-satisfied formulas clearly holds, so consider a satisfiable formula.

We split the $2mn^2$ iterations of the algorithm into blocks of $2n^2$ iterations. For a given block i, let Z be the number of iteration from the start of the block until a solution is found. By the preceding lemma and Markov's inequality, the probability for failing to find a solution within block i is at most

$$\Pr(Z \geq 2n^2) \leq \frac{\mathbf{E}[Z]}{2n^2} = \frac{1}{2}.$$

The probability that no block includes a solution is at most $(1/2)^m$. □

Example: 3-SAT [M&U Section 7.1.2]

Unlike 2-SAT, we know that 3-SAT is NP-complete, so we don't expect to find an efficient randomized algorithm. However, we can use the ideas from our 2-SAT algorithm to get something that is much better than brute force.

We start with a straightforward attempt.

1. Assign all variables arbitrarily.
2. Repeat m times or until the formula is satisfied:
   (a) Choose a random unsatisfied clause.
   (b) Choose a random literal from the clause
   and change its value (making the clause satisfied).
3. If the formula is satisfied, return the assignment.
   Otherwise return "not satisfiable."

We analyse as previously. Let S be a satisfying assignment, A_i the assignment of the algorithm after iteration i, and X_i the number of variables that are assigned the same value in S and A_i.

Since A_i does not satisfy any of the literals in the clause chosen in 2(a), but S satisfies at least one, we have

$$\begin{aligned} \Pr(X_{i+1} = j+1 \mid X_i = j) &\geq 1/3 \\ \Pr(X_{i+1} = j-1 \mid X_i = j) &\leq 2/3. \end{aligned}$$

Again we estimate X_i by a Markov chain:

$$\begin{aligned} \Pr(Y_{i+1} = 1 \mid Y_i = 0) &= 1 \\ \Pr(Y_{i+1} = j+1 \mid Y_i = j) &= 1/3 \quad \text{if } j \geq 1 \\ \Pr(Y_{i+1} = j-1 \mid Y_i = j) &= 2/3 \quad \text{if } j \geq 1. \end{aligned}$$

Unfortunately we can see that Y is more likely to decrease than to increase.

Nevertheless, we can get some kind of an estimate. Again, let h_j be the expected number of iterations to find a satisfying assignment, if initially there are j variables whose assignment agrees with S. As previously, we get

$$\begin{aligned} h_n &= 0 \\ h_j &= \tfrac{2}{3}h_{j-1} + \tfrac{1}{3}h_{j+1} + 1, \quad 1 \le j \le n-1 \\ h_0 &= h_1 + 1. \end{aligned}$$

Again it is useful to substitute $\Delta_j = h_j - h_{j-1}$, giving the recursion

$$\Delta_{j+1} = 2\Delta_j - 3$$

with the solution, taking into account also the boundary conditions,

$$\Delta_j = 3 - 2^{j+1}.$$

Hence, for all j we have

$$0 = h_n = h_j + \sum_{i=j+1}^{n} \Delta_i = h_j - 2^{n+2} - 2^{j+2} + 3(n-j),$$

so

$$h_j = 2^{n+2} - 2^{j+2} - 3(n-j).$$

This straightforward analysis gives an estimate $O(2^n)$ for the number of iterations. This of course is not interesting, as we could as well search all the assignments deterministically by brute force.

We make two observations that help improve the algorithm.

1. If the initial assignment is not quite arbitrary, but uniformly random, then the initial number of matching variables is distributed as $\mathrm{Bin}(n, 1/2)$. Hence, with some small but non-zero probability we get an initial value j that is clearly larger than $n/2$.

2. Iterating for too long descreases the probability of finding a good assignment.

Therefore, we change the strategy so that we have more and shorter iteration rounds.

We get the modified algorithm.

1. Repeat m times or until the formula is satisfied:
   (a) Assign the variables uniformly at random.
   (b) Repeat $3n$ times or until the formula is satisfied:
      (i) Choose a random unsatisfied clause.
      (ii) Choose a random literal in the clause
      and change its value (making the clause satisfied).
2. If the formula is satisfied, return the assignment.
   Otherwise return "not satisfiable."

Suppose that at some iteration there are j variables whose assigned value is different from S. Let q_j be the probability that in at most $3n$ iterations the algorithm finds S. Now

$$q_j \geq \binom{3j}{j} \left(\frac{2}{3}\right)^j \left(\frac{1}{3}\right)^{2j}.$$

We get this by considering the case where Y_i increases exactly $2j$ times and decreases exactly j times during $3j \leq 3n$ iterations.

We estimate the binomial coefficient by using a simple form of Stirling's formule: for all $m > 0$ we have

$$\sqrt{2\pi m}\left(\frac{m}{\mathrm{e}}\right)^m \leq m! \leq 2\sqrt{2\pi m}\left(\frac{m}{\mathrm{e}}\right)^m.$$

When $j > 0$, we get

$$\begin{aligned}
\binom{3j}{j} &= \frac{(3j)!}{j!(2j)!} \\
&\geq \frac{1}{4\sqrt{2\pi}}\sqrt{\frac{3j}{j \cdot 2j}}\left(\frac{3j}{\mathrm{e}}\right)^{3j}\left(\frac{\mathrm{e}}{2j}\right)^{2j}\left(\frac{\mathrm{e}}{j}\right)^{j} \\
&= \frac{1}{8}\sqrt{\frac{3}{\pi j}}\left(\frac{27}{4}\right)^{j} \\
&= \frac{c}{\sqrt{j}}\left(\frac{27}{4}\right)^{j}
\end{aligned}$$

where $c = (1/8)\sqrt{3/\pi}$.

Hence, for $j > 0$ we get

$$q_j \geq \frac{c}{\sqrt{j}}\left(\frac{27}{4}\right)^{j}\left(\frac{2}{3}\right)^{j}\left(\frac{1}{3}\right)^{2j} \geq \frac{c}{2^j\sqrt{j}}.$$

Furthermore, $q_0 = 1$.

Now the probability that a random initial assignment A leads to S (or another satisfying assignment) in $3n$ iterations is at least

$$\begin{aligned} q &\geq \sum_{j=0}^{n} q_j \Pr(\text{assignment } A \text{ differs in exactly } j \text{ variables}) \\ &\geq \frac{1}{2^n} + \sum_{j=1}^{n} \binom{n}{j} \left(\frac{1}{2}\right)^n \frac{c}{2^j\sqrt{j}} \\ &\geq \frac{c}{\sqrt{n}} \left(\frac{1}{2}\right)^n \sum_{j=0}^{n} \binom{n}{j} \left(\frac{1}{2}\right)^j 1^{n-j} \\ &= \frac{c}{\sqrt{n}} \left(\frac{1}{2}\right)^n \left(1+\tfrac{1}{2}\right)^n \\ &= \frac{c}{\sqrt{n}} \left(\frac{3}{4}\right)^n. \end{aligned}$$

Therefore, the number of initial assignments A we need to try is upper bounded by Geom(q). Hence, in expectation we try $1/q$ initial assignment, and for each initialization we iterate at most $3n$ steps, so in expectation we make $O(n^{3/2}(4/3)^n)$ iteration steps.

Classification of states [M&U Section 7.2]

State j is accessible from state i, if $P^n_{i,j} > 0$ for some $n \geq 0$. If i and j are accessible from each other, they communicate, which we denote by $i \leftrightarrow j$.

Hence, $\leftrightarrow$ is the equivalence relation where the equivalence classes are the strongly connected components of the graph representation of the chain. If there is only one equivalence class, i.e., the graph is strongly connected, the chain is irreducible.

Let $r^t_{i,j}$ be the probability that if the process starts from state i, it will at time t enter state j for the first time:

$$r^t_{i,j} = \Pr(X_t = j \text{ and } X_n \neq j \text{ for } n < t \mid X_0 = i).$$

State i is recurrent, if $\sum_{t \geq 1} r^t_{i,i} = 1$. Then with probability 1, the state will be repeated infinitely often, assuming it's entered at all.

If $\sum_{t \geq 1} r^t_{i,i} < 1$, the state is transient. Then with probability 1 it occurs only a finite number of times, and the number of occurences has geometric distribution.

A Markov chain is recurrent, if its every state is.

Let

$$h_{i,j} = \sum_{t \geq 1} t r_{i,j}^t$$

be the expected transition time from state i to state j.

If i is recurrent and $h_{i,i}$ is finite, we say that i is positive recurrent.

If i is recurrent but $h_{i,i} = \infty$, we say that i is null recurrent.

Example 1.4: Consider a Markov chain with an infinite number of states $1, 2, 3, \ldots$. Let.

$$\begin{aligned} P_{i,i+1} &= \frac{i}{i+1} \\ P_{i,1} &= \frac{1}{i+1}. \end{aligned}$$

Starting from state 1, the probability of avoiding returning to state 1 within the first t steps is

$$\prod_{j=1}^{t} \frac{j}{j+1} = \frac{1}{t+1}.$$

Hence, the probability of never returning to state 1 is $\lim_{t\to\infty} 1/(t+1) = 0$, so state 1 is recurrent. However,

$$r_{1,1}^{t} = \frac{1}{t(t+1)},$$

so

$$h_{1,1} = \sum_{t\geq 1} t r_{1,1}^{t} = \sum_{t\geq 1} \frac{1}{t+1} = \infty$$

and state 1 is null recurrent. □

More generally, it turns out the null recurrent states exist only in chains with an infinite number of states.

Lemma 1.5 [M&U Lemma 7.5]**:** In a finite Markov chain there is at least one recurrent state, and all recurrent states are positive recurrent.

(Proof is left as an exercise. □)

State j is periodic if there is some integer $\Delta > 1$ such that $\Pr(X_{t+s} = j \mid X_t = j) = 0$ when s is not divisible by Δ. A chain is periodic, if it contains at least one periodic state. A state or chain that is not periodic is aperiodic.

A state is ergodic if it is aperiodic and positive recurrent. A chain is ergodic, if all of its states are.

Corollary 1.6 [M&U Corollary 7.6]**:** A finite, irreducible and aperiodic Markov chain is ergodic. □

Example: Gambler's Ruin [M&U Section 7.2.1]

Consider playing repeatedly the following fair zero-sum game between two players. With probability $1/2$ player A gives one euro to player B, and with probability $1/2$ player B gives one euro to player A.

Let W_t be the amount that player A has won in the first t rounds. If player A has lost money, this value is negative. We assume that player A has ℓ_1 euros available, and player B has ℓ_2 euros.

If one player runs out of money, he has lost and the game is stopped. If, for example, player A loses in t rounds, we define $W_t = -\ell_1$ for $W_t \geq t$.

We model this as a Markov chain (W^t) with a finite state space $\{-\ell_1, \ldots, \ell_2\}$, initial state state 0, and transition probabilities

$$\begin{aligned} P_{i,i+1} &= 1/2, & -\ell_1 < i < \ell_2 \\ P_{i,i-1} &= 1/2, & -\ell_1 < i < \ell_2 \\ P_{-\ell_1,-\ell_1} &= 1 \\ P_{\ell_2,\ell_2} &= 1. \end{aligned}$$

What is the probability q for the event that player A wins?

Clearly states $-\ell_1$ and ℓ_2 are recurrent and the other states transient. The probability of the chain ending up in state ℓ_2 is

$$q = \lim_{t\to\infty} P^t_{0,\ell_2},$$

and the probability of ending up in state $-\ell_1$ is $1-q$.

Each round is fair, meaning that the expected change of the wealth of player A is zero. Therefore, at any time t we have

$$0 = \mathbf{E}[W^t] = \sum_{i=-\ell_1}^{\ell_2} i P^t_{0,i}.$$

Since $\lim_t P^t_{0,i} = 0$ for all $-\ell_1 < i < \ell_2$, we get

$$\lim_{t\to\infty} \mathbf{E}[W^t] = \ell_2 q - \ell_1(1-q) = 0,$$

from which we can solve

$$q = \frac{\ell_1}{\ell_1 + \ell_2}.$$

Another way to solve this is to denote by q_j the probability that A reaches wealth ℓ_2 before reaching wealth $-\ell_1$, if initially his wealth is j. Clearly $q_{-\ell_1} = 0$, $q_{\ell_2} = 1$ and

$$q_j = \frac{q_{j-1}}{2} + \frac{q_{j+1}}{2}, \quad -\ell_1 < j < \ell_2.$$

By writing the recurrence as $\Delta_j = \Delta_{j+1}$, where $\Delta_j = q_j - q_{j-1}$, we easily see that

$$q_j = \frac{\ell_1 + j}{\ell_1 + \ell_2}.$$

Stationary distribution [M&U Section 7.3]

A vector $\boldsymbol{\pi}$ is a probability vector, if $\pi_i \geq 0$ for all i, and $\sum_i \pi_i = 1$. A probability vector π is a stationary distribution of a Markov chain with transition matrix $\mathbf{P}$, if

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}.$$

The following is a fundamental result about finite Markov chains.

Theorem 1.7 [M&U Thm 7.7]**:** If a finite Markov chain with transition matrix $\mathbf{P}$ is irreducible and ergodic, the following conditions hold:

1. The chain has a unique stationary distribution $\boldsymbol{\pi}$.
2. For all j and i the limit $\lim_{t\to\infty} P^t_{j,i}$ exists and is the same for all j.
3. $\pi_i = \lim_{t\to\infty} P^t_{j,i} = 1/h_{i,i}$.

Intuitively, under the given conditions the effect of the initial state vanishes as time goes by.

Aperiodicity is not a necessary condition for the existence of a stationary distribution. For example, in a two-state chain with transition probabilities $P_{1,2} = P_{2,1} = 1$, the stationary distribution is $(1/2, 1/2)$. However, there is no convergence to this distribution. Depending on the initial state, we get either states $(1, 2, 1, 2, 1, 2, \ldots)$ or states $(2, 1, 2, 1, 2, 1, \ldots)$.

In a finite chain there is always at least one recurrent state, and if the chain enters the component containing this state, it will never leave. Thus, the chain will always end up in the distribution corresponding to the stationary distribution of one of the components. However, if there are several components with recurrent states, again we do not converge to any fixed stationary distribution. The initial transient phase decides in which component we end up, and each component has its own stationary distribution.

The textbook proves Theorem 1.7 assuming the following lemma.

Lemma 1.8: If the chain is irreducible and ergodic, then for all i the limit $\lim_{t\to\infty} P_{i,i}^t$ exists and

$$\lim_{t\to\infty} P_{i,i}^t = \frac{1}{h_{i,i}}.$$

We omit the proof. See for example Chung: *Markov Chains with Stationary Transition Probabilities*.

Intuitively it is not surprising that if the limit $\lim_{t\to\infty} P_{i,i}^t$ exists, its value can only be $1/h_{i,i}$, because the state i must on the average appear once every $h_{i,i}$ steps. To prove the existence of the limit, we need ergodicity and irreducibility.

Proof of Theorem 1.7: We first show that for all i and j we have

$$\lim_{t\to\infty} P_{j,i}^t = \lim_{t\to\infty} P_{i,i}^t = \frac{1}{h_{i,i}}.$$

Recall that $r_{j,i}^t$ is the probability of entering state i for the first time at time t having started from j. Since the chain is irreducible and all the states are recurrent, the probability of never entering i after having started from j is 0, so

$$\sum_{t=1}^{\infty} r_{j,i}^t = 1.$$

Let $\varepsilon > 0$, and let t_1 be such that

$$\sum_{t=1}^{t_1} r_{j,i}^t \geq 1 - \varepsilon.$$

For $i \neq j$ we can write

$$P_{j,i}^t = \sum_{k=1}^{t} r_{j,i}^k P_{i,i}^{t-k}.$$

When $t \geq t_1$, we get

$$\sum_{k=1}^{t_1} r_{j,i}^k P_{i,i}^{t-k} \leq \sum_{k=1}^{t} r_{j,i}^k P_{i,i}^{t-k} = P_{j,i}^t.$$

By Lemma 1.8 we can now take the limit

$$\begin{aligned}
\lim_{t\to\infty} P_{j,i}^t &\geq \lim_{t\to\infty} \sum_{k=1}^{t_1} r_{j,i}^k P_{i,i}^{t-k} \\
&= \sum_{k=1}^{t_1} r_{j,i}^k \lim_{t\to\infty} P_{i,i}^{t-k} \\
&= \lim_{t\to\infty} P_{i,i}^t \sum_{k=1}^{t_1} r_{j,i}^k \\
&\geq (1-\varepsilon) \lim_{t\to\infty} P_{i,i}^t.
\end{aligned}$$

Similarly,

$$P_{j,i}^t = \sum_{k=1}^{t} r_{j,i}^k P_{i,i}^{t-k} \leq \sum_{k=1}^{t_1} r_{j,i}^k P_{i,i}^{t-k} + \sum_{k>t_1} r_{j,i}^k \leq \sum_{k=1}^{t_1} r_{j,i}^k P_{i,i}^{t-k} + \varepsilon,$$

which implies

$$\begin{aligned} \lim_{t\to\infty} P_{j,i}^t &\leq \lim_{t\to\infty} \left(\sum_{k=1}^{t_1} r_{j,i}^k P_{i,i}^{t-k} + \varepsilon \right) \\ &= \sum_{k=1}^{t_1} r_{j,i}^k \lim_{t\to\infty} P_{i,i}^{t-k} + \varepsilon \\ &\leq \lim_{t\to\infty} P_{i,i}^t + \varepsilon. \end{aligned}$$

In the limit $\varepsilon \to 0$ we now get

$$\lim_{t\to\infty} P_{j,i}^t = \lim_{t\to\infty} P_{i,i}^t,$$

and by Lemma 1.8 this limit is $1/h_{i,i}$.

Define now $\pi_i = 1/h_{i,i}$. We are going to show that this defines a stationary distribution. For a finite state set $\{0, \ldots, n\}$, we get for any j

$$\sum_{i=0}^{n} \pi_i = \sum_{i=0}^{n} \lim_{t\to\infty} P_{j,i}^t = \lim_{t\to\infty} \sum_{i=0}^{n} P_{j,i}^t = \lim_{t\to\infty} 1 = 1,$$

so $\boldsymbol{\pi}$ is a probability vector.

The stationarity then follows:

$$\pi_i = \lim_{t\to\infty} P_{j,i}^{t+1} = \lim_{t\to\infty} \sum_{k=0}^{n} P_{j,k}^t P_{k,i} = \sum_{k=0}^{n} \lim_{t\to\infty} P_{j,k}^t P_{k,i} = \sum_{k=0}^{n} \pi_k P_{k,i}.$$

Finally, we check the uniqueness. Let $\boldsymbol{\phi}$ be a stationary distribution, so

$$\boldsymbol{\phi} = \boldsymbol{\phi}\mathbf{P}^t$$

for all t. In particular,

$$\phi_i = \lim_{t\to\infty} \sum_{k=0}^{n} \phi_k P_{k,i}^t = \sum_{k=0}^{n} \phi_k \lim_{t\to\infty} P_{k,i}^t = \sum_{k=0}^{n} \phi_k \pi_i = \pi_i,$$

because $\sum_k \phi_k = 1$. $\square$

We can find the stationary distribution by solving the system of equations $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ under the constraint $\sum_i \pi_i = 1$.

The following result is often helpful.

Theorem 1.9 [M&U Thm 7.9]**:** Let S be a set of states in a finite Markov chain with a stationary distribution. Then in the stationary distribution, the probability of leaving set S is the same as of entering set S.

Proof: Let $\boldsymbol{\pi}$ be the stationary distribution. Then

$$\sum_{i \in S} \sum_{j=0}^{n} \pi_j P_{j,i} = \sum_{i \in S} \pi_i = \sum_{i \in S} \pi_i \sum_{j=0}^{n} P_{i,j}.$$

Eliminating terms that appear on both sides yields

$$\sum_{i \in S} \sum_{j \notin S} \pi_j P_{j,i} = \sum_{i \in S} \sum_{j \notin S} \pi_i P_{i,j}.$$

□

Example 1.10: Consider Markov chain with transition probabilities

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

By applying the previous result about the stationary distribution (π_o, π_1) to $S = \{0\}$ we get

$$\pi_0 p = \pi_1 q.$$

Since $\pi_0 + \pi_1 = 1$, we get $\pi_0 = q/(p+q)$ ja $\pi_1 = p/(p+q)$.

In some cases, the next theorem gives an easy way of calculating the stationary distribution.

Theorem 1.11 [M&U Thm 7.10]: If $\sum_i \pi_i = 1$ and

$$\pi_i P_{i,j} = \pi_j P_{j,i} \qquad (*)$$

for all i, j, then $\boldsymbol{\pi}$ is the stationary distribution.

Proof:

$$\sum_{i=0}^{n} \pi_i P_{i,j} = \sum_{i=0}^{n} \pi_j P_{j,i} = \pi_j \sum_{i=0}^{n} P_{j,i} = \pi_j.$$

□

The condition $(*)$ is **not** a necessary condition for the stationary distribution. If it holds, the chain is said to be time reversible.

Notice that the condition involves $n(n-1)/2$ equations for n unknowns.

We state without proof how this generalises to infinite chains.

Theorem 1.12 [M&U Thm 7.11]**:** If a Markov chain with transition matrix $\mathbf{P}$ is irreducible and aperiodic, then exactly one of the following holds.

1. The chain is ergodic with a unique stationary distribution $\boldsymbol{\pi}$ that satisfies $\pi_i = \lim_{t\to\infty} P^t_{j,i} > 0$ for all i, j.

2. No state in the chain is positive recurrent, $\lim_{t\to\infty} P^t_{j,i} = 0$ for all i, j, and the chain has no stationary distribution.

□

Theorems 1.9 and 1.11 hold also for infinite chains.

Example: simple queue [M&U Section 7.3.1]

A queue can hold up to n customers. At each time step, exactly one of the following takes place:

- If the queue currently has less than n customers, with probability λ a new customer will join the queue.
- If the queue currently has at least one customer, with probability μ the first customer is served and leaves the queue.
- Otherwise there is no change.

The queue length X_t is a Markov chain with

$$
\begin{aligned}
P_{i,i+1} &= \lambda, \qquad i < n \\
P_{i,i-1} &= \mu, \qquad i > 0 \\
P_{i,i} &= \begin{cases} 1-\lambda & \text{for } i = 0 \\ 1-\lambda-\mu & \text{for } 0 < i < n \\ 1-\mu & \text{for } i = n. \end{cases}
\end{aligned}
$$

To find out the stationary distribution, we have the system of equations

$$\begin{aligned} \pi_0 &= (1-\lambda)\pi_0 + \mu\pi_1 \\ \pi_i &= \lambda\pi_{i-1} + (1-\lambda-\mu)\pi_i + \mu\pi_{i+1}, \qquad 0 < i < n \\ \pi_n &= \lambda\pi_{n-1} + (1-\mu)\pi_n. \end{aligned}$$

The first equation gives $\pi_1 = (\lambda/\mu)\pi_0$. By substituting this into the second one we get $\pi_2 = (\lambda/\mu)\pi_1$. From this, we guess

$$\pi_i = \pi_0 \left(\frac{\lambda}{\mu}\right)^i,$$

which is easily verified by induction. The normalization constraint $\sum_i \pi_i = 1$ then gives us

$$\pi_i = \frac{1}{Z}\left(\frac{\lambda}{\mu}\right)^i,$$

where $Z = \sum_{i=0}^{n} (\lambda/\mu)^i$.

Another method is to consider a partitioning to two sets, $S = \{0, \ldots, i\}$ and $\{i+1, \ldots, n\}$. We know that $\pi_i P_{i,i+1} = \pi_{i+1} P_{i+1,i}$, implying $\pi_i \lambda = \pi_{i+1} \mu$. Again, by induction we get $\pi_i = \pi_0 (\lambda/\mu)^i$.

If the queue length is not bounded, we have

$$\begin{aligned} \pi_0 &= (1-\lambda)\pi_0 + \mu\pi_1 \\ \pi_i &= \lambda\pi_{i-1} + (1-\lambda-\mu)\pi_i + \mu\pi_{i+1}, i \geq 1, \end{aligned}$$

and as previously we get $\pi_i = \pi_0 (\lambda/\mu)^i$. The normalization condition is now

$$\pi_0 \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i = 1.$$

For $\lambda < \mu$ we get the stationary distribution

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \left(1 - \frac{\lambda}{\mu}\right).$$

For $\lambda \geq \mu$, the series diverges and there is no stationary distribution. Intuitively, customers enter faster than they leave and the queue length tends towards infinity. It is possible to show that in the case $\lambda > \mu$ all states are transient and in the case $\lambda = \mu$ null recurrent.

Random walk in an undirected graph [M&U Section 7.4]

Let $G = (V, E)$ be a connected undirected graph and $d(v)$ the degree of vertex v. A random walk in G is the Markov chain with state space V and transition probabilities $P_{ij} = 1/d(i)$ for all i, j such that $(i, j) \in E$.

Lemma 1.13: The random walk in G is aperiodic, if and only if G is not bipartite, in other words, there is no partitioning $V = V_1 \cup V_2$ such that $E \subseteq V_1 \times V_2$.

Proof: The "only if" direction is clear.

If the graph is not bipartite, it has an odd-length cycle. Since there is also a path of lenght two from any vertex back to itself, the walk is aperiodic. □

From now on we assume that G is not bipartite. Then the Markov chain is finite, irreducible and aperiodic. Therefore it will converge towards the stationary distribution.

Theorem 1.14: The stationary distribution $\boldsymbol{\pi}$ of the random walk in G satisfies

$$\pi_v = \frac{d(v)}{2\,|E|}.$$

Proof: Since $\sum_v d(v) = 2\,|E|$, this defines $\boldsymbol{\pi}$ which is a probability vector.

Furthermore,

$$(\boldsymbol{\pi}\boldsymbol{P})_u = \sum_v \pi_v P_{v,u} = \sum_{v:(v,u)\in E} \frac{d(v)}{2\,|E|}\frac{1}{d(v)} = \pi_u.$$

□

Recall that the stationary distribution satisfies $\pi_i = 1/h_{i,i}$ where $h_{i,j}$ is the expected time to get from vertex i to vertex j.

Corollary 1.15: For all v we have $h_{v,v} = 2|E|/d(v)$. □

Lemma 1.16: If $(u,v) \in E$, then $h_{v,u} < 2|E|$.

Proof: Since

$$\frac{2|E|}{d(u)} = h_{u,u} = 1 + \sum_v P_{u,v} h_{v,u} = \frac{1}{d(u)} \sum_{v:(u,v)\in E} (1 + h_{v,u}),$$

we must have $h_{v,u} < 2|E|$ whenever $(u,v) \in E$. □

The cover time of a graph is the maximum over vertices v of the expected time for the random walk starting from v to visit all vertices.

Lemma 1.17: The cover time of $G = (V, E)$ is at most $4\,|V|\,|E|$.

Proof: Take any spanning tree of G and choose vertex v as a root. Traverse it in depth-first order so that each edge in the tree is traversed once in each direction. The tree has $|V| - 1$ edges, so this traversal gives a sequence of vertices $v_0, v_1, \ldots, v_{2|V|-2}$ where $v_0 = v_{2|V|-2} = v$ and $(v_i, v_{i+1}) \in E$ for all i.

By the previous lemma, the cover time is upper bounded by

$$\sum_{i=0}^{2|V|-3} h_{v_i,v_{i+1}} < (2\,|V| - 2)(2\,|E|) < 4\,|V|\,|E|\,.$$

□

Example: reachability

We are given a graph $G = (V, E)$ and vertices $s, t \in V$. The problem is to decide whether there is a path between s and t.

We can easily do this for example by depth-first search. This however requires $O(|V|)$ work space.

We consider a randomized algorithm.

1. Start a random walk from s.
2. If within $4\,|V|^3$ steps vertex t is visited,
   answer "yes;" otherwise answer "no."

If there is no path between s and t, the algorithm always answers "no."

Assume that a path exists, and the connected component containing s and t is not bipartite. (If needed, we can as pre-processing add one edge to create a triangle.) The expected time for the random walk to reach t is at most $4|V||E| \leq 2|V|^3$. Hence, a path is found in time $4|V|^3$ with probability at least $1/2$ (Markov).

The algorithm needs only $O(\log |V|)$ bits for book keeping (assuming there's no issue in making the random choices).

Parrando's Paradox [M&U Section 7.5]

We consider a situation where two games in which the expected gain is negative can be combined into one in which the expected gain is positive.

All the games consist of repeating flips of a biased coin. If the result is heads, the player wins one euro. Otherwise, he loses one euro.

In game A, we flip a coin a with probability of heads $p_a < 1/2$; for example, $p_a = 0.49$. The expected loss of the player is then $1 - 2p_a$ euros per round.

The game B uses coin b, if the net profit of the player up to now is divisible by three (in euros). Otherwise, coin c is used. The probabilities of heads for these coins are p_b and p_c, respectively.

For concreteness, let us choose $p_b = 0.09$, $p_c = 0.74$. **If** it were the case that 1/3 of the time, the player's profit is divisible by three, we could calculate the winning probability as

$$\frac{1}{3} \cdot \frac{9}{100} + \frac{2}{3} \cdot \frac{74}{100} = \frac{157}{300} > \frac{1}{2}.$$

However, this assumption is **false**. Typically, the player loses the first round, since to start with his profit is 0 and we use coin b. Then in the second round, coin c is used, and the playes is likely to win, getting back to 0 total profit. The game may oscillate quite long between -1 and 0.

We need to evaluate the winning probability in game B over a longer period of time. Consider a Markov chain that has a states $\{-3, \ldots, 3\}$, which we interpret as amounts of profit. The expected profit over long sequences of play is negative, if starting from 0 we are more likely to end up in -3 than in 3.

Let z_i be the probability of reaching state -3 before state 3 starting from state i. We wish to evaluate z_0. We can solve it from the boundary conditions $z_{-3} = 1$ and $z_3 = 0$ and the recursion

$$\begin{aligned} z_{-2} &= (1-p_c)z_{-3} + p_c z_{-1} \\ z_{-1} &= (1-p_c)z_{-2} + p_c z_0 \\ z_0 &= (1-p_b)z_{-1} + p_b z_1 \\ z_1 &= (1-p_c)z_0 + p_c z_2 \\ z_2 &= (1-p_c)z_1 + p_c z_3 \end{aligned}$$

which yields

$$z_0 = \frac{(1-p_b)(1-p_c)^2}{(1-p_b)(1-p_c)^2 + p_b p^2}.$$

Thus, with the parameter values we chose, we have $z_0 \approx 0{,}555$, making the expected profit negative for long games.

We present another proof for the same fact.

Let s be a sequence of states starting from 0 and ending in 3. Let $f(s)$ be another sequence which is the same as s except that the signs are flipped after the last 0. Thus, for example,

$$f(0,-1,0,1,2,1,0,1,2,3) = (0,-1,0,1,2,1,0,-1,-2,-3).$$

Clearly f is a bijection from sequences ending in 3 to sequences ending in -3.

Lemma 1.18: For any s that ends in 3 we have

$$\frac{\Pr(s)}{\Pr(f(s))} = \frac{p_b p_c^2}{(1-p_b)(1-p_c)^2}.$$

Before proving Lemma 7.18, we notice that it implies

$$\frac{\Pr(\text{state 3 occurs before } -3)}{\Pr(\text{state } -3 \text{ occurs before 3})} = \frac{\sum_s \Pr(s)}{\sum_s \Pr(f(s))} = \frac{p_b p_c^2}{(1-p_b)(1-p_c)^2}.$$

Here summation is over sequences s that end in 3 without entering -3.

Proof of Lemma 7.18: Divide the transitions into four classes:

$$\begin{array}{ll} A_1 & 0 \to 1 \\ A_2 & 0 \to -1 \\ A_3 & -2 \to -1, \quad -1 \to 0, \quad 1 \to 2, \quad 2 \to 3 \\ A_4 & -2 \to -3, \quad -1 \to -2, \quad 1 \to 0, \quad 2 \to 1. \end{array}$$

Let t_i be the number of transitions belonging to A_i in s.

The transformation $s \mapsto f(s)$ changes one transition from class A_1 into a transition from class A_2. Additionally, the number of transitions from A_3 decreases and the number of transitions from A_4 increases by 2. Therefore,

$$\begin{aligned} \Pr(s) &= p_b^{t_1}(1-p_b)^{t_2}p_c^{t_3}(1-p_c)^{t_4} \\ \Pr(f(s)) &= p_b^{t_1-1}(1-p_b)^{t_2+1}p_c^{t_3-2}(1-p_c)^{t_4+2}. \end{aligned}$$

□

We can also analyse this via the stationary distribution. Consider a Markov chain with three states $\{0,1,2\}$, which represent the profit of the player modulo 3. Let $\boldsymbol{\pi}$ be the stationary distribution. The probability to win one euro approaches in a long game the value

$$p_b\pi_0 + p_c\pi_1 + p_c\pi_2 = p_b\pi_0 + p_c(1-\pi_0) = p_c - (p_c - p_b)\pi_0.$$

We get for the stationary distribution a system of equations

$$\begin{aligned}
\pi_0 &= (1-p_c)\pi_1 + p_c\pi_2 \\
\pi_1 &= p_b\pi_0 + (1-p_c)\pi_2 \\
\pi_2 &= (1-p_b)\pi_0 + p_c\pi_1.
\end{aligned}$$

Together with the normalization constraint $\sum_i \pi_i = 1$ this yields

$$\begin{aligned}
\pi_0 &= \frac{1}{Z}(1 - p_c + p_c^2) \\
\pi_1 &= \frac{1}{Z}(p_b p_c - p_c + 1) \\
\pi_2 &= \frac{1}{Z}(p_b p_c - p_b + 1)
\end{aligned}$$

where $Z = 3 - 2p_c - p_b + 2p_b p_c + p_c^2$. In our example case, $\pi_0 = 673/1759 \approx 0{,}3826$ and $p_c - (p_c - p_b)\pi_0 = 86421/175900 < 1/2$.

Consider now a game C in which in each round, first a fair coin is flipped, and based on the outcome, we play either game A or B. Equivalently, we could play B so that after choosing between b and c, we would with probability $1/2$ decide to use a instead. Thus, the winning probabilities in game C are obtained from the winning probabilities in B by replacing p_b and p_c by $p_b^* = \frac{1}{2}(p_a + p_b)$ and $p_c^* = \frac{1}{2}(p_a + p_c)$.

Hence, the win ratio in game C is with our parameter values

$$\frac{p_b^*(p_c^*)^2}{(1-p_b^*)(1-p_c^*)^2} = \frac{438741}{420959} > 1$$

making the expected profit positive. We can also derive this via the stationary distribution:

$$p_c^* - (p_c^* - p_b^*)\pi_0^* = \frac{4456523}{8859700} > \frac{1}{2},$$

where $\boldsymbol{\pi}^*$ is the stationary distribution for B calculated using p_b^* and p_c^*.

Intuitively, game B leads to loss because of its particular structure. Mixing game A into it breaks the structure and in particular helps the player leave the "hole" around zero profit.

More formally we can notice that the paradox does not violate the linearity of expectation. For all states s we have

$$\mathbf{E}[X_C \mid s] = \tfrac{1}{2}\mathbf{E}[X_A \mid s] + \tfrac{1}{2}\mathbf{E}[X_B \mid s]$$

where X_Z is the profit in game Z at a given time. However, conditioning on s changes the winning probabilities in games A and B.

2. Continuous random variables

We consider random variables that have as range the real numbers, or some real interval. From application point of view, we are particularly interested in random variables that represent the time when some even occurs.

We avoid going into general theory, since in applications we usually deal with the basic setting where there is a density function. In practice, continuity then often just means that we calculate integrals instead of sums. However there are some technical issues that need to be taken into account.

Outline of this chapter:

- basic properties, some important distributions
- Poisson processes
- Markov processes with continuous time.

Recall that a random variable is a mapping $X : \Omega \to \mathbb{R}$, where $(\Omega, \mathcal{F}, \Pr)$ is a probability space and $\Pr(X(\omega) \leq a)$ is defined for all $a \in \mathbb{R}$.

Often in practice the original sample space Ω is not interesting and we deal just with the distribution function F defined by

$$F(a) = \Pr(X \leq a).$$

Previously we mainly considered discrete random variables with a finite or countably infinite range. We say that X is continuous if the distribution function F is continuous. Then in particular $\Pr(X = a) = 0$ for all $a \in \mathbb{R}$.

If there is a function f for which

$$F(a) = \int_{-\infty}^{a} f(t)\, dt,$$

we call f the density function of X. Then $F'(a) = f(a)$.

The expected value of a random variable with a density function is defined as

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} t f(t)\, dt.$$

More generally,

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(t) f(t)\, dt.$$

As before, we are in particular interested in the variance

$$\mathbf{Var}[X] = \int_{-\infty}^{\infty} (t - \mathbf{E}[X])^2 f(t)\, dt$$

and moments

$$\mathbf{E}[X^i] = \int_{-\infty}^{\infty} t^i f(t)\, dt.$$

In the following we assume that continuous random variables have a density function.

Lemma 2.1 [M&U Lemma 8.1]**:** If X is continuous and gets only non-negative values, we have

$$\mathbf{E}[X] = \int_0^\infty \Pr(X \geq t)\, dt.$$

The discrete analogue is [M&U Lemma 2.9] (also page 42 of lecture notes for *Randomized Algorithms I*).

Proof: Let f be the density function of X. Then

$$\begin{aligned}
\int_0^\infty \Pr(X \geq t)\, dt &= \int_0^\infty \int_t^\infty f(s)\, ds\, dt \\
&= \int_0^\infty \int_0^s f(s)\, dt\, ds \\
&= \int_0^\infty s f(s)\, ds \\
&= \mathbf{E}[X].
\end{aligned}$$

□

If X and Y are continuous random variables, their joint distribution function is

$$F(x, y) = \mathsf{Pr}(X \leq x, Y \leq y)$$

and joint density function

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

(if the partial derivatives exist). Thus,

$$F(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y)\, dy\, dx.$$

This is generalized to three or more variables in the obvious manner.

Example Consider the distribution

$$F(x,y) = 1 - \mathrm{e}^{-ax} - \mathrm{e}^{-by} + \mathrm{e}^{-ax-by}, \quad x,y \geq 0,$$

where $a, b > 0$, and $F(x,y) = 0$ if $x < 0$ or $y < 0$. Now

$$f(x,y) = \frac{\partial}{\partial x}(0 + 0 - b\mathrm{e}^{-by} - b\mathrm{e}^{-ax-by}) = ab\mathrm{e}^{-ax-by}, \quad x,y > 0,$$

and $f(x,y) = 0$ if $x < 0$ or $y < 0$. As a check we can verify that

$$\begin{aligned}
\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,dv\,du &= \int_{0}^{x}\int_{0}^{y} ab\mathrm{e}^{-au-bv}\,dv\,du \\
&= ab(\int_{0}^{x}\mathrm{e}^{-au}\,du)(\int_{0}^{y}\mathrm{e}^{-bv}\,dv) \\
&= ab\cdot\frac{1}{-a}(\mathrm{e}^{-ax}-1)\cdot\frac{1}{-b}(\mathrm{e}^{-by}-1) \\
&= F(x,y).
\end{aligned}$$

From the joint distribution function F we can obtain the marginal distribution functions

$$F_X(x) = \Pr(X \leq x), \qquad F_Y(y) = \Pr(Y \leq y).$$

We denote the corresponding density functions by f_X and f_Y. The marginar density function can be obtained from the joint density function by "integrating away" the other variable:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

Random variables X and Y are independent, if

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x) \Pr(Y \leq y),$$

that is,

$$F(x, y) = F_X(x) F_Y(y),$$

for all x, y. If the relevant derivatives exist, this is equivalent with

$$f(x, y) = f_X(x) f_Y(y)$$

for all x, y.

Example Consider again

$$F(x,y) = 1 - \mathrm{e}^{-ax} - \mathrm{e}^{-by} + \mathrm{e}^{-ax-by}.$$

Since $F(x,y) = P(x)Q(y)$, where

$$\begin{aligned} P(x) &= 1 - \mathrm{e}^{-ax} \\ Q(y) &= 1 - \mathrm{e}^{-by} \end{aligned}$$

and P and Q are valid distribution functions, we conclude that $F_X = P$ and $F_Y = Q$, and X and Y are independent. For the density functions we get

$$f(x,y) = (a\mathrm{e}^{-ax})(b\mathrm{e}^{-by}) = f_X(x)f_Y(y).$$

In defining conditional probabilities we run into technical problems. If we take the definition $\Pr(A \mid B) = \Pr(A \cap B)/\Pr(B)$ as it is, we get expressions of the form 0/0 for example if B is of the form "$Y = y$". Intuitively, however, such conditional probabilities would seem to make sense. For example, in the previous example we would like to say that

$$\Pr(X + Y \leq 5 \mid Y = 3) = \Pr(X \leq 2 \mid Y = 3) = \Pr(X \leq 2) = 1 - \mathrm{e}^{-2a},$$

because X and Y are independent.

For discrete random variables, when $\Pr(Y = y) > 0$,

$$\Pr(X \leq x \mid Y = y) = \sum_{u \leq x} \frac{\Pr(X = u, Y = y)}{\Pr(Y = y)}.$$

For continuous random variables we define analogously

$$\Pr(X \leq x \mid Y = y) = \int_{-\infty}^{x} \frac{f(u, y)}{f_Y(y)}\, du,$$

when $f_Y(y) > 0$. The conditional probability is thus defined via the conditional density function

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

To motivate the previous definition, notice that it satisfies the property $\Pr(X \leq x \mid Y = y) = \lim_{h \to 0+} \Pr(X \leq x \mid y \leq Y \leq y + h)$, when conditioning over a set with strictly positive probability is defined in the usual manner:

$$
\begin{aligned}
&\lim_{h \to 0+} \Pr(X \leq x \mid y \leq Y \leq y + h) \\
&\quad = \lim_{h \to 0+} \frac{\Pr(X \leq x, y \leq Y \leq y + h)}{\Pr(y \leq Y \leq y + h)} \\
&\quad = \lim_{h \to 0+} \frac{F(x, y + h) - F(x, y)}{F_Y(y + h) - F_Y(y)} \\
&\quad = \lim_{h \to 0+} \left(\frac{F(x, y + h) - F(x, y)}{h} \cdot \frac{h}{F_Y(y + h) - F_Y(y)} \right) \\
&\quad = \frac{\partial F(x, y)}{\partial y} \left(\frac{dF_Y(y)}{dy} \right)^{-1} \\
&\quad = \frac{\int_{-\infty}^{x} f(u, y)\, du}{f_Y(y)}.
\end{aligned}
$$

Uniform distribution [M&U Section 8.2]

A random variable X has uniform distribution over the interval $[a, b]$, denoted $X \sim U(a, b)$, if

$$\Pr(c \leq X \leq d) = \frac{d-c}{b-a}$$

for all $a \leq c \leq d \leq b$. The distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } b \leq x \end{cases}$$

and the density function

$$f(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } b < x. \end{cases}$$

A straightforward integration gives the expected value and variance as

$$\mathbf{E}[X] = \frac{b+a}{2}, \qquad \mathbf{E}[X^2] = \frac{b^2+ab+a^2}{3}, \qquad \mathbf{Var}[X] = \frac{(b-a)^2}{12}.$$

Lemma 2.2 [M&U Lemma 8.2]: If $X \sim U(a,b)$, then for all $c \leq d$ we have

$$\Pr(X \leq c \mid X \leq d) = \frac{c-a}{d-a}.$$

That is, X conditioned on $X \leq d$ has distribution $U(a,d)$.

Proof:

$$\begin{aligned}
\Pr(X \leq c \mid X \leq d) &= \frac{\Pr(X \leq c,\ X \leq d)}{\Pr(X \leq d)} \\
&= \frac{\Pr(X \leq c)}{\Pr(X \leq d)} \\
&= \frac{c-a}{d-a}.
\end{aligned}$$

□

Lemma 2.3: Assume $X_1, \ldots, X_n \sim U(0,1)$ are mutually independent. Define random variables $Y_1, \ldots, Y_n$ that have the same values as $X_1, \ldots, X_n$ but ordered from smallest to largest. Then $\mathbf{E}[Y_k] = k/(n+1)$.

Proof: To calculate $\mathbf{E}[Y_1]$ we notice that

$$\begin{aligned} \Pr(Y_1 \geq y) &= \Pr(X_1 \geq y, \ \ldots, \ X_n \geq y) \\ &= (1-y)^n. \end{aligned}$$

Therefore the distribution function of Y_1 is $1-(1-y)^n$. Differentiating, we see that the density function is $f(y) = n(1-y)^{n-1}$. We obtain the expected value by integration by parts:

$$\mathbf{E}[Y_1] = \int_0^1 yn(1-y)^{n-1}\,dy = \Big|_0^1(-y(1-y)^n) + \int_0^1 (1-y)^n\,dy = \frac{1}{n+1}.$$

We could generalize this technique to obtain the other expected values $\mathbf{E}[Y_i]$, too. Instead of doing that, however, we save some calculation effort by a symmetry argument.

We place random points $P_0, \ldots, P_n$ uniformly on a circle with circumference 1. Let X_i be the distance counterclockwise along the circle from P_0 to P_i. Clearly $X_i \sim U(0,1)$ and the random variables X_i are mutually independent. The points P_i divide the circumference of the circle into $n+1$ parts, and the length of part number j is $Y_j - Y_{j-1}$ where Y_i is as in the statement of the lemma (and we define $Y_0 = 0$). By symmetry, all parts have the same length distribution, and in particular the same expected length. Since the sum of the lengths of the parts is 1, by linearity of expectation each part has expected length $1/(n+1)$. Therefore

$$\mathbf{E}[Y_k] = \mathbf{E}[Y_0] + \sum_{i=1}^{k} \mathbf{E}[Y_i - Y_{i-1}] = \frac{k}{n+1}.$$

□

Exponential distribution [M&U Section 8.3]

To motivate the exponential distribution, recall Poisson distribution. Consider a process where some events occur in such a way that

1. the density of events per time unit stays constant and
2. in expectation we have θ events per time unit.

We can approximate this by a discrete-time process where at time j/n, for $j = 0, 1, 2, \ldots$, an event occurs with probability θ/n independently of the other events. When $T = j/n$ for some j, the number of events before time T is distributed as Bin$(Tn, \theta/n)$. When we make the time scale finer by letting $n \to \infty$, we get as limit the distribution Poisson$(T\theta)$.

Thus, Poisson($T\theta$) describes the number of events in a continuous-time process during the interval $[0, T)$. Let X be the time when the first event occurs. The distribution of X is then the exponential distribution with parameter θ, which we denote by $X \sim \text{Expon}(\theta)$.

Now $X \geq T$ if and only if no even occurred during the interval $[0, T)$. In other words, this means $Y = 0$ where $Y \sim \text{Poisson}(T\theta)$. Hence,

$$\Pr(X \leq T) = 1 - \Pr(Y = 0) = 1 - \mathrm{e}^{-T\theta}.$$

We have thus "derived" for the exponential distribution the distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \mathrm{e}^{-\theta x} & \text{if } x \geq 0. \end{cases}$$

The density function is therefore

$$f(x) = \theta \mathrm{e}^{-\theta x}, \quad x \geq 0.$$

Integration by parts yields

$$\begin{aligned} \mathbf{E}[X] &= \frac{1}{\theta} \\ \mathbf{E}[X^2] &= \frac{2}{\theta^2} \\ \mathbf{Var}[X] &= \frac{1}{\theta^2}. \end{aligned}$$

The exponential distribution is in some sense the continuous counterpart of the geometric distribution. In particular, it also has the property of being memoryless.

Lemma 2.4: If $X \sim \text{Expon}(\theta)$, then

$$\Pr(X > s + t \mid X > t) = \Pr(X > s).$$

Proof: Since $\Pr(X > r) = \mathrm{e}^{-\theta r}$, we have

$$\begin{aligned} \Pr(X > s + t \mid X > t) &= \frac{\Pr(X > s + t)}{\Pr(X > t)} \\ &= \frac{\exp(-\theta(s + t))}{\exp(-\theta t)} \\ &= \Pr(X > s). \end{aligned}$$

□

An exponentially distributed random variable can also be interpreted as the time a device keeps working, if its mean time between failures is $1/\theta$. This applies to malfunctions due to random disturbances, not due to wearing etc. which get worse over time.

Suppose now a device has n components, all of which must be working for the whole device to work. We assume the components have mutually independent failures, and the mean times between failure are $1/\theta_1, \ldots, 1/\theta_n$. We show that the time the device works is still exponentially distributed, and the mean time between failures is $1/(\theta_1 + \ldots + \theta_n)$.

Lemma 2.5 [M&U Lemma 8.5]**:** Let $X_1, \ldots, X_n$ be mutually independent with $X_i \sim \text{Expon}(\theta_i)$, and let $Y = \min\{X_1, \ldots, X_n\}$. Now $Y \sim \text{Expon}(\theta)$, where $\theta = \sum_{i=1}^{n} \theta_i$. Furthermore, $\Pr(Y = X_i) = \theta_i/\theta$.

Proof: We prove the case $n = 2$. The case $n > 2$ follows easily by induction.

Since

$$\begin{aligned} \Pr(\min\{X_1, X_2\} > x) &= \Pr(X_1 > x)\Pr(X_2 > x) \\ &= e^{-\theta_1 x} e^{-\theta_2 x} \\ &= e^{-(\theta_1+\theta_2)x}, \end{aligned}$$

we have $\min\{X_1, X_2\} \sim \text{Expon}(\theta_1 + \theta_2)$.

The joint density function of X_1 ja X_2 is $f(x,y) = \theta_1 \mathrm{e}^{-\theta_1 x} \theta_2 \mathrm{e}^{-\theta_2 x}$. Therefore,

$$\begin{aligned}
\Pr(X_1 < X_2) &= \int_0^\infty \int_{x_1}^\infty f(x_1, x_2)\, dx_2\, dx_1 \\
&= \int_0^\infty \theta_1 \mathrm{e}^{-\theta_1 x_1} \left(\int_{x_1}^\infty \theta_2 \mathrm{e}^{-\theta_2 x_2}\, dx_2 \right) dx_1 \\
&= \int_0^\infty \theta_1 \mathrm{e}^{-\theta_1 x_1} \mathrm{e}^{-\theta_2 x_1}\, dx_1 \\
&= \theta_1 \int_0^\infty \mathrm{e}^{-(\theta_1+\theta_2) x_1}\, dx_1 \\
&= \frac{\theta_1}{\theta_1 + \theta_2}.
\end{aligned}$$

□

Balls and bins with feedback [M&U Section 8.3.2]

As an application of combining exponential distributions, consider placing balls into two bins. However, this time the balls are not independent, but there is feedback: the next ball is more likely to go to the bin that has more balls. This is a crude model for, say, software market with two competitors, when compatibility reasons favor the one with the larger market share.

Suppose at a given time there are x balls in bin 1 and y balls in bin 2. We first consider a model where the next ball goes to bin 1 with probability $x/(x+y)$ and to bin 2 with probability $y/(x+y)$. Initially we place one ball to each bin. It is known ([M&U Exercise 1.6]; *Randomized Algorithms I*, homework 1.1) that when there are n balls, the number of balls in bin 1 is distributed uniformly over $\{1, \ldots, n-1\}$.

Consider now a situation in which the feedback is stronger. Let's have the next ball going to bin 1 with probability $x^p/(x^p + y^p)$ and to bin 2 with probability $y^p/(x^p + y^p)$, where $p > 1$, and we let the process continue indefinitely.

Theorem 2.6: Let $p > 1$, and assume that initially both bins have at least one ball. With probability 1 there is some finite value c such that one of the bins never gets more than c balls.

Proof: Let (x, y) denote the situation with x balls in bin 1 and y balls in bin 2. If we start from $(1, 1)$, then for any $x, y \geq 1$ there is a non-zero probability of reaching (x, y). Thus, if from some (x, y) we would have a non-zero probability of getting an unbounded number of balls in both bins, this would happen also starting from $(1, 1)$. Thus, without loss of generality we may assume that we start from $(1, 1)$.

As a helpful tool we consider a continuous-time process where the bins are independent:

- If bin 1 receives its ball number z at time t, then it receives its ball number $z+1$ at time $t+T_z$, where $T_z \sim \text{Expon}(z^p)$,
- If bin 2 receives its ball number z at time t, then it receives its ball number $z+1$ at time $t+U_z$, where $U_z \sim \text{Expon}(z^p)$,
- all random variables T_z and U_z are mutually independent

Perhaps surprisingly, this turns out to represent exactly the original process.

Consider the situation when a ball has just been added and there are x balls in bin 1 and y balls in bin 2. Because the exponential distribution is memoryless, it does not matter which bin received the latest ball. In any case, the expected remaining time until the next ball to bin 1 is $\text{Expon}(x^p)$, and to bin 2, $\text{Expon}(y^p)$. By Lemma 8.5, the next ball to arrive is for bin 1 with probability $x^p/(x^p+y^p)$.

Hence, the sequence of ball placements in this exponential model has the same distribution as in the original model.

Define the saturation times of the bins as $F_1 = \sum_{z=1}^{\infty} T_z$ and $F_2 = \sum_{z=1}^{\infty} U_z$. If $F_1 < \infty$, then

- if $t < F_1$, then at time t bin 1 contains a finite number of balls and
- as $t \to F_1-$, the number of balls in bin 1 at time t approaches infinity.

On the other hand, if $F_1 = \infty$, then the number of balls in bin 1 is finite at all times. The same characterization applies to the number of balls in bin 2.

Since $p > 1$, we have

$$\mathbf{E}[F_1] = \sum_{z=1}^{\infty} \mathbf{E}[T_z] = \sum_{z=1}^{\infty} \frac{1}{z^p} < \infty.$$

In particular, F_1 is finite with probability 1, and by the same argument, so is F_2.

Hence, we may assume that F_1 and F_2 are finite. If $F_1 = F_2$, then

$$T_1 = \sum_{z=1}^{\infty} U_z - \sum_{z=2}^{\infty} T_z.$$

Whatever value the right-hand side has, the probability of T_1 hitting the same value is zero. Therefore, $\Pr(Z_1 \neq Z_2) = 1$.

Consider the case $F_1 < F_2$; the case $F_1 > F_2$ is similar. Then there is n such that

$$\sum_{z=1}^{n} U_z < F_1 \leq \sum_{z=1}^{n+1} U_z.$$

Hence, there is m_0 such that for large enough $m \geq m_0$ we have

$$\sum_{z=1}^{n} U_z < \sum_{z=1}^{m} T_z < \sum_{z=1}^{n+1} U_z.$$

This means that bin 1 receives m balls before bin 2 receives $n+1$ balls. Since this holds for arbitrarily large m, bin 2 never receives more than n balls. □

Poisson processes [M&U Section 8.4]

Consider some events that take place over a continuous time interval. If $N(t)$ is the number of events during the interval $[0, t]$ for $t \geq 0$, we call $(N(t))_t$ a stochastic counting process.

A stochastic counting process N is a Poisson process with parameter λ, if $N(0) = 0$ and

1. increments are mutually independent: the differences $N(t_2) - N(t_1)$ and $N(t_4) - N(t_3)$ are independent if the intervals $[t_1, t_2]$ and $[t_3, t_4]$ are disjoint
2. increments are stationary: for all s and t, the difference $N(t + s) - N(s)$ has the same distribution as $N(t)$
3. $\lim_{t \to 0} \Pr(N(t) = 1)/t = \lambda$ and
4. $\lim_{t \to 0} \Pr(N(t) \geq 2)/t = 0$.

Condition 4 entails that occurence times of different events are independent. It would be violated for example if the events always occured two at a time.

Let N be a Poisson process with parameter λ and

$$P_n(t) = \mathsf{Pr}(N(t+s) - N(s) = n)$$

the probability of excetly n events in t time units. Because of stationarity, the number of events has Poisson distribution.

Theorem 2.7 [M&U Thm 8.7]: For all n we have

$$P_n(t) = \mathsf{e}^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

Proof: We use Conditions 3 and 4 to obtain a differential equation for P_n.

By independence, $P_0(t+h) = P_0(t)P_0(h)$, so

$$\begin{aligned}
\frac{P_0(t+h) - P_0(t)}{h} &= P_0(t)\frac{P_0(h) - 1}{h} \\
&= P_0(t)\frac{1 - \Pr(N(h) = 1) - \Pr(N(h) \geq 2) - 1}{h} \\
&\to P_0(t)(-\lambda + 0)
\end{aligned}$$

as $h \to 0$. Therefore,

$$P_0'(t) = -\lambda P_0(t).$$

By considering the initial condition $P_0(0) = 1$ we get the solution

$$P_0(t) = \mathrm{e}^{-\lambda t}.$$

For $n \geq 1$ we similarly get

$$P_n(t+h) - P_n(t) = P_n(t)(P_0(h) - 1) + P_{n-1}(t)P_1(h) + \sum_{k=2}^{n} P_{n-k}(t)P_k(h).$$

Based on the above, we have

$$P_0(h) = 1 - \lambda h + o(h).$$

By Condition 3,

$$P_1(h) = \lambda h + o(h).$$

Condition 4 then implies

$$0 \leq \sum_{k=2}^{n} P_{n-k}(t)P_k(h) \leq \Pr(N(h) \geq 2) = o(h).$$

Therefore,

$$P_n'(t) = \lim_{h \to 0} \frac{P_n(t+h) - P_n(h)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t).$$

To solve the equation

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$$

we write it as

$$\mathrm{e}^{\lambda t}(P_n'(t) + \lambda P_n(t)) = \lambda \mathrm{e}^{\lambda t} P_{n-1}(t)$$

and further as

$$\frac{d}{dt}\left(\mathrm{e}^{\lambda t} P_n(t)\right) = \lambda \mathrm{e}^{\lambda t} P_{n-1}(t).$$

We show by induction that for all n we have

$$P_n(t) = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

The base case $n = 0$ was done above. For $n \geq 1$, from the induction assumption

$$P_{n-1}(t) = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}$$

we get based on the above

$$\frac{d}{dt}\left(\mathrm{e}^{\lambda t} P_n(t)\right) = \lambda \mathrm{e}^{\lambda t} P_{n-1}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!}.$$

Integration yields

$$\mathrm{e}^{\lambda t} P_n(t) = \frac{(\lambda t)^n}{n!} + c,$$

and $P_n(0) = 0$ implies $c = 0$. □

Conversely, Poisson distributed event counts are also a sufficient condition for the process being Poisson.

Theorem 2.8 [M&U Thm 8.8]: If $\{ N(t) \mid t \geq 0 \}$ is a stochastic process with $N(0) = 0$ such that

1. increments are independent and
2. the number of events in t time units is Poisson(λt),

then $(N(t))$ is a Poisson process with parameter λ.

Proof: We have

$$\lim_{t\to 0} \frac{\Pr(N(t)=1)}{t} = \lim_{t\to 0} \frac{\mathrm{e}^{-\lambda t}\lambda t}{t} = \lambda$$

and

$$\lim_{t\to 0} \frac{\Pr(N(t)\geq 2)}{t} = \lim_{t\to 0} \sum_{k=2}^{\infty} \mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!t} = 0,$$

so Conditions 3 and 4 hold. Conditions 1 and 2 follow directly from the assumptions. □

Let X_1 be the time of the first event in a Poisson process. For $n \geq 1$, let X_n be the time between events $n-1$ and n We call the random variables X_n interarrival times.

Theorem 2.9 [M&U Thm 8.9]: The distribution of X_1 is Expon(λ).

Proof:

$$\Pr(X_1 \leq t) = 1 - \Pr(N(t) = 0) = 1 - \mathrm{e}^{-\lambda t}.$$

□

More generally,

Theorem 2.10 [M&U Thm 8.10]: The interarrival times X_n are mutually independent and all have distribution Expon(λ).

Proof:

$$\begin{aligned}
&\Pr(X_k \geq t_k \mid X_1 = t_1,\ \ldots,\ X_{k-1} = t_{k-1}) \\
&\quad = \quad \Pr\left(N\left(\sum_{i=1}^{k} t_i\right) - N\left(\sum_{i=1}^{k-1} t_i\right) = 0 \right) \\
&\quad = \quad \mathrm{e}^{-\lambda t_k}.
\end{aligned}$$

Hence, $\Pr(X_k \leq t) = 1 - \mathrm{e}^{-\lambda t}$ regardless of the values of $X_1, \ldots, X_{k-1}$. □

The reverse of the above also holds. Exponential interarrival times thus give a third characterization for a Poisson process.

Theorem 2.11 [M&U Thm 8.11]: If $\{ N(t) \mid t \geq 0 \}$ is a stochastic process such that $N(0) = 0$ and interarrival times are mutually mutually independent and Expon(λ), then $(N(t))$ is a Poisson process with parameter λ.

(Proof omitted.)

As we combined exponential distributions by Lemma 2.5 [M&U Lemma 8.5], we can combine Poisson processes. We say that processes $(N_1(t))$ ja $(N_2(t))$ are independent, if $N_1(t)$ and $N_2(s)$ are independent for all t, s.

Theorem 2.12 [M&U Thm 8.12]: If $(N_1(t))$ and $(N_2(t))$ are independent Poisson processes with parameters λ_1 and λ_2, respectively, then $(N_1(t) + N_2(t))$ is a Poisson process with parameter $\lambda_1 + \lambda_2$.

If we further interpret the process $(N_1(t) + N_2(t))$ as a combination of the events that constitute $(N_1(t))$ and $(N_2(t))$, then each event in the process $(N_1(t) + N_2(t))$ is with probability $\lambda_1/(\lambda_1 + \lambda_2)$ from process $(N_1(t))$.

Proof: The first part follows directly from the characterization of Poisson processes in terms of the event counts (Theorem 2.8 [M&U Thm 8.8]) and the fact that the sum of two independent Poisson random variables also has Poisson distribution (*Randomized Algorithms I*, Corollary 5.4 [M&U p. 97]).

The second part follows from the characterization of Poisson processes in terms of exponential interarrival distributions (Theorem 2.9) and the result about combining exponential distributions (Lemma 2.5) □

By induction, this generalises to more than two processes.

We can also split a Poisson process into two independent subprocesses.

Theorem 2.13 [M&U Thm 8.13]**:** Assume that $(N(t))$ is a Poisson process with parameter λ. Assign each event independently of each other to type 1 with probability p and type 2 with probability $1-p$. Let $(N_1(t))$ and $(N_2(t))$ be the processes constituted by the events of types 1 and 2, respectively. Then $(N_1(t))$ is a Poisson process with parameter $p\lambda$, and $(N_2(t))$ is a Poisson process with parameter $(1-p)\lambda$. Furthermore, $(N_1(t))$ and $(N_2(t))$ are independent.

Proof: Clearly $N_1(0) = 0$ and the increments of $(N_1(t))$ are independent. For the first part it therefore suffices to show that $N_1(t) \sim \text{Poisson}(p\lambda)$.

We get the distribution of N_1 as

$$\begin{aligned}
\Pr(N_1(t) = k) &= \sum_{j=k}^{\infty} \Pr(N_1(t) = k \mid N(t) = j) \Pr(N(t) = j) \\
&= \sum_{j=k}^{\infty} \frac{j!}{k!(j-k)!} p^k (1-p)^{j-k} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^j}{j!} \\
&= \mathrm{e}^{-\lambda p t} \frac{(\lambda p t)^k}{k!} \sum_{j=k}^{\infty} \mathrm{e}^{-\lambda(1-p)t} \frac{(\lambda(1-p)t)^{j-k}}{(j-k)!} \\
&= \mathrm{e}^{-\lambda p t} \frac{(\lambda p t)^k}{k!}
\end{aligned}$$

which is Poisson. The process N_2 is handled similarly.

To show independence, we first notice that $N_1(t)$ and $N_2(t)$ are independent for all t:

$$\begin{aligned}
\Pr(N_1(t) = m,\ N_2(t) = n) &= \Pr(N(t) = m + n,\ N_2(t) = n) \\
&= \mathrm{e}^{-\lambda t}\frac{(\lambda t)^{m+n}}{(m+n)!}\binom{m+n}{n}p^m(1-p)^n \\
&= \mathrm{e}^{-\lambda tp}\frac{(\lambda tp)^m}{m!}\cdot \mathrm{e}^{-\lambda t(1-p)}\frac{(\lambda t(1-p))^n}{n!} \\
&= \Pr(N_1(t) = m)\Pr(N_2(t) = n).
\end{aligned}$$

This implies independence of $N_1(t)$ and $N_2(u)$ also for $t \neq u$. For example, assume $u > t$. Since $N(u) - N(t)$ and $N(t)$ are independent, also $N_2(u) - N_2(t)$ is independent of $N_1(t)$ and $N_2(t)$. Therefore,

$$\begin{aligned}
&\Pr(N_1(t) = m,\ N_2(u) = n) \\
&\quad = \sum_{k=0}^{n} \Pr(N_1(t) = m,\ N_2(t) = k,\ N_2(u) - N_2(t) = n - k) \\
&\quad = \sum_{k=0}^{n} \Pr(N_1(t) = m,\ N_2(t) = k) \Pr(N_2(u) - N_2(t) = n - k) \\
&\quad = \Pr(N_1(t) = m) \sum_{k=0}^{n} \Pr(N_2(t) = k) \Pr(N_2(u) - N_2(t) = n - k) \\
&\quad = \Pr(N_1(t) = m) \Pr(N_2(u) = n).
\end{aligned}$$

□

The interarrival times of a Poisson process have exponential distribution.

However, it turns out that if we know the number of events during a given time interval, the arrival times are distributed uniformly over this interval.

As a preliminary, consider one event during $(0, t]$:

$$\begin{aligned}
\Pr(X_1 < s \mid N(t) = 1) &= \frac{\Pr(X_1 < s,\ N(t) = 1)}{\Pr(N(t) = 1)} \\
&= \frac{\Pr(N(s) = 1,\ N(t) - N(s) = 0)}{\Pr(N(t) = 1)} \\
&= \frac{(\lambda s \mathrm{e}^{-\lambda s})\mathrm{e}^{-\lambda(t-s)}}{\lambda t \mathrm{e}^{-\lambda t}} \\
&= \frac{s}{t},
\end{aligned}$$

for $s \leq t$.

More generally, when $X_1, \ldots, X_n$ are mutually independent, their order statistics are the random variables $Y_{(1)}, \ldots, Y_{(n)}$ where the value of $Y_{(i)}$ is the ith largest of the values of $X_1, \ldots, X_n$. We assume $X_i \neq X_j$ for $i \neq j$, which for continuous distributions holds with probability 1. Hence, if $X_i = x_i$ and $Y_{(i)} = y_i$, then $\{ x_1, \ldots, x_n \} = \{ y_1, \ldots, y_n \}$ ja $y_1 < \ldots < y_n$.

Theorem 2.14 [M&U Thm 8.14]: The arrival times during $(0, t]$ of a Poisson process $(N(t))$ with condition $N(t) = n$ have the same distribution as the order statistics of n mutually independent random variables $X_1, \ldots, X_n$ where X_i is uniform over $[0, t]$.

Proof: Denote the order statistics by $Y_{(1)}, \ldots, Y_{(n)}$. We first calculate the joint distribution of $Y_{(i)}$, then the conditional distribution of arrival times, and notice that the distributions are the same.

Let $\mathcal{E}$ denote the event

$$Y_{(1)} \le s_1,\ Y_{(2)} \le s_2,\ \ldots,\ Y_{(n)} \le s_n.$$

For a permutation σ of the set $\{1, \ldots, n\}$, let $\mathcal{E}_\sigma$ denote the event

$$X_{\sigma(1)} \le s_1,\ X_{\sigma(1)} \le X_{\sigma(2)} \le s_2,\ \ldots,\ X_{\sigma(n-1)} \le X_{\sigma(n)} \le s_n.$$

Omitting the cases where $X_i = X_j$ for some $i \neq j$, we gets $\mathcal{E} = \cup_\sigma \mathcal{E}_\sigma$. By symmetry, $\Pr(\mathcal{E}) = n!\Pr(\mathcal{E}_\sigma)$ for any σ. In particular,

$$\begin{aligned}\Pr(\mathcal{E}) &= n!\Pr(X_1 < s_1,\ X_1 < X_2 < s_2,\ \ldots,\ X_{n-1} < X_n < s_n) \\ &= \frac{n!}{t^n}\int_0^{s_1}\int_{u_1}^{s_2}\cdots\int_{u_{n-1}}^{s_n} du_n \ldots du_2\, du_1,\end{aligned}$$

where $1/t$ comes from the uniform density over $[0, t]$.

On the other hand, let $S_1, \ldots, S_{n+1}$ be the $n+1$ first arrival times, $T_1 = S_1$ and $T_i = S_i - S_{i-1}$ for $i > 1$. We know that without the conditioning, the interarrival times T_i are independent and exponentially distributed, with density $\lambda \mathrm{e}^{-\lambda t}$. Therefore,

$$\begin{aligned}
&\Pr(S_1 \le s_1,\ S_2 \le s_2,\ \ldots,\ S_n \le s_n,\ N(t) = n) \\
&\quad = \Pr\left(T_1 \le s_1,\ T_2 \le s_2 - T_1,\ \ldots,\ T_n \le s_n - \sum_{i=1}^{n-1} T_i,\ T_{n+1} > t - \sum_{i=1}^{n} T_i\right) \\
&\quad = \int_0^{s_1} \int_0^{s_2 - t_1} \cdots \int_0^{s_n - \sum_{i=1}^{n-1} t_i} \int_{t - \sum_{i=1}^{n}}^{\infty} \lambda^{n+1} \exp\left(-\lambda \sum_{i=1}^{n+1} t_i\right) dt_{n+1}\, dt_n \ldots dt_2 dt_1.
\end{aligned}$$

We evaluate the innermost integral:

$$\int_{t - \sum_{i=1}^{n}}^{\infty} \lambda^{n+1} \exp\left(-\lambda \sum_{i=1}^{n+1} t_i\right) dt_{n+1} = \left|_{t_{n+1} = t - \sum_{i=1}^{n} t_i}^{\infty} \left(-\lambda^n \exp\left(-\lambda \sum_{i=1}^{n+1} t_i\right)\right) = \lambda^n \mathrm{e}^{-\lambda t}.\right.$$

By writing $u_j = \sum_{i=1}^{j} t_i$ we get

$$\begin{aligned}
&\Pr(S_1 \le s_1,\ S_2 \le s_2,\ \ldots,\ S_n \le s_n,\ N(t) = n) \\
&\quad = \int_0^{s_1} \int_0^{s_2 - t_1} \cdots \int_0^{s_n - \sum_{i=1}^{n-1} t_i} \lambda^n \mathrm{e}^{-\lambda t} dt_n \ldots dt_2 dt_1 \\
&\quad = \lambda^n \mathrm{e}^{-\lambda t} \int_0^{s_1} \int_{u_1}^{s_2} \cdots \int_{u_{n-1}}^{s_n} du_n \ldots du_2 du_1.
\end{aligned}$$

Since $\Pr(N(t) = n) = \mathrm{e}^{-\lambda t}(\lambda t)^n / n!$, we get

$$\begin{aligned}
&\Pr(S_1 \le s_1,\ S_2 \le s_2,\ \ldots,\ S_n \le s_n \mid N(t) = n) \\
&\quad = \frac{\Pr(S_1 \le s_1,\ S_2 \le s_2,\ \ldots,\ S_n \le s_n,\ N(t) = n)}{\Pr(N(t) = n)} \\
&\quad = \frac{n!}{t^n} \int_0^{s_1} \int_{u_1}^{s_2} \cdots \int_{u_{n-1}}^{s_n} du_n \ldots du_2 du_1
\end{aligned}$$

which is the same as $\Pr(\mathcal{E})$. □

Continuous time Markovin processes [M&U Section 8.5]

A process $\{ X(t) \mid t \geq 0 \}$ is a (time-homogeneous) Markov process, if for all $s, t \geq 0$ we have

$$\Pr(X(t+s) = x \mid X(u), 0 \leq u \leq t) = \Pr(X(t+s) = x \mid X(t))$$

and this probability is the same for all t.

We are interested in discrete space Markov processes, for which the range of $X(t)$ is countable.

In discrete time, the properties of a Markov chain were represented in a transition matrix. With a continuous time Markov process, we can identify two subprocesses:

1. the embedded Markov chain, where the element $p_{i,j}$ of the transition matrix gives the probability that the next state is j, if current state is i.
2. parameters θ_i which determine the times the process spends in each state, so that the time spent in state i before moving on has distribution Expon(θ_i).

The memoryless exponential distribution is essential to keep the whole process Markovian.

We will just have a quick look into stationary distributions without going much into the theory.

As with discete time, the stationary distribution is such that if the process has that distribution at some time, it keeps the same distribution in the future. Further, under some additional assumptions, the process converges towards the stationary distribution regardless of where it started.

Let $P_{j,i}(t)$ be the probability of being in state i at time t, when at time 0 the process is at state j. Hence, under suitable assumptions, the stationary distribution $\boldsymbol{\pi}$ satisfies

$$\lim_{t\to\infty} P_{j,i}(t) = \pi_i$$

for all i and j. If the process approaches the stationary distribution, then in particular the derivative must satisfy

$$\lim_{t\to\infty} P'_{j,i}(t) = 0.$$

To find the stationary distribution, we calculate $\lim_{t\to\infty} P'(t)$ in another way in terms of the stationary distribution $\boldsymbol{\pi}$.

$$
\begin{aligned}
P'_{j,i}(t) &= \lim_{h\to 0} \frac{P_{j,i}(t+h) - P_{j,i}(t)}{h} \\
&= \lim_{h\to 0} \frac{\sum_k P_{j,k}(t) P_{k,i}(h) - P_{j,i}(t)}{h} \\
&= \lim_{h\to 0} \left(\sum_{k\neq i} \frac{P_{k,i}(h)}{h} P_{j,k}(t) - \frac{1 - P_{i,i}(h)}{h} P_{j,i}(t) \right).
\end{aligned}
$$

We can think of the process leaving state k as a Poisson process with parameter θ_k. Hence, as $h \to 0$, the probability of making one transition in time h is asymptotically $\theta_k h$, and the probability of making two transitions is zero. Therefore,

$$
\lim_{h\to\infty} \frac{P_{k,i}(h)}{h} = \theta_k p_{k,i} \qquad \text{and} \qquad \lim_{h\to\infty} \frac{1 - P_{i,i}(h)}{h} = \theta_i (1 - p_{i,i}).
$$

Assuming we can change the order of limit and summation (which is clear at least for finite state spaces), we get

$$\begin{aligned} P'_{j,i}(t) &= \sum_{k \neq i} \theta_k p_{k,i} P_{j,k}(t) - P_{j,i}(t)(\theta_i - \theta_i p_{i,i}) \\ &= \sum_{k} \theta_k p_{k,i} P_{j,k}(t) - P_{j,i}(t)\theta_i. \end{aligned}$$

Assuming $\lim_{t\to\infty} P_{j,i}(t) = \pi_i$ we get

$$\lim_{t\to\infty} P'_{j,i}(t) = \sum_{k} \theta_k p_{k,i} \pi_k - \theta_i \pi_i.$$

Since this must be 0, we get

$$\pi_i \theta_i = \sum_{k} \pi_k \theta_k p_{k,i}.$$

Thus, the stationary distribution satisfies rate equations

$$\pi_i \theta_i = \sum_k \pi_k \theta_k p_{k,i}$$

which have an intuitive interpretation:

- $\pi_i \theta_i$ is the rate of transitions leaving state i
- $\sum_k \pi_k \theta_k p_{k,i}$ is the rate of transitions entering state i.

In particular, if $\theta_i = \theta$ for all i, we get

$$\pi_i = \sum_k \pi_k p_{k,i}.$$

In this special case, the stationary distribution of the continuous time Markov process is the same as that of the embedded discrete time chain.

Markovian queues [M&U Section 8.6]

Consider a single server with a queue. The arrival times of customers follow a Poisson process. Arriving customers join the end of the queue. The customer at the front of the queue gets served, and the service time he needs has exponential distribution.

We denote this model by $M/M/1$:

- first M stands for memoryless arrival distribution
- second M stands for memoryless service time distribution
- there is 1 server.

We model this as Markov process where queue lengths are the states.

We denote by λ the parameter for the Poisson process giving the arrival times, and by μ the parameter of the exponential distribution of the service times. Let $M(t)$ be the queue length at time t, and $P_k(t) = \mathsf{Pr}(M(t) = k)$.

Consider first the stationary distribution. Since in a time interval of length h, the probability of a customer arriving is $\lambda h + o(h)$, and of leaving, $\mu h + o(h)$, we get

$$\begin{aligned} P_0'(t) &= \lim_{h\to 0} \frac{P_0(t+h) - P_0(t)}{h} \\ &= \lim_{h\to 0} \frac{P_0(t)(1-\lambda h) + P_1(t)\mu h - P_0(h)}{h} \\ &= -\lambda P_0(t) + \mu P_1(t) \end{aligned}$$

and for $k \geq 1$ similarly

$$P_k'(t) = \lambda P_{k-1}(t) - (\lambda + \mu) P_k(t) + \mu P_{k+1}(t).$$

In the stationary situation, $P_k'(t) = 0$ for all k. Hence, if $\boldsymbol{\pi}$ is the stationary distribution, we have

$$\mu\pi_1 = \lambda\pi_0.$$

From this we get easily by induction

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^k$$

for all k. The normalization $\sum_k \pi_k = 1$ then implies $\pi_0 = 1 - \lambda/\mu$ assuming $\lambda/\mu < 1$. It can be shown that in this case the process does converge to this stationary distribution.

If $\lambda/\mu \geq 1$, then $\sum_k \pi_k$ does not converge for any positive π_0. Therefore, there is no stationary distribution.

To calculate the expected queue length, we write the stationary probability for queue length $k \geq 1$ as

$$\pi_k = \frac{\lambda}{\mu} p_k$$

where

$$p_k = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{k-1}.$$

Notice that p_k is the probability that a random variable with Geom$(1 - \lambda/\mu)$ distribution gets value k. Since the expected value of Geom(q) is $1/q$, the expected queue length is

$$L = \frac{\lambda}{\mu} \cdot \frac{1}{1 - \lambda/\mu} = \frac{\lambda}{\mu - \lambda}.$$

(Again, we assume $\lambda < \mu$.)

All this holds regardless the order in which customers are served. Let us now analyse from an individual customer's point of view, how long he spends in the queue. For this we assume that customers are served on a first in, first out principle.

Let $L(k)$ be the event that the queue length is k when the customer arrives. The expected time W that a customer spends in the system is

$$\begin{aligned} W &= \sum_{k=0}^{\infty} \mathbf{E}[W \mid L(k)] \Pr(L(k)) \\ &= \sum_{k=0}^{\infty} \frac{k+1}{\mu} \Pr(L(k)), \end{aligned}$$

where we used the fact that because of the distributions are memoryless, the expected remaining service time for all customers is $1/\mu$. We next calculate $\Pr(L(k))$.

When the process is stationary, the rate of leaving state k is $\pi_k\theta_k$, where $\theta_0 = \lambda$ and $\theta_k = \lambda + \mu$ for $k \geq 1$. By the result about combining exponential distributions (Lemma 2.5 [M&U Lemma 8.5]), the probability that leaving state k is due to an arrival of a new customer is λ/θ_k. Hence, the rate of events where a customer arrives to find a queue length k is

$$\pi_k\theta_k \cdot \frac{\lambda}{\theta_k} = \pi_k\lambda.$$

By again applying Lemma 2.5, we see that the arrival event is with probability π_k such that the queue length is k:

$$\Pr(L(k)) = \pi_k.$$

This is an instance of so-called PASTA principle (Poisson Arrivals See Time Averages).

We can now calculate the expected waiting time:

$$
\begin{aligned}
W &= \sum_{k=0}^{\infty} \frac{k+1}{\mu} \pi_k \\
&= \frac{1}{\mu} \left(1 + \sum_{k=0}^{\infty} k \pi_k \right) \\
&= \frac{1}{\mu} (1 + L) \\
&= \frac{1}{\mu} \left(1 + \frac{\lambda}{\mu - \lambda} \right) \\
&= \frac{1}{\mu - \lambda} \\
&= \frac{L}{\lambda}.
\end{aligned}
$$

The end result $L = \lambda W$ holds also more generally in various stable queue systems.

By joining and splitting Poisson processes we can reduce more general queue systems into $M/M/1$; for example several independent arrival processes, or several independent servers.

We could also consider queue systems of type $M/M/1/K$, where the queue length has an upper limit K. If a customer arrives when the queue length is K, he will leave immediately.

Essentially the same calculation as earlier gives then

$$\pi_k = \begin{cases} \pi_0(\lambda/\mu)^k & \text{for } k \leq K \\ 0 & \text{for } k > K. \end{cases}$$

Then

$$\pi_0 = \left(\sum_{k=0}^{K} \left(\frac{\lambda}{\mu} \right)^k \right)^{-1}$$

and a stationary distribution exists regardless of whether $\lambda < \mu$ or not.

Consider now a situation in which a customer can enter some service, spend some time there, and then leave. There can be an arbitrary number of customers in the service at any time. If the arrival and service times are again memoryless, this can be modelled as an $M/M/\infty$ queue. Every customer has his own queue, and there are infinitely many queues available.

Consider first the stationary distribution. If the process is in state k (meaning that there are k customers in service), the next event comes from a combination of $k+1$ Poisson processes. There are k processes representing each customer leaving, and one process for arrivals.

Hence, the time for the next event at state k has Poisson distribution with parameter

$$\theta_k = k\mu + \lambda,$$

and the event is an arrival with probability $\lambda/(k\mu + \lambda)$. The transition probabilities between states are

$$\begin{aligned} p_{k,k+1} &= \frac{\lambda}{k\mu + \lambda} \\ p_{k,k-1} &= \frac{k\mu}{k\mu + \lambda}. \end{aligned}$$

The stationary distribution $\boldsymbol{\pi}$ satisfies conditions

$$\pi_k \theta_k = \pi_{k-1}\theta_{k-1}p_{k-1,k} + \pi_{k+1}\theta_{k+1}p_{k+1,k}$$

which thus become

$$\pi_k(k\mu + \lambda) = \pi_{k-1}\lambda + \pi_{k+1}(k+1)\mu$$

(where for the case $k = 0$ we define $\pi_{-1} = 0$).

Write the condition

$$\pi_k(k\mu + \lambda) = \pi_{k-1}\lambda + \pi_{k+1}(k+1)\mu$$

as

$$\pi_{k+1}(k+1)\mu - \pi_k\lambda = \pi_k k\mu - \pi_{k-1}\lambda.$$

Because $\pi_0 \cdot 0 \cdot \mu - \pi_{-1}\lambda = 0$, for all k we have

$$\pi_{k+1}(k+1)\mu - \pi_k\lambda = 0.$$

Therefore,

$$\pi_{k+1} = \frac{\lambda}{\mu(k+1)}\pi_k,$$

so

$$\pi_k = \pi_0 \prod_{j=1}^{k} \frac{\lambda}{\mu j} = \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}.$$

The normalization condition now gives the final result

$$\pi_k = \mathrm{e}^{-\lambda/\mu}\frac{(\lambda/\mu)^k}{k!},$$

so the stationary distribution is Poisson(λ/μ).

As an alternative solution, let $M(t)$ be the number of customers being served at time t. Let $N(t)$ be the number of customers who arrived during time interval $[0, t]$. Since $N(t)$ is a Poisson process with parameter λ, we have

$$\Pr(M(t) = j) = \sum_{n=j}^{\infty} \Pr(M(t) = j \mid N(t) = n) \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

By Theorem 2.14 [M&U Thm 8.14] about conditional arrival times, the arrival times of the n first customers conditioned on $N(t) = n$ are uniform over $[0, t]$.

A customer who arrived at time $x < t$ is at time t still in the system with probability $\mathrm{e}^{-\mu(t-x)}$.

Thus, if we consider one fixed customer among the first n customers to arrive, then the probability of his still being in the system at time t is

$$p = \int_0^t \mathrm{e}^{-\mu(t-x)} \cdot \frac{1}{t} \cdot dx = \frac{1}{\mu t} \left(1 - \mathrm{e}^{-\mu t}\right).$$

Since the customers are independent,

$$\Pr(M(t) = j \mid N(t) = n) = \binom{n}{j} p^j (1-p)^{n-j}.$$

Therefore,

$$\begin{aligned}
\Pr(M(t) = j) &= \sum_{n=j}^{\infty} \binom{n}{j} p^j (1-p)^{n-j} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
&= \mathrm{e}^{-\lambda t} \frac{(\lambda t p)^j}{j!} \sum_{n=j}^{\infty} \frac{(\lambda t (1-p))^{n-j}}{(n-j)!} \\
&= \mathrm{e}^{-\lambda t} \frac{(\lambda t p)^j}{j!} \mathrm{e}^{\lambda t (1-p)} \\
&= \mathrm{e}^{-\lambda t p} \frac{(\lambda t p)^j}{j!}.
\end{aligned}$$

The number of customers at time t is Poisson(λtp). The parameter of the distribution approaches the value

$$\lim_{t \to \infty} \lambda t p = \lim_{t \to \infty} \lambda t \frac{1}{\mu t} \left(1 - \mathrm{e}^{-\mu t}\right) = \frac{\lambda}{\mu}.$$

3. The Monte Carlo method

Monte Carlo method is a generic name for a style of randomized algorithms which most typically estimate some numerical quantity by

1. defining a random variable with the desired quantity as its expected value and
2. calculating the average of a sufficiently large sample of independent draws of the random variable.

We need to choose the random variable so the its values are reasonably well concentrated around the expected value. In particular, we must pay attention to the fact that if the target value we try to estimate is close to zero, the relative errors can easily become very large.

Because of these considerations, the suitable distributions may be complicated and sampling from them difficult. One general technique for this is Markov Chain Monte Carlo (MCMC), where sampling is done using a Markov chain with a suitable stationary distribution.

Basics of Monte Carlo [M&U Section 10.1]

The classic introductory example is estimating the value of π. Choose $X \in [-1, 1]$ and $Y \in [-1, 1]$ from the uniform distribution and let

$$Z = \begin{cases} 1 & \text{if } X^2 + Y^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The square $[-1, 1] \times [-1, 1]$ has area 4, and the unit disk has area π, so

$$\mathbf{E}[Z] = \Pr(Z = 1) = \frac{\pi}{4}.$$

Let $W = \sum_{i=1}^{m} Z_i$, where Z_i are independent copies of Z and m is a sample size to be determined later. Let

$$W' = \frac{4}{m} W = \frac{4}{m} \sum_{i=1}^{m} Z_i.$$

Then $\mathbf{E}[W'] = \pi$, and we can apply a Chernoff bound to get

$$\begin{aligned} \Pr\left(\frac{|W' - \pi|}{\pi} \geq \varepsilon\right) &= \Pr\left(\left|W - \frac{m\pi}{4}\right| \geq \frac{m\pi\varepsilon}{4}\right) \\ &= \Pr(|W - \mathbf{E}[W]| \geq \varepsilon \mathbf{E}[W]) \\ &\leq 2\exp\left(-m\pi\varepsilon^2/12\right). \end{aligned}$$

Hence, for any $\varepsilon, \delta > 0$, if we choose

$$m \geq \frac{12\ln(2/\delta)}{\pi\varepsilon^2},$$

then with probability at least $1 - \delta$ the value W' approximates π with relative error at most ε.

More generally, if we estimate a quantity V with a randomized algorithm such that the output X of the algorithm satisfies

$$\Pr(|X - V| \leq \varepsilon V) \geq 1 - \delta,$$

we say the algorithm is an (ε, δ) approximation algorithm.

Hence, the previous sampling method gives an (ε, δ) approximation for π, as long as $m \geq 12 \ln(2/\delta)/(\pi\varepsilon^2)$. From Chernoff bounds we get more generally

Theorem 3.1 [M&U Thm 10.1]: Let X_i, $i = 1, \ldots, m$ be independent identically distributed random variables with $\mathbf{E}[X_i] = \mu$, and let $X = (1/m) \sum_{i=1}^{m} X_i$. If

$$m \geq \frac{3}{\varepsilon^2 \mu} \ln \frac{2}{\delta},$$

then X is an (ε, δ) approximation for μ. □

For appying the previous result it is important to notice that μ appears in the denominator of the sample size m. Since we use relative error as our criterion, small quantities are difficult to estimate accurately.

Example 3.2: Let $B_n = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i^2 \leq 1 \right\}$ be the n dimensional unit ball and V_n its volume. We draw m random points X_i from the uniform distribution over $[-1,1]^n$. Let $Z_i = 1$ if $X_i \in B_n$, and $Z_i = 0$ otherwise. Then $\mu = \mathbf{E}[Z_i] = V_n/2^n$.

It is well known that

$$V_n = \frac{\pi^{n/2}}{\Gamma(1+n/2)} \approx (2\pi)^{-1/2} \left(\frac{2e\pi}{n} \right)^{n/2},$$

where $\Gamma(n+1) = n!$ and we used Stirling's approximation. We conclude that the sampling method will not give a good estimate for the volume of V_n if the number of dimensions n is high. □

This is a basic example of a fairly common scenario, where the size of the "interesting" set (here B_n) is vanishigly small compared to the "obvious" sample space (here$[-1,1]^n$). Hence, more refined sampling strategies are needed.

More generally, let $V(x)$ be some function depending on input x, and let $|x|$ denote the size of x.

An algorithm $A(\cdot, \cdot, \cdot)$ is a polynomial randomized approximation scheme for V, if for any x and for all $\varepsilon, \delta > 0$, the algorithm $A(x, \varepsilon, \delta)$ runs in time poly($|x|$) and gives an (ε, δ) approximation for $V(x)$.

An algorithm $A(\cdot, \cdot, \cdot)$ is a fully polynomial randomized approximation scheme (FPRAS), if additionally the run time of $A(x, \varepsilon, \delta)$ is also polynomial in $1/\varepsilon$ and $\ln(1/\delta)$.

The main difference in polynomial and fully polynomial schemes is that the former allows running times such as $O(n^{1/\varepsilon})$, the latter does not. The dependence on δ is usually not an issue.

Counting satisfying assignments [M&U Section 10.2]

For a Boolean formula $\varphi(x_1, \ldots, x_n)$, let $c(\varphi)$ be the number of truth value assignments $\boldsymbol{x} \in \{0,1\}^n$ that satisfy it. For example, $c(\varphi) \geq 1$ for satisfiable formulas, and $c(\varphi) = 2^n$ for tautologically true formulas. There are two important computational problems related to c.

CNF counting:
Given: formula φ in conjunctive normal form (CNF)
Task: compute $c(\varphi)$

DNF counting
Given: formula φ in disjunctive normal form (DNF)
Task: compute $c(\varphi)$.

If φ is a CNF formula, then its negation $\overline{\varphi}$ is a DNF formula of roughly same size, and $c(\overline{\varphi}) = 2^n - c(\varphi)$. Hence, if an exact answer is required, the problems have essentially the same computational complexity.

The DNF and CNF counting problems in exact form are known to be complete for a class called $\sharp P$ ("number P" or "sharp P"). A function f belongs to $\sharp P$, if there is a polynomial time nondeterministic Turing machine M such that the value $f(x)$ is the same as the number of accepting computations of M with input x.

However, from an approximation point of view, the DNF and CNF counting problems are very different. The NP complete SAT problem is the same as asking whether $c(\varphi) > 0$ holds for a CNF formula φ. Hence, approximating CNF counting with any relative error strictly less than 100% would solve SAT.

In the DNF case, the NP complete problem is deciding between the cases $c(\varphi) = 2^n$ and $c(\varphi) \leq 2^n - 1$, and the relative difference between 2^n and $2^n - 1$ is very small. (The satisfiability problem is trivial for DNF formulas.)

We give a FPRAS for DNF counting. For CNF counting, the existence of a FPRAS would imply that any problem in NP could be solved in polynomial time at least if some reasonable model of randomization is allowed.

Our first attempt is a straight generalization from the Monte Carlo algorithm for π. For $i = 1, \ldots, m$, let $\boldsymbol{V}_i \in \{0,1\}^n$ be independent random value assingments, and $X_i = 1$ if $\varphi(\boldsymbol{V}_i) = 1$.

Now $\mathbf{E}[X_i] = c(\varphi)/2^n$. Hence, for

$$Y = \frac{2^n}{m} \sum_{i=1}^{m} X_i,$$

we have $\mathbf{E}[Y] = c(\varphi)$. Then Y gives a (ε, δ) approximation, if

$$m \geq \frac{3 \ln(2/\delta)}{\varepsilon^2} \cdot \frac{2^n}{c(\varphi)}.$$

If $c(\varphi)$ is polynomial, this only gives an exponential upper bound. A closer analysis (which we omit here) shows that this is not an artefact of any loose approximations in our proof, and an exponential sample size really is necessary.

For an improved sampling method, write $\varphi = C_1 \vee \ldots \vee C_t$, where each term C_i is a conjunction with ℓ_i literals. Let $R(\psi) \subseteq \{0,1\}^n$ be the number of assignments that satisfy ψ. So in general, $c(\psi) = |R(\psi)|$, and here in particular,

$$c(C_i) = |R(C_i)| = 2^{n-\ell_i}.$$

Therefore, it is easy to calculate

$$\sum_{i=1}^{t} c(C_i) = \sum_{i=1}^{t} 2^{n-\ell_i}.$$

Since $R(\varphi) = \cup_i R(C_i)$, we have

$$\frac{c(\varphi)}{\sum_{i=1}^{t} c(C_i)} = \alpha$$

for some $0 \leq \alpha \leq 1$, and therefore $c(\varphi) = \alpha \sum_i 2^{n-\ell_i}$. Our plan is to

1. define a set U such that $|U| = \sum_{i=1}^{t} c(C_i)$,
2. define a set $S \subseteq U$ such that $|S| = c(\varphi)$ and
3. estimate $\alpha = |S| / |U|$ by sampling from U.

First, we define

$$U = \{\, (i, \boldsymbol{x}) \in \{\, 1, \ldots, t \,\} \times \{\, 0, 1 \,\}^n \mid \boldsymbol{x} \in R(C_i) \,\}\,.$$

Clearly

$$|U| = \sum_{i=1}^{t} |R(C_i)| = \sum_{i=1}^{t} 2^{n-\ell_i}.$$

Now we define

$$S = \{\, (i, \boldsymbol{x}) \in U \mid (j, \boldsymbol{x}) \notin U \text{ for } j < i \,\}\,.$$

Then

$$|S| = |\cup_i R(C_i)| = c(\varphi),$$

so

$$c(\varphi) = \frac{|S|}{|U|} \sum_{i=1}^{t} 2^{n-\ell_i}.$$

Additionally, $|U| \leq t\,|S|$, so the ratio $\alpha = |S|\,/\,|U|$ can be approximated efficiently, if we know how to sample uniformly from U.

We claim that the following sampling procedure produces pairs $(i, \boldsymbol{x})$ according to the uniform distribution over U:

1. Choose $i \in \{1, \ldots, t\}$ randomly so that the probability of choosing i is

$$\frac{c(C_i)}{\sum_{j=1}^{t} c(C_j)} = \frac{c(C_i)}{|U|}.$$

2. Choose $\boldsymbol{x} \in \{0,1\}^n$ so that the ℓ_i literals in C_i are satisfied and the remaining $n - \ell_i$ variables are assigned random values uniformly and independently.

Clearly stage 2 samples $\boldsymbol{x}$ uniformly from $R(C_i)$, so

$$\begin{aligned}
\Pr((i,\boldsymbol{x}) \text{ is chosen}) &= \Pr(i \text{ is chosen}) \cdot \Pr((i,\boldsymbol{x}) \text{ chosen} \mid i \text{ is chosen}) \\
&= \frac{c(C_i)}{|U|} \cdot \frac{1}{c(C_i)} \\
&= \frac{1}{|U|}.
\end{aligned}$$

Hence, the distribution is uniform.

We have the following algorithm.

$X := 0$
Repeat for $k = 1, \ldots, m$:
    Choose a random $(i, \boldsymbol{x}) \in U$.
    If $C_j(\boldsymbol{x}) = 0$ for all $j < i$, then $X := X + 1$.
Return $(X/m) \sum_{i=1}^{t} 2^{n-\ell_i}$.

Based on the above, this given an (ε, δ) approximation when

$$m \geq \frac{3t \ln(2/\delta)}{\varepsilon^2}.$$

Hence, we have an FPRAS.

From sampling to counting [M&U Section 10.3]

The DNF counting algorithm was an example of how the cardinality $|S|$ of a set S is estimated as

$$|S| = \frac{|S|}{|U|}\,|U|\,,$$

where U is such that

- $|U|$ is known
- we have an efficient method for uniform sampling from U and
- $|S|\,/\,|U|$ is not too small.

We consider a general method for finding a suitable sample space U.

If the random output w of a sampling algorithm A satisfies

$$\left| \Pr(w \in S) - \frac{|S|}{|\Omega|} \right| \leq \varepsilon$$

for all $S \subseteq \Omega$, we say that A generates an ε uniform sample of Ω.

If instances x of some computational problem are associated with a sample space $\Omega(x)$, we call A a fully polynomial almost uniform sampler (FPAUS) for this problem, if A, given as input any $\varepsilon > 0$ and x, generates an ε uniform sample of $\Omega(x)$ in running time which is polynomial in $\ln(1/\varepsilon)$ and the size of x

We are interested in the setting where $\Omega(x)$ is the set of solutions to a problem instance x and we wish to estimate $|\Omega(x)|$.

As an example, consider the number of independent sets in a graph $G = (V, E)$. Thus, we take as $\Omega(G)$ the set of independent sets in G.

Let $E = \{ e_1, \ldots, e_m \}$, where $m = |E|$, and $G_i = (V, \{ e_1, \ldots, e_i \})$ for $i = 0, \ldots, m$. Hence, $\Omega(G_{i+1}) \subset \Omega(G_i)$, and

$$|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \cdot \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \cdot \frac{|\Omega(G_{m-2})|}{|\Omega(G_{m-3})|} \cdot \ldots \cdot \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \cdot |\Omega(G_0)| \,.$$

We know that $\Omega(G_0) = 2^n$. We shall next show that the ratio

$$r_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$$

can be estimated with sufficient accuracy, if we have access to an almost uniform sampler of $\Omega(G_{i-1})$. We shall later return to the question of constructing such samplers.

Let $\tilde{r}_i$ be the estimate we got for r_i by sampling. Thus, the output of the algorithm is $2^n \prod_{i=1}^m \tilde{r}_i$, whereas the correct answer is $2^n \prod_{i=1}^m r_i$. The relative error is

$$\left| \frac{2^n \prod_{i=1}^m \tilde{r}_i - 2^n \prod_{i=1}^m r_i}{2^n \prod_{i=1}^m r_i} \right| = \left| \prod_{i=1}^m \frac{\tilde{r}_i}{r_i} - 1 \right|.$$

We first estimate this total error in terms of the errors related to each individual estimate $\tilde{r}_i$.

Lemma 3.3: If $\tilde{r}_i$ is an $(\varepsilon/(2m), \delta/m)$ approximation for all i, and $0 < \varepsilon, \delta < 1$, then

$$\Pr\left(\left| \prod_{i=1}^m \frac{\tilde{r}_i}{r_i} - 1 \right| \geq \varepsilon \right) \leq \delta.$$

Proof: We assume

$$\Pr\left(|\tilde{r}_i - r_i| > \frac{\varepsilon}{2m} r_i\right) < \frac{\delta}{m}$$

for all i. Hence, by the union bound, the probability that $|\tilde{r}_i - r_i| > r_i\varepsilon/(2m)$ holds for at least one i is at most δ. With probability $1 - \delta$ we have

$$1 - \frac{\varepsilon}{2m} \leq \frac{\tilde{r}_i}{r_i} \leq 1 + \frac{\varepsilon}{2m}$$

for all i Then

$$\left(1 - \frac{\varepsilon}{2m}\right)^m \leq \prod_{i=1}^{m} \frac{\tilde{r}_i}{r_i} \leq \left(1 + \frac{\varepsilon}{2m}\right)^m.$$

We define $f(\varepsilon) = (1 + \varepsilon/(2m))^m$ and use Taylor's formula to approximate

$$\begin{aligned} f(\varepsilon) &= f(0) + \varepsilon f'(0) + \frac{\varepsilon^2}{2} f''(z) \\ &= 1 + \frac{\varepsilon}{2} + \frac{\varepsilon^2}{2} \cdot \frac{m(m-1)}{(2m)^2} (1 + z/(2m))^{m-2} \\ &\leq 1 + \varepsilon \cdot \frac{1}{2} + \frac{\varepsilon^2}{8} \mathrm{e}^{z/2} \\ &< 1 + \varepsilon, \end{aligned}$$

where $0 \leq z \leq \varepsilon < 1$. By similarly estimating $1 - \varepsilon < (1 - \varepsilon/(2m))^m$, we get

$$1 - \varepsilon \leq \prod_{i=1}^{m} \frac{\tilde{r}_i}{r_i} \leq 1 + \varepsilon$$

with probability $1 - \delta$. $\square$

We now consider estimating an individual r_i. Since we have access only to an *almost* uniform sampler of $\Omega(G_{i-1})$, we cannot directly apply Theorem 3.1.

We use the following method.

> **Assumption:** A is an $\varepsilon/(6m)$ uniform sampler of $\Omega(G_{i-1})$.
>
> Repeat for $k = 1, \ldots, M$:
> Choose $Z_k \in \Omega(G_{i-1})$ using A.
> If $Z_k \in \Omega(G_i)$, then $X_k = 1$; else $X_k = 0$.
> Return $\tilde{r}_i = (1/M) \sum_{k=1}^{m} X_k$.

Lemma 3.4: For all $m \geq 1$ and $0 < \varepsilon, \delta \leq 1$, the algorithm above returns an $(\varepsilon/(2m), \delta/m)$ approximation $\tilde{r}_i$ for r_i, assuming

$$M \geq \frac{1296 m^2 \ln(2m/\delta)}{\varepsilon^2}.$$

Proof: We first show that r_i is not too small, which is the basis of the whole idea.

Let (u, v) be the edge that is included in G_i but not in G_{i-1}. If $I \in \Omega(G_{i-1}) - \Omega(G_i)$, then I includes both vertices u and v. If we define $f(I) = I - \{ v \}$, we have $f(I) \in \Omega(G_i)$.

Since f is a one-to-one mapping from $I \in \Omega(G_{i-1}) - \Omega(G_i)$ to $\Omega(G_i)$, we get $|\Omega(G_{i-1}) - \Omega(G_i)| \leq |\Omega(G_i)|$ and

$$r_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|} = \frac{|\Omega(G_i)|}{|\Omega(G_i)| + |\Omega(G_{i-1}) - \Omega(G_i)|} \geq \frac{|\Omega(G_i)|}{|\Omega(G_i)| + |\Omega(G_i)|} = \frac{1}{2}.$$

The rest is technicalities to show that the given sample size is sufficient for the desired approximation accuracy.

The first part is to show that $\mathbf{E}[\tilde{r}_i]$ is sufficiently close to r_i. By our assumptions about A, we have

$$\left|\mathbf{E}[X_k] - \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}\right| = \left|\Pr(X_k = 1) - \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}\right| \leq \frac{\varepsilon}{6m}$$

for all k. Hence,

$$|\mathbf{E}[\tilde{r}_i] - r_i| = \left|\frac{1}{M}\sum_{k=1}^{m}\mathbf{E}[X_k] - \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}\right| \leq \frac{\varepsilon}{6m}.$$

Since $r_i \geq 1/2$ and $\varepsilon \leq 1$, we get in particular

$$\mathbf{E}[\tilde{r}_i] \geq \frac{1}{2} - \frac{\varepsilon}{6m} \geq \frac{1}{3}.$$

By applying Theorem 3.1 [M&U Thm 10.1] we now see that $\tilde{r}_i$ is an $(\varepsilon/(12m), \delta/m)$ approximation for the expected value $\mathbf{E}[\tilde{r}_i]$, when

$$M \geq \frac{3\ln(2m/\delta)}{(\varepsilon/12m)^2 \cdot (1/3)} = \frac{1296m^2 \ln(2m/\delta)}{\varepsilon^2}.$$

Hence, with probability at least $1 - \delta$, for all i we have

$$1 - \frac{\varepsilon}{12m} \leq \frac{\tilde{r}_i}{\mathbf{E}[\tilde{r}_i]} \leq 1 + \frac{\varepsilon}{12m}.$$

On the other hand, we saw earlier that $|\mathbf{E}[\tilde{r}_i] - r_i| \leq \varepsilon/(6m)$, so

$$1 - \frac{\varepsilon}{6mr_i} \leq \frac{\mathbf{E}[\tilde{r}_i]}{r_i} \leq 1 + \frac{\varepsilon}{6mr_i}.$$

By taking into account $r_i \geq 1/2$, we obtain

$$1 - \frac{\varepsilon}{3m} \leq \frac{\mathbf{E}[\tilde{r}_i]}{r_i} \leq 1 + \frac{\varepsilon}{3m}.$$

Hence, with probability at least $1 - \delta/m$ we have

$$\left(1 - \frac{\varepsilon}{3m}\right)\left(1 - \frac{\varepsilon}{12m}\right) \leq \frac{\mathbf{E}[\tilde{r}_i]}{r_i} \cdot \frac{\tilde{r}_i}{\mathbf{E}[\tilde{r}_i]} \leq \left(1 + \frac{\varepsilon}{3m}\right)\left(1 + \frac{\varepsilon}{12m}\right),$$

which implies

$$1 - \frac{\varepsilon}{2m} \leq \frac{\tilde{r}_i}{r_i} \leq 1 + \frac{\varepsilon}{2m}.$$

□

Markov Chain Monte Carlo (MCMC) [M&U Section 10.4]

We saw that if we have a fully polynomial almost uniform sampler for independent sets of a graph G, we can use it to obtain a fully polynomial approximation scheme for the number of independent sets.

It is fairly easy to show that such a sampling method exists if the graph in question has degree at most 4 [M&U Section 11.6]. Since our approximation algorithm only requires sampling for the original graph and some of its subgraphs, if the original graph G has degree at most 4, we can approximate the number of its independent sets.

The only other parts in the proof that were specific to independent sets were calculating the initial value $|\Omega(G_0)|$ and deriving the lower bound $r_i \geq 1/2$. Hence, the method is quite general.

Markov Chain Monte Carlo (MCMC) is a method for obtaining an almost uniformly distributed independent sample $(X_1, X_2, \ldots)$ from Ω:

1. Construct a Markov chain $(Y_0, Y_1, Y_2, \ldots)$ with state space Ω and a uniform stationary distribution.
2. Choose $X_1 = Y_r$, $X_2 = Y_{2r}$, $X_3 = Y_{3r}$, $\ldots$, where r is large enough.

Here the sampling interval r needs to be sufficiently long for the distribution to be sufficiently close to the stationary distribution regardless of the initial state. This is usually hard to analyse. (Notice that for estimating the mean or some other statistical quantity, it may be better to use $X_r, X_{r+1}, X_{r+2}, \ldots$, even if this makes the individual sample points highly correlated.)

However, there are fairly standard methods for finding a Markov chain with a desired stationary distribution, in particular the Metropolis algorithm we shall study next.

Therefore, it is fairly common to use MCMC as a heuristic without proof of approximation quality. In practice this often works well.

This can also be generalized to non-uniform distributions.

Assume that Ω is finite. The first part of the construction is defining for each $x \in \Omega$ its neighborhood $N(x) \subseteq \Omega$. We assume that the neighborhood structure is symmetric ($x \in N(y) \Leftrightarrow y \in N(x)$), and $x \notin N(x)$.

For example, if Ω is the set of all independent sets in G, we might choose the neighborhood $N(I)$ of an independent set I to consist of all independent sets of the form $I \cup \{ v \}$ or $I - \{ v \}$ for $v \in V$.

Choose some upper bound $M \geq \max_{x \in \Omega} |N(x)|$ and set the following transition probabilities for the Markov chain:

$$P_{x,y} = \begin{cases} 1/M & \text{if } y \in N(x) \\ 0 & \text{if } y \neq x \text{ and } y \notin N(x) \\ 1 - |N(x)|/M & \text{if } y = x. \end{cases}$$

Theorem 3.5 [M&U Lemma 10.7]**:** If the Markov chain constructed above is irreducible and aperiodic, then its has the uniform distribution as its unique stationary distribution.

Proof: For all $x \neq y$ we have either $P_{x,y} = P_{y,x} = 0$ or $P_{x,y} = P_{y,x} = 1/M$, so the uniform distribution $\boldsymbol{\pi}$ satisfies

$$\pi_x P_{x,y} = \pi_y P_{y,x} \qquad \text{for all } x, y.$$

Hence, it is a stationary distribution.

The uniqueness follows from irreducibility and aperiodicity (*Randomized Algorithms I* [M&U Thm 7.7]) □

For example, let Ω be the set of independent sets in a graph $G = (V, E)$ and the neighborhoods $N(x)$ as above. The desired Markov chain, with $M = |V|$, can be implemented as follows.

1. Choose $X_0 = \emptyset$ (or any other independent set).
2. If $X_k = I$, then
   Choose $v \in V$ uniformly at random.
   If $v \in I$, then $X_{k+1} = I - \{ v \}$.
   Else if $I \cup \{ v \}$ is an independent set, then $X_{k+1} = I \cup \{ v \}$.
   Else $X_{k+1} = I$.

The chain is clearly irreducible. If $E \neq \emptyset$, then $P_{I,I} > 0$ for at least some I, so the chain is aperiodic.

The Metropolis algorithm

We generalize the previous idea to sampling from a non-uniform distribution. Suppose we are given for each state x some weight $b(x) > 0$, with the intention that the stationary distribution π should satisfy $\pi_x = b(x)/B$ for some constant B. We then of course have $B = \sum_x b(x)$, but we do not require that the explicit value of B is known or easily computable.

For example, in the case on independent sets we might choose $b(I) = \exp(c|I|)$ for some $c > 0$, which favors large sets.

We define the transition matrix as follows:

$$P_{x,y} = \begin{cases} \frac{1}{M} \min\left\{1, \frac{b(y)}{b(x)}\right\} & \text{if } y \in N(x) \\ 0 & \text{if } y \neq x \text{ and } y \notin N(x) \\ 1 - \sum_{z \neq x} P_{x,z} & \text{if } y = x. \end{cases}$$

The previous construction is obtained as the special case where $b(x)$ is the same for all x.

For example, for independent sets with weight function $b(I) = \exp(c\,|I|)$, $c > 0$, this can be implemented as

1. Choose $X_0 = \emptyset$.
2. If $X_k = I$, then
   Choose $v \in V$ from the uniform distribution.
   If $v \in I$, then $X_{k+1} = I - \{\, v \,\}$ with probability e^{-c} and
   $X_{k+1} = I$ with probability $1 - \mathrm{e}^{-c}$.
   Else if $I \cup \{\, v \,\}$ is independent, then $X_{k+1} = I \cup \{\, v \,\}$.
   Else $X_{k+1} = I$.

Theorem 3.6 [M&U Lemma 10.8]: If the Markov chain constructed above is irreducible and aperiodic, it has the unique stationary distribution $\boldsymbol{\pi}$ where $\pi_x = b(x)/B$ for some constant B.

Proof: Let $\pi_x = b(x)/B$, where $B = \sum_{x \in \Omega} b(x)$. Hence, $\boldsymbol{\pi}$ is a probability distribution and satisfies $b(x)/b(y) = \pi_x/\pi_y$ for all x, y.

For all $x \neq y$, one of the following holds:

1. $b(x) = b(y)$ and $P_{x,y} = 1/M = P_{y,x}$,
2. $b(x) > b(y)$ and $P_{x,y} = (1/M) \cdot (b(y)/b(x))$ and $P_{y,x} = 1/M$,
3. $b(x) < b(y)$ and $P_{x,y} = 1/M$ and $P_{y,x} = (1/M) \cdot (b(x)/b(y))$.

Hence, in all cases we have $b(x)P_{x,y} = b(y)P_{y,x}$, which implies

$$\pi_x P_{x,y} = \pi_y P_{y,x} \qquad \text{for all } x, y.$$

Therefore, $\boldsymbol{\pi}$ is a stationary distribution.

The uniqueness follows again from finite state space, irreducibility and aperiodicity. □

4. Coupling of Markov chains

Coupling is one method of analysing the speed with which a Markov chain converges towards the stationary distribution. The analysis of this convergence speed, called mixing time, in realistic applications is beyond the scope of this course. We give examples of the coupling technique to give an idea of how such results might be accomplished.

(The 1996 Gödel prize was given to Mark Jerrum and Alistair Sinclair for their work on rapidly mixing Markov chains. Their method was based on analysing the conductance of the Markov chains, which is a different technique.)

We consider ergodic, irreducible finite-state discrete-time Markov chains. Hence, a unique stationary distribution exists.

For two probability measures D_1 and D_2 over a countable sample space S, their variation distance is

$$\|D_1 - D_2\| = \frac{1}{2} \sum_{x \in S} |D_1(\{x\}) - D_2(\{x\})| .$$

The variation distance has a useful alternative formulation:

Lemma 4.1 [M&U Lemma 11.1]**:** If D_1 and D_2 are probability measures over S, we have

$$\|D_1 - D_2\| = \max_{A \subseteq S} |D_1(A) - D_2(A)| .$$

Proof: Let $S^+ = \{\, x \in S \mid D_1(\{\, x \,\}) \geq D_2(\{\, x \,\}) \,\}$ and $S^- = \{\, x \in S \mid D_2(\{\, x \,\}) > D_1(\{\, x \,\}) \,\}$. Clearly

$$\begin{aligned} \max_{A \subseteq S} (D_1(A) - D_2(A)) &= D_1(S^+) - D_2(S^+) \\ \max_{A \subseteq S} (D_2(A) - D_1(A)) &= D_2(S^-) - D_1(S^-). \end{aligned}$$

Furthermore,

$$D_1(S^+) + D_1(S^-) = D_1(S) = 1 = D_2(S) = D_2(S^+) + D_2(S^-),$$

so

$$D_1(S^+) - D_2(S^+) = D_2(S^-) - D_1(S^-).$$

Therefore,

$$\begin{aligned} \max_{A \subseteq S} |D_1(A) - D_2(A)| &= \max\left\{\, D_1(S^+) - D_2(S^+), D_2(S^-) - D_1(S^-) \,\right\} \\ &= \frac{1}{2}\left((D_1(S^+) - D_2(S^+)) + (D_2(S^-) - D_1(S^-))\right) \\ &= \frac{1}{2}\sum_{x \in S} |D_1(\{\, x \,\}) - D_2(\{\, x \,\})| \,. \end{aligned}$$

□

Let A be an algorithm that produces random samples of a space Ω according to a distribution D_A, and let U be the uniform distribution over Ω. The previous result shows that A gives ε uniform samples if and only if $\|D_A - U\| \leq \varepsilon$.

Let $\bar{\boldsymbol{\pi}}$ be the stationary distribution of chain (X_t) with state space S, and let $\boldsymbol{p}_x^t$ be the distribution of X_t under condition $X_0 = x$. We define

$$\Delta_x(t) = \left\|\bar{\boldsymbol{\pi}} - \boldsymbol{p}_x^t\right\| \qquad \text{and} \qquad \Delta(t) = \max_{x \in S} \Delta_x(t).$$

Furthermore,

$$\tau_x(\varepsilon) = \min\{\, t \mid \Delta_x(t) \leq \varepsilon \,\} \qquad \text{ja} \qquad \tau(\varepsilon) = \max_{x \in S} \tau_x(\varepsilon).$$

The function $\tau(\varepsilon)$ is called the mixing time of the chain. If the mixing time is polynomial in $\log(1/\varepsilon)$ and the problem size, the chain is rapidly mixing.

Consider a Markov chain (M_t) with state space S. A coupling of this Markov chain is a Markov chain $(Z_t) = ((X_t, Y_t))$ with state space $S \times S$, such that

$$\begin{aligned} \Pr(X_{t+1} = x' \mid Z_t = (x, y)) &= \Pr(M_{t+1} = x' \mid M_t = x) \\ \Pr(Y_{t+1} = y' \mid Z_t = (x, y)) &= \Pr(M_{t+1} = y' \mid M_t = y). \end{aligned}$$

Thus (X_t) and (Y_t) are both copies of the original chain:

$$\Pr(X_t = r \mid X_0 = s) = \Pr(Y_t = r \mid Y_0 = s) = \Pr(M_t = r \mid M_0 = s).$$

Trivial examples of a coupling would be two independent copies of the original chain, or two identical copies. More useful examples are obtained by having non-trivial dependencies between (X_t) and (Y_t).

When (X_t) and (Y_t) have entered the same state, we say they have coupled. We can include a dependence that keeps the chains in the same state after they are coupled.

Example 4.2: Consider a Markov chain that represents shuffling a deck of n cards by picking at each step a random card from the deck and moving it to the top. The states are the $n!$ permutations of the deck, and each state has n equally probable successors (one of which is itself).

We create a coupling $((X_t, Y_t))$ where the initial distributions X_0 and Y_0 may be arbitrary. The transition from state (X_t, Y_t) is determined as follows.

1. In the permutation represented by X_t, pick a random position. Let the C be the card in that position. Obtain X_{t+1} by moving card C to the top.

2. Obtain Y_{t+1} by finding the card C in permutation Y_t and moving it to the top.

Clearly X_t and Y_t are copies of the same chain, and if $X_T = Y_T$ then $X_t = Y_t$ for all $t \geq T$. □

Lemma 4.3 [M&U Lemma]: Let $((X_t, Y_t))$ be a coupling of a Markov chain with state space S and T such that

$$\Pr(X_T \neq Y_T \mid X_0 = x, Y_0 = y) \leq \varepsilon$$

for all $x, y \in S$. Then

$$\tau(\varepsilon) \leq T.$$

Proof: Let $\boldsymbol{\pi}$ be the stationary distribution of the original chain. The choice of initial distributions X_0 and Y_0 does not affect the assumption or claim of the lemma. Hence, we consider (Y_t) that has $\boldsymbol{\pi}$ as the initial distribution. Therefore, also Y_t has distribution $\boldsymbol{\pi}$, regardless of how (X_t) is chosen.

By the assumptions, $\Pr(X_T \neq Y_T) \leq \varepsilon$ for all initial distributions (X_0, Y_0), so for all $A \subseteq S$ we have

$$\begin{aligned} \Pr(X_T \in A) &\geq \Pr((X_T = Y_T) \cap (Y_T \in A)) \\ &\geq 1 - \Pr(Y_T \notin A) - \Pr(X_T \neq Y_T) \\ &\geq \Pr(Y_T \in A) - \varepsilon. \end{aligned}$$

Similarly, $\Pr(X_T \notin A) \geq \Pr(Y_T \notin A) - \varepsilon$. Since Y_T follows the stationary distribution, the variation distance between X_T and the stationary distribution is at most ε. □

Example continued: Consider the mixing time for shuffling a deck of cards, using the previously introduced coupling $((X_t, Y_t))$.

If each card C has been selected at least once, the decks represented by X_t and Y_t are in the same order. The problem is thus reduced to coupon collecting.

After $n \ln n + cn$ steps, the probability that a given card C has never been selected is

$$\left(1 - \frac{1}{n}\right)^{n \ln n + cn} \leq \mathrm{e}^{-(\ln n + c)} = \frac{\mathrm{e}^{-c}}{n}.$$

Hence, after $n \ln n + n \ln(1/\varepsilon)$ steps, the probability that the decks are not in the same order is at most

$$n \cdot \frac{\mathrm{e}^{-\ln(1/\varepsilon)}}{n} = \varepsilon.$$

Hence, $\tau(\varepsilon) \leq n \ln n + n \ln(1/\varepsilon)$ and the chain is rapidly mixing. $\square$

Example 4.4: Random walk in a hypercube

Consider the familiar graph where the vertex set is $V = \{0,1\}^n$ and two vertices have an edge between them if they differ by exactly one bit.

We create a Markov chain by making transitions in the hypercube so that we first pick randomly one of the bit positions $1,\ldots,n$ and then with probability $1/2$ flip that bit. Hence, with probability $1/2$ we stay in the same state.

Alternatively, one can think that in a state $\boldsymbol{x} = (x_1,\ldots,x_n)$ we perform one random operation chosen from the $2n$ operations $x_i := b$, $i = 1,\ldots,n$, $b \in \{0,1\}$. Again, half the operations actually leave the state unchanged.

The chain is ergodic, since the self-loops prevent periodicity. Hence, a unique stationary distribution exists.

Construct a coupling $((X_t, Y_t))$ such that the same operation is always performed in chain (X_t) and chain (Y_t). Therefore, after each bit position has been operated at least once, the chains are in the same state.

Again, we have an instance of coupon collecting. As in the previous example, after $O(n \ln(n/\varepsilon))$ step the probability that the chains have coupled is at least $1 - \varepsilon$. The mixing time of the original chain is therefore $O(n \ln(n/\varepsilon))$. □

Fixed-size independent sets

We construct a Markov chain where the states are those independent sets in a graph $G = (V, E)$ that have exactly k vertices.

When $X \subset V$ is an indepenent set with $|X| = k$ we define $m(v, w, X)$, for any $v \in X$ and $w \in V$, to be the independent set with k vertices as follows:

- if $w \notin X$ and $X \cup \{ w \} - \{ v \}$ is an independent set, then $m(v, w, X) = X \cup \{ w \} - \{ v \}$,
- else $m(v, w, X) = X$.

We construct a Markov chain (X_t) by setting $X_{t+1} = m(v, w, X_t)$, where $v \in X_t$ and $w \in V$ are chosen from the uniform distributions. The chain is ergodic, with uniform stationary distribution (left as exercise).

We create a coupling $Z_t = (X_t, Y_t)$ as follows:

- Given X_t and Y_t, pick some bijection $M : X_t - Y_t \to Y_t - X_t$.
- Choose random $v \in X_t$ and $w \in V$ and set $X_{t+1} = m(v, w, X_t)$.
- If $v \in Y_t$, then $v' = v$; else $v' = M(v)$. Let $Y_{t+1} = m(v', w, Y_t)$.

Since each pair (v, w) and (v', w) has the same probability $1/(kn)$ of being chosen, (X_t) and (Y_t) are copies of the original chain. Furthermore, if $X_t = Y_t$, then $X_{t+1} = Y_{t+1}$.

How soon do we get $X_t = Y_t$? Let's look into the size of the set difference

$$d_t = |X_t - Y_t| = k - |X_t \cap Y_t| \,.$$

If $d_t = 0$, then $d_{t+1} = 0$.

Suppose now the degree of the graph is at most a constant Δ, and $k \leq n/(3\Delta + 3)$. We first show that for some value $0 < c < 1$, which depends on n, k, and Δ, we have

$$\mathbf{E}[d_{t+1} \mid d_t] \leq (1 - c)d_t.$$

By considering different alternatives we see that in any case $d_{t+1} \in \{\, d_t - 1, d_t, d_t + 1 \,\}$. If $d_t = 0$, then $d_{t+1} = 0$. Consider the probabilities of events $d_{t+1} = d_t + 1$ and $d_{t+1} = d_t - 1$ assuming $d_t > 0$. Write $X_{t+1} = m(v, w, X_t)$ and $Y_{t+1} = m(v', w, Y_t)$ as above.

If $d_{t+1} = d_t + 1$, then $|X_{t+1} \cap Y_{t+1}| = |X_t \cap Y_t| - 1$. This can only happen if $v = v' \in X_t \cap Y_t$, but $v \notin X_{t+1} \cap Y_{t+1}$ and $w \notin X_{t+1} \cap Y_{t+1}$.

Therefore, exactly one of the conditions $m(v, w, X_t) = X_t$ and $m(v', w, Y_t) = Y_t$ holds. The vertex w or some of its neighbors is in $(X_t - Y_t) \cup (Y_t - X_t)$.

The probability of this event is

$$\Pr(d_{t+1} = d_t + 1 \mid d_t > 0) \leq \frac{k - d_t}{k} \cdot \frac{2d_t(\Delta + 1)}{n}.$$

On the other hand, $d_{t+1} = d_t - 1$ holds at least if $v \notin Y_t$, and neither w nor any of its neighbors is in $(X_t \cup Y_t) - \{ v, v' \}$. Since $|X_t \cup Y_t| = k + d_t$, we get

$$\Pr(d_{t+1} = d_t - 1 \mid d_t > 0) \geq \frac{d_t}{k} \cdot \frac{n - (k + d_t - 2)(\Delta + 1)}{n}.$$

Therefore, for all $m \geq 1$ we have

$$\begin{aligned}
\mathbf{E}[d_{t+1} \mid d_t = m] &= m + \Pr(d_{t+1} = m+1 \mid d_t = m) - \Pr(d_{t+1} = m-1 \mid d_t = m) \\
&\leq m + \frac{k-m}{k} \cdot \frac{2m(\Delta+1)}{n} - \frac{m}{k} \cdot \frac{n-(k+m-2)(\Delta+1)}{n} \\
&= m\left(1 - \frac{n-(3k-m-2)(\Delta+1)}{kn}\right) \\
&\leq (1-c)m,
\end{aligned}$$

where

$$c = \frac{n-(3k-3)(\Delta+1)}{kn}.$$

Furthermore, $\mathbf{E}[d_{t+1} \mid d_t = 0] = 0$. Hence,

$$\mathbf{E}[d_{t+1}] = \mathbf{E}[\mathbf{E}[d_{t+1} \mid d_t]] \leq (1-c)\mathbf{E}[d_t].$$

By induction, we get

$$\mathbf{E}[d_t] \leq \mathbf{E}[d_0](1-c)^t \leq \mathbf{E}[d_0]\mathrm{e}^{-ct}.$$

Since $\mathbf{E}[d_0] \leq k$ and d_t only gets non-negative integer values,

$$\Pr(X_t \neq Y_t) = \Pr(d_t > 0) = \Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq \mathbf{E}[d_0]\mathrm{e}^{-ct}.$$

Therefore, $\Pr(X_t \neq Y_t) \leq \varepsilon$ for

$$t \geq \frac{\ln(k/\varepsilon)}{c}.$$

The mixing time is $\tau(\varepsilon) \leq (\ln(k/\varepsilon)/c)$, so the chain is rapidly mixing.

Monotonicity of variation distance [M&U Section 11.3]

Recall that we use $\boldsymbol{p}_x^t$ to denote the distribution at time t if initial state is x, and

$$\Delta(t) = \max_{x \in S} \left\| \bar{\boldsymbol{\pi}} - \boldsymbol{p}_x^t \right\|.$$

We show that the convergence of a Markov chain towards its stationary distribution is monotone in the sense that

$$\Delta(t+1) \leq \Delta(t).$$

We will need the following lemma.

Lemma 4.5 [M&U Lemma 11.3]: Let $Z = (X, Y)$ be a random variable with range $S \times S$, where S is finite, and X and Y have marginal distributions σ_X ja σ_Y. Then

$$\Pr(X \neq Y) \geq \|\sigma_X - \sigma_Y\|.$$

Furthermore, for any given marginal distributions σ_X and σ_Y there is a joint distribution for which this holds as equality.

Proof: Since $\Pr(X = Y = s) \leq \min\{\Pr(X = s), \Pr(Y = s)\}$, we have

$$\begin{aligned}
\Pr(X \neq Y) &= \sum_{s \in S} (\Pr(X = s) - \Pr(X = Y = s)) \\
&\geq \sum_{s \in S} (\Pr(X = s) - \min\{\Pr(X = s), \Pr(Y = s)\}) \\
&= \sum_{s \in S} (\sigma_X(s) - \min\{\sigma_X(s), \sigma_Y(s)\}) \\
&= \sum_{s \in S^+} (\sigma_X(s) - \sigma_Y(s)),
\end{aligned}$$

where $S^+ = \{ s \in S \mid \sigma_X(s) \geq \sigma_Y(s) \}$ as in the proof of Lemma 4.1 [M&U Lemma 11.1] The same argument as in the proof of Lemma 4.1 shows that

$$\sum_{s \in S^+} (\sigma_X(s) - \sigma_Y(s)) = \|\sigma_X - \sigma_Y\|.$$

Assume now that σ_X and σ_Y are given. We want a joint distribution with

$$\Pr(X \neq Y) = \|\sigma_X - \sigma_Y\| .$$

Based on the above, this is equivalent with having

$$\Pr(X = Y = s) = \min\{\sigma_X(s), \sigma_Y(s)\} .$$

for all $s \in S$. Define $m(s) = \min\{\sigma_X(s), \sigma_Y(s)\}$. We claim that the desired joint distribution is obtained as

$$\Pr(X = x, Y = y) = \begin{cases} m(x) & \text{if } x = y \\ \frac{(\sigma_X(x) - m(x))(\sigma_Y(y) - m(y))}{1 - \sum_z m(z)} & \text{if } x \neq y. \end{cases}$$

The basic idea is that of the probability mass $\sigma_X(x)$ related to the event $(X = x)$, we reserve $m(x)$ for the event $(X = Y = x)$. The remaining part $\sigma_X(x) - m(x)$ is shared among events $(X = x, Y = y)$ respecting the desired marginal distribution for Y.

The distribution Pr has the desired margins. If $m(x) = \sigma_X(x)$, then $\Pr(X = x, Y \neq x) = 0$, and

$$\Pr(X = x) = \sum_{y \in S} \Pr(X = x, Y = y) = \Pr(X = Y = x) = m(x).$$

On the other hand, if $m(x) = \sigma_Y(x)$, then

$$\begin{aligned}
\Pr(X = x) &= \sum_{y \in S} \Pr(X = x, Y = y) \\
&= m(x) + \sum_{y \neq x} \frac{(\sigma_X(x) - m(x))(\sigma_Y(y) - m(y))}{1 - \sum_z m(z)} \\
&= m(x) + \frac{(\sigma_X(x) - m(x)) \sum_{y \neq x} (\sigma_Y(y) - m(y))}{1 - \sum_z m(z)} \\
&= m(x) + \frac{(\sigma_X(x) - m(x))(1 - \sigma_Y(x) - (\sum_y m(y) - m(x)))}{1 - \sum_z m(z)} \\
&= \sigma_X(x).
\end{aligned}$$

□

Theorem 4.6 [M&U Thm 11.4]: Any ergodic Markov chain satisfies $\Delta(T+1) \leq \Delta(T)$ for all T.

Proof: Consider an arbitrary state x. Let p_x^t be the distribution of the chain at time t when the initial state is x. Let p_*^t be the distribution of the chain at time t when the initial state follows the stationary distribution. Hence, p_*^t is still the stationary distribution, and

$$\Delta_x(t) = \left\| p_x^t - p_*^t \right\|.$$

Let (X_t) and (Y_t) be Markov chains with distributions p_x^t and p_*^t.

Consider some fixed time T. By Lemma 4.5, we can construct for $Z_T = (X_T, Y_T)$ a joint distribution such that $\Pr(X_T \neq Y_T) = \left\|p_x^T - p_*^T\right\|$ (and the margin distributions remain p_x^T and p_*^T). Additionally, we can define $Z_{T+1} = (X_{T+1}, Y_{T+1})$ by making a coupling such that if $X_T = Y_T$, then $X_{T+1} = Y_{T+1}$.

Hence, by the construction of the distributions of (X_T, Y_T) and (X_{T_1}, Y_{T+1}), we have

$$\Delta_x(T) = \left\|p_x^T - p_*^T\right\| = \Pr(X_T \neq Y_T) \geq \Pr(X_{T+1} \neq Y_{T+1}).$$

By Lemma 4.5, in any case we have

$$\Pr(X_{T+1} \neq Y_{T+1}) \geq \left\|p_x^{T+1} - p_*^{T+1}\right\| = \Delta_x(T+1).$$

Since x was arbitrary, the claim follows. $\square$

Geometric convergence [M&U Section 11.4]

Theorem 4.7 [M&U Thm 11.5]: Consider a finite-state irreducible aperiodic Markov chain with transition matrix $\mathbf{P}$ and stationary distribution $\boldsymbol{\pi}$. Let $m_j = \min_i p_{ij}$ for all j, and $m = \sum_j m_j$. For all x and T we have

$$\left\| p_x^T - \boldsymbol{\pi} \right\| \leq (1-m)^T.$$

Proof: We create two copies of the chain, (X_t) and (Y_t). No matter what the values of X_{t-1} and Y_{t-1} are, the assumptions imply

$$\Pr(X_t = Y_t = j) \geq m_j$$

for all j. Hence $\Pr(X_t \neq Y_t) \leq 1 - m$, and the probability of making at least T without coupling is at most $(1-m)^T$. □

In practice we often have $m_j = 0$ for all j, in which case the previous result is not directly useful. However, under the assumptions, for all i, j there is t_{ij} such that $P_{ij}^{t_{ij}} > 0$. If additionally $P_{ii} > 0$, then $P_{ij}^{t} > 0$ for all $t \geq t_{ij}$. Then $P_{ij}^{T} > 0$ for all i, where $T = \max_i t_{ij}$. Hence, we can apply the result to the chain with transition matrix $\boldsymbol{P}^T$. This chain has the same stationary distribution as the original one.

Instead of assuming $P_{ii} > 0$, it is sufficient to make the weaker assumption that state i is aperiodic.

We have an even more general result that shows geometric convergence.

Theorem 4.8 [M&U Thm 11.6]: Consider a finite-state irreducible aperiodic Markov chain (M_t) for which $\tau(c) \leq T$ for some $c < 1/2$. Then for all ε we have

$$\tau(\varepsilon) \leq \left\lceil \frac{\ln \varepsilon}{\ln(2c)} T \right\rceil .$$

Proof: Let the transition matrix of the chain be $\boldsymbol{P}$, and stationary distribution $\boldsymbol{\pi}$. Fix states x and y. By assumption, $\left\| p_x^T - \boldsymbol{\pi} \right\| \leq c$ and $\left\| p_y^T - \boldsymbol{\pi} \right\| \leq c$, so $\left\| p_x^T - p_y^T \right\| \leq 2c$. By Lemma 4.5 [M&U Lemma 11.3], we can construct a random variable $Z_{T,x,y} = (X_T, Y_T)$ such that X_T and Y_T have distributions p_x^T and p_y^T, and $\Pr(X_T \neq Y_T) \leq 2c$.

Let now (M'_t) be a Markov chain with transition matrix $\boldsymbol{P}^T$. We make a coupling for this chain using the same coupling that gave us $Z_{T,x,y}$. More specifically, if the coupling at time t is in state (x', y'), then the distribution at time $t+1$ is the same as for $Z_{T,x',y'}$.

Then the probability that (M'_t) makes k transitions without coupling is at most $(2c)^k$. Hence, after k transitions (M'_t) has variation distance at most ε from the stationary distribution if

$$(2c)^k \leq \varepsilon,$$

which means

$$k \geq \left\lceil \frac{\ln \varepsilon}{\ln(2c)} \right\rceil.$$

The claim follows since T transitions in (M_t) correspond to one transition in (M'_t) and the chains have identical stationary distributions. □

Sampling graph colorings [M&U Section 11.5]

A (vertex) coloring of a graph $G = (V, E)$ with c colors is a mapping $h\colon V \to \{1, \ldots, c\}$. The coloring is proper if $h(u) \neq h(v)$ whenever $(u, v) \in E$.

If the graph has degree Δ, it can easily be colored properly using $\Delta + 1$ colors by just considering the vertices one after another and always picking a color that has not been used in the neighborhood of the vertex.

The chromatic number $\chi(G)$ of graph G is the smallest number of colors in a proper coloring. Determining the chromatic number is a well-known NP-hard problem. Here we consider sampling from proper colorings where the number of colors is well above the chromatic number.

We define the Markov chain in a straightforward manner. To make a transition, we choose a vertex $v \in V$ and a color $\ell \in \{1, \ldots, c\}$ from the uniform distribution. If changing the color of v into ℓ results in a proper coloring, we make the change. Otherwise the coloring remains unchanged.

The chain is clearly aperiodic. It is also irreducible assuming $c \geq \Delta + 2$. To move from coloring X to coloring Y, we fix the colors of vertices in some arbitrary order. If changing the color of a vertex v to match Y would lead to an improper coloring, this must be because of some later vertex v'. Since $c \geq \Delta + 2$, we can fix the situation by changing the color of v' to a different one that is not used in any of its neighbors.

We first give a simple coupling that show rapid mixing when $c \geq 4\Delta + 1$. We then improve the construction to get down to $c \geq 2\Delta + 1$.

Theorem 4.9: If a graph has n vertices and maximum degree Δ, then for $c \geq 4\Delta + 1$ the mixing time of the Markov chain of its proper c-colorings is

$$\tau(\varepsilon) \leq \left\lceil \frac{nc}{c - 4\Delta} \ln \frac{n}{\varepsilon} \right\rceil .$$

Proof: We construct a coupling $((X_t, Y_t))$ such that at each transition, both chains choose the same pair (v, ℓ). Let D_t be the set of vertices that at time t have different color in chains (X_t) ja (Y_t), and $d_t = |D_t|$. We'll show that if $d_t > 0$, then d_t is more likely to decrease than increase. The proof strategy is similar to the one we used for fixed-size indepent sets.

Assume that $d_t > 0$, and consider first the case $v \in D_t$, which has probability d_t/n. If further $v \in D_{t+1}$, then the chosen color ℓ has appeared in a neighbor of v in at least one of the colorings X_t and Y_t. There are at most 2Δ such colors. Hence, the probability that D_t decreases is

$$\Pr(d_{t+1} = d_t - 1 \mid d_t > 0) \geq \frac{d_t}{n} \cdot \frac{c - 2\Delta}{c}.$$

On the other hand, suppose $v \notin D_t$. If $v \in D_{t+1}$, then the chosen color ℓ is such that some u gets color ℓ in exactly one of the colorings X_t and Y_t. Each $u \in D_t$ can in this manner affect at most Δ vertices v and two colors ℓ. Hence,

$$\Pr(d_{t+1} = d_t + 1 \mid d_t > 0) \leq d_t \cdot \frac{\Delta}{n} \cdot \frac{2}{c}.$$

If $d_t = 0$, then $d_{t+1} = 0$, so these estimates hold (as equalities) also if we change the condition to be $d_t = 0$.

Therefore,

$$\begin{aligned}
\mathbf{E}[d_{t+1} \mid d_t] &= d_t + \Pr(d_{t+1} = d_t + 1) - \Pr(d_{t+1} = d_t - 1) \\
&\leq d_t + d_t \cdot \frac{\Delta}{n} \cdot \frac{2}{c} - \frac{d_t}{n} \cdot \frac{c - 2\Delta}{c} \\
&= d_t \left(1 - \frac{c - 4\Delta}{nc}\right)
\end{aligned}$$

and

$$\mathbf{E}[d_{t+1}] = \mathbf{E}[\mathbf{E}[d_{t+1} \mid d_t]] \leq \mathbf{E}[d_t] \left(1 - \frac{c - 4\Delta}{nc}\right).$$

By induction, we get

$$\mathbf{E}[d_t] \leq d_0 \left(1 - \frac{c - 4\Delta}{nc}\right)^t.$$

Since $d_0 \leq n$ and d_t is a non-negative integer,

$$\Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq n\left(1 - \frac{c - 4\Delta}{nc}\right)^t \leq n \exp\left(-\frac{t(c-4\Delta)}{nc}\right).$$

Hence, $\Pr(d_t = 0) \geq 1 - \varepsilon$ for

$$t \geq \frac{nc}{c - 4\Delta} \ln \frac{n}{\varepsilon}.$$

□

The preceding analysis can be made sharper by considering how many of the neighbors of v are in D_t. For this to be useful, we also need to change the coupling a bit. This results in a slightly smaller requirement for the number of colors.

Theorem 4.10 [M&U Thm 11.8]**:** If a graph has n vertices and degree at most Δ, then the mixing time for its proper c-colorings is

$$\tau(\varepsilon) \leq \left\lceil \frac{n(c-\Delta)}{c-2\Delta} \ln \frac{n}{\varepsilon} \right\rceil$$

assuming $c \geq 2\Delta + 1$.

Proof: Again we make a coupling $((X_t, Y_t))$. We choose a random $v \in V$ as earlier. If v has different color in X_t and Y_t, we choose a random $\ell \in \{1, \ldots, c\}$ and try in both chains to switch the color of v to ℓ as previously. The case when v_t has the same color in X_t and Y_t is more delicate and will be handled later.

Again, let D_t be the set of vertices that have different color in X_t and Y_t, and let $d_t = |D_t|$. Additionally, we define $A_t = V - D_t$.

Denote the set of neighbors of v by $N(v)$. We define

$$d'(v) = \begin{cases} |N(v) \cap D_t| & \text{if } v \in A_t \\ |N(v) \cap A_t| & \text{if } v \in D_t. \end{cases}$$

Then

$$\sum_{v \in D_t} d'(v) = \sum_{w \in A_t} d'(v) =: m'.$$

If $v \in D_t$, then there are at least $c - (2\Delta - d'(v))$ colors that can properly be given to v both in X_t and Y_t. Therefore,

$$\Pr(d_{t+1} = d_t - 1 \mid d_t > 0) \geq \sum_{v \in D_t} \frac{1}{n} \cdot \frac{c - 2\Delta + d'(v)}{c} = \frac{(c - 2\Delta)d_t + m'}{cn}.$$

We now move to the case $v \in A_t$. Consider first the situation where exactly one of the neighbors of v have different color in X_t and Y_t. Denote these colors by ℓ_x and ℓ_y. We make the coupling so that if in (X_t) we chose (v, ℓ_x), then in (Y_t) we choose (v, ℓ_y), and vice versa. Now only one of these choices will give v different color in X_{t+1} and Y_{t+1}, when in our original coupling both choices did this.

More generally, let S_1 be the set of colors that appear among the neighbors of v in X_t but not in Y_t. Similarly, let S_2 be the set of colors that appear among the neighbors of v in Y_t but not in X_t.

We make as many pairs $(\ell_1, \ell_2) \in S_1 \times S_2$ as possible and make the coupling so that if in X_{t+1} we attempt to assign color ℓ_1 to v, then in Y_{t+1} we choose ℓ_2, and vice versa. Now, since we assumed $v \in A_t$, exactly one of the colors $\{\ell_1, \ell_2\}$ is such that choosing it in (X_t) causes v to get different color in X_{t+1} and Y_{t+1}. Hence,

$$\Pr(d_{t+1} = d_t + 1 \mid d_t > 0) \leq \sum_{v \in A_t} \frac{1}{n} \cdot \frac{d'(v)}{c} = \frac{m'}{cn}.$$

As previously, we now notice that

$$\mathbf{E}[d_{t+1} \mid d_t] \leq d_t + \frac{m'}{cn} - \frac{(c-2\Delta)d_t + m'}{cn} = d_t\left(1 - \frac{c-2\Delta}{cn}\right)$$

and

$$\Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq n\left(1 - \frac{c-2\Delta}{cn}\right)^t \leq n\exp\left(-\frac{t(c-2\Delta)}{cn}\right).$$

Hence, $\Pr(d_t = 0) \geq 1 - \varepsilon$ for

$$t \geq \frac{cn}{c-2\Delta}\ln\frac{n}{\varepsilon}.$$

□

5. Martingales

Martingales are a useful class of random processes that allow us to generalize many results from independent random variables to certain types of dependencies.

We shall only introduce the basic concepts. As an example application we introduce the generalization of Chernoff bounds for large deviations of $\sum_i X_i$ where the random variables X_i don't need to be independent (but still must satisfy some conditions).

(The term "martingale" in this context originally comes from a betting strategy in which the bet is doubled after each loss.)

Let $(Z_0, Z_1, Z_2, \ldots)$ and $(X_0, X_1, X_2, \ldots)$ be finite or countably infinite sequences of random variables.

We call (Z_t) a martingale with respect to (X_t) if for all n

1. the values of $X_0, \ldots, X_n$ determine the value of Z_n
2. $\mathbf{E}[|Z_n|] < \infty$ and
3. $\mathbf{E}[Z_{n+1} \mid X_0, \ldots, X_n] = Z_n$.

The sequence (Z_n) is a martingale if it is a martingale with respect to itself.

Example 5.1: Let X_i be results of independent throws of a 6-sided die. Before throw i, the player places a bet $R_i \in [0, c]$. If X_i is odd, the player wins R_i units, otherwise he loses R_i units of money. Let $Y_i \in [-c, c]$ be the amount the player wins in round i, and $Z_t = \sum_{i=1}^{t} Y_i$.

If R_i is constant, the random variables Y_i are independent and we can apply the familiar Chernoff bounds to the total profit Z_t.

If R_i may depend on the previous gains $Z_0, \ldots, Z_{i-1}$, then (Z_t) is a martingale.

If R_i may depend on the results of the earlier throws $X_0, \ldots, X_{i-1}$, then (Z_t) is a martingale with respect to (X_t). □

Let $X_0, \ldots, X_n$ be a sequence of random varibles and Y such that $\mathbf{E}[|Y|] < \infty$ and the values $X_0, \ldots, X_n$ determine Y. Define

$$Z_i = \mathbf{E}[Y \mid X_0, \ldots, X_i].$$

Hence, $(Z_0, Z_1, \ldots, Z_n)$ gives a sequence of increasingly well-informed estimates of Y. In particular $Z_n = Y$, and if X_0 is some constant "fake variable" then $Z_0 = \mathbf{E}[Y]$ is constant.

This is a martingale with respect to (X_t):

$$\begin{aligned} \mathbf{E}[Z_{i+1} \mid X_0, \ldots, X_i] &= \mathbf{E}[\mathbf{E}[Y \mid X_0, \ldots, X_{i+1}] \mid X_0, \ldots, X_i] \\ &= \mathbf{E}[Y \mid X_0, \ldots, X_i] \\ &= Z_i, \end{aligned}$$

where we used the property

$$\mathbf{E}[\mathbf{E}[Y \mid U, V] \mid U] = \mathbf{E}[Y \mid U].$$

Martingales like this are called Doob martingales.

If a martingale represents a gambling scenario like above, the game is fair in the sense that the expected wealth of the player at any given time is the same as his initial wealth:

Lemma 5.2 [M&U Lemma 12.1]: If $Z_0, Z_1, \ldots$ is a martingale with respect to $X_0, X_1, \ldots$, then

$$\mathbf{E}[Z_t] = \mathbf{E}[Z_0]$$

for all t.

Proof: By known properties,

$$\mathbf{E}[Z_i] = \mathbf{E}[\mathbf{E}[Z_{i+1} \mid X_0, \ldots, X_i]] = \mathbf{E}[Z_{i+1}].$$

The claim follows by induction. □

As an example of applying martingales, we consider generalizing Chernoff bounds.

Theorem 5.3 (Azuma-Hoeffding, [M&U Thm 12.4]): Let $X_0, X_1, \ldots$ be a martingale such that

$$|X_k - X_{k-1}| \leq c_k.$$

Then for all t we have

$$\Pr(|X_t - X_0| \geq \lambda) \leq 2 \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^{t} c_k^2}\right).$$

We omit the proof. It is based on a similar estimation of generating functions as the Chernoff bounds for independent variables.

Corollary 5.4 [M&U Corollary 12.5]: Let $X_0, X_1, \ldots$ be a martingale such that

$$|X_k - X_{k-1}| \leq c.$$

Then for all t we have

$$\Pr(|X_t - X_0| \geq c\lambda\sqrt{t}) \leq 2 \exp\left(-\frac{\lambda^2}{2}\right).$$

Example 5.5: Let X be a random string of n symbols over an alphabet Σ where $|\Sigma| = s$. Hence, $X = X_1 \dots X_n$, where X_i are independent and uniformly distributed over Σ.

An occurence of a string $B = b_1 \dots b_k$ in X is an index $1 \leq i \leq n - k + 1$ such that $X_{i+j-1} = b_j$ for all j. Let $F_i = 1$ if i is an occurence of B, and $\sum_i F_i$ the number of occurences. Therefore,

$$\mathbf{E}[F] = \sum_{i=1}^{n-k+1} \Pr(X[i \dots i+k-1] = B) = \frac{n-k+1}{s^k}.$$

Define a Doob martingale with $Z_0 = \mathbf{E}[F]$ and

$$Z_{i+1} = \mathbf{E}[F \mid X_1, \ldots, X_{i+1}].$$

Since F_i can depend on X_{j+1} only if $i \leq j+1 \leq i+k-1$, for any given j there are at most k indices i such that the difference

$$\Delta_{i,j} = \mathbf{E}[F_i \mid X_1, \ldots, X_{j+1}] - \mathbf{E}[F_i \mid X_1, \ldots, X_j]$$

in non-zero. Since $F_i \in \{0, 1\}$, we have in any case

$$-1 \leq \Delta_{i,j} \leq 1.$$

Therefore,

$$|Z_{j+1} - Z_j| = \sum_{i=1}^{n-k+1} |\Delta_{i,j}| \leq k.$$

Corollary 5.4 now implies

$$\Pr(|F - \mathbf{E}[F]| \geq k\lambda\sqrt{n}) \leq 2\exp\left(-\frac{\lambda^2}{2}\right).$$

□

6. Summary

We have seen various ways of applying probabilities in designing and analysing algorithms:

- algorithms that use randomness
- average case analysis
- some other random environment (for example, queue systems).

Things we can analyze about algorithms:

- the expected cases: clearly more difficult than worst-case analysis, but linearity of expectation often makes this manageable
- variance: usually difficult to analyse
- large deviations: usually difficult, but in particular cases we may be able to apply powerful tools such as Chernoff bounds.

Ways to use randomness in algorithm desing:

- avoiding the worst case (as in Las Vegas algorithms)
- sampling: works for example for counting; helpful techniques include Chernoff bounds, MCMC and rapidly mixing Markov chains.
- fingerprinting: we use randomness to create a small fingerprint for a large object (for example, hashing).
- load balancing: packet routing in networks etc.
- symmetry breaking, in particular in distributed computing.

The End