582691 Randomized Algorithms I

Spring 2013, period III

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
- continues in *Randomized algorithms II* which can also be taken separately
- 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. 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.

Why randomization?

Randomness is an important tool in modeling natural and other fenomena.

In design and analysis of algorithms, sources of randomness include

- randomized algorithms: the computations even on the same input may vary depending on internal randomization of the algorithm ("coin tosses")
- the algorithm may act in a random environment (average case analysis, data communications, ...)

Theory of probability is a powerful general tool for these situations.

Randomizations may allow us to find an algorithm that compared to a deterministic one is

- faster or more memory efficient or
- easier to implement.

Basic techniques and situations include

- random sampling, Monte Carlo methods
- randomized search, simulated annealing
- fingerprinting technique.

In some situation randomization is compulsory to get any acceptable solution:

- hiding information from an adversary (cryptography, games)
- distributed systems: load balancing, leader election etc.

Randomization may make the algorithm more robust against unwanted situations.

- Example: randomized quicksort avoids having any particular worst-case input.

Typical questions

Usually a randomized algorithm has some non-zero probability of giving an incorrect result.

- if the result is yes/no: what's the probability of error?
- if the result is a numerical value: what's the probability of a large error?

Some randomized algorithms (sometimes called Las Vegas algorithms) always give the correct result, but the run time is random.

- what's the expected run time?
- what's the probability of the run time exceeding some boundary?

Contents of the course

Main topics for *Randomized algorithms I* include

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"

Randomized algorithms II continues with

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

1. Probability

Let Ω be any set and $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ a collection of subsets of Ω. (We use $\mathcal{P}(\Omega)$ to denote the power set of Ω.) A function $\Pr\colon \mathcal{F} \to \mathbb{R}$ is a probability function (or probability measure) [M&U Def. 1.2] if

1. $\Pr(E) \geq 0$ for all $E \in \mathcal{F}$,
2. $\Pr(\Omega) = 1$ and
3. if $E_1, E_2, E_3, \ldots$ is a sequence of pairwise disjoint sets (meaning $E_i \cap E_j = \emptyset$ when $i \neq j$) such that $E_i \in \mathcal{F}$ for all i, we have

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$$

(countable additivity).

For the conditions we just set for a probability function Pr to be interesting, its domain $\mathcal{F}$ must have some closure properties.

A collection of subsets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ algebra if

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F}$ implies $\overline{A} \in \mathcal{F}$, where $\overline{A} = \Omega - A$
3. if $A_1, A_2, A_3, \ldots$ is a sequence such that $A_i \in \mathcal{F}$ for all $i \in \{1, 2, 3, \ldots\}$, we have

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Remark No assumption is made about the union $\cup_{i \in I} A_i$ of a family $\{ A_i \mid i \in I \}$ if the index set I is not countable.

We now define a probability space as a triple $(\Omega, \mathcal{F}, \Pr)$ where

1. the sample space Ω is an arbitrary set
2. $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ algebra over Ω
3. $\Pr\colon \mathcal{F} \to \mathbb{R}$ is a probability function.

Subsets $E \subseteq \Omega$ of the sample space are called events. In particular, the sets $E \in \mathcal{F}$ are allowable events.

For any property ϕ of the elements of the sample space, we write simply $\Pr(\phi(x)) = \Pr(\{\, x \in \Omega \mid \phi(x) \,\})$. For example, $\Pr(g(x) = 3)$ denotes the probability $\Pr(\{\, x \in \Omega \mid g(x) = 3 \,\})$.

Example 1.1: For a finite sample space with $|\Omega| = n \in \mathbb{N}$, we define the symmetrical (or uniform) probability for Ω as $(\Omega, \mathcal{P}(\Omega), \Pr)$ where $\Pr(E) = |E|/n$ kaikilla $E \subseteq \Omega$.

More generally, if a probability space is of the form $(\Omega, \mathcal{P}(\Omega), \Pr)$ for a finite or countably infinite Ω, we say it's discrete. A discrete probability space can be specified by just giving all the probabilities $\Pr(\{\, x \,\})$ of the individual elements $x \in \Omega$. □

On this course we'll mainly deal with discrete spaces. Therefore we often don't mention assumptions of the type "if $E \in \mathcal{F}$" (which often would anyway be clear from the context).

Occasionally even in a finite or countable Ω it's useful to consider other σ algebras than just $\mathcal{P}(\Omega)$.

Example 1.2: Let $\Omega = \mathbb{R}$, and let $\mathcal{F}$ be the smallest σ algebra that includes all the closed intervals $[a, b]$, $a, b \in \mathbb{R}$. The elements of this σ algebra are called Borel sets.

We define the uniform probability over the interval $[0, 1]$ by setting for all $0 \leq a \leq b \leq 1$ the probability of the interval $[a, b]$ to be the same as its length: $\Pr([a, b]) = b - a$. The probabilities of other Borel sets follow by the properties of a probability function.

Remark We have $\Pr(\{ x \}) = 0$ for any individual $x \in \mathbb{R}$, so the probability of any countable set is zero, too. However, this implies nothing about uncountable sets. □

It might seem nice to pick $\mathcal{F} = \mathcal{P}(\mathbb{R})$, in other words make any set of real numbers an allowable event. However, it turns out to be impossible to define $\Pr(A)$ for all $A \subseteq \mathbb{R}$ so that all the requirements of a probability function would be satisfied. Luckily, in practice, we have little need to go beyond Borel sets.

Probability of the union

Straight from the definition, for any two allowable events we have

$$\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(E \cap F).$$

For any countable I and a sequence of allowable events $(E_i)_{i \in I}$ we have

$$\Pr\left(\bigcup_{i \in I} E_i\right) \leq \sum_{i \in I} \Pr(E_i)$$

which is called the union bound [M&U Lemma 1.2]. This is a very useful but sometimes quite loose inequality.

When $|I| = n \in \mathbb{N}$, the exact probability of the union can be obtained as

$$\Pr\left(\bigcup_{i \in I} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{J \subseteq I, |J|=k} \Pr\left(\bigcap_{j \in J} E_j\right)$$

which is known as the inclusion-exclusion principle [M&U Lemma 1.3].

By taking the sum only up to some limit $k < n$ we get alternating upper and lower bounds:

For odd ℓ we have

$$\Pr\left(\bigcup_{i\in I} E_i\right) \leq \sum_{k=1}^{\ell} (-1)^{k+1} \sum_{J\subseteq I, |J|=k} \Pr\left(\bigcap_{j\in J} E_j\right).$$

For even ℓ we have

$$\Pr\left(\bigcup_{i\in I} E_i\right) \geq \sum_{k=1}^{\ell} (-1)^{k+1} \sum_{J\subseteq I, |J|=k} \Pr\left(\bigcap_{j\in J} E_j\right)$$

(Bonferroni inequalities).

Independence

Two events E and F are independent [M&U Def. 1.3] if

$$\Pr(E \cap F) = \Pr(E)\Pr(F).$$

More generally, events $E_1, \ldots, E_k$ are mutually independent (or just independent) if for all $I \subseteq \{1, \ldots, k\}$ we have

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$

Events $E_1, \ldots, E_k$ are pairwise independent if for all $i \neq j$ the events E_i and E_j are independent.

Remark Pairwise independence does not in general imply mutual independence for more than two events.

If $\Pr(F) > 0$, we define the conditional probability of E given F as

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

Thus, if $\Pr(F) > 0$, then E and F are independent iff $\Pr(E \mid F) = \Pr(E)$.

Given two probability spaces $(\Omega_1, \mathcal{F}_1, \Pr_1)$ and $(\Omega_2, \mathcal{F}_2, \Pr_2)$, we define their product space as

$$(\Omega_1, \mathcal{F}_1, \Pr_1) \times (\Omega_2, \mathcal{F}_2, \Pr_2) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \Pr_1 \times \Pr_2)$$

where $\mathcal{F}_1 \times \mathcal{F}_2$ is the smallest σ algebra that includes all the sets $E_1 \times E_2$ where $E_i \in \mathcal{F}_i$, and $(\Pr_1 \times \Pr_2)(E_1 \times E_2) = \Pr_1(E_1) \times \Pr_2(E_2)$ for $E_i \in \mathcal{F}_i$. This means that if we draw a pair (x, y) from $\Pr_1 \times \Pr_2$, then the events $x \in E_1$ and $y \in E_2$ are independent for any $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$.

An important special case is the n-fold product of a space with itself $(\Omega, \mathcal{F}, \Pr)^n = (\Omega^n, \mathcal{F}^n, \Pr^n)$. This represents n independent repetitions of the random experiment represented by the original space. In this case we often (inaccurately) simply use $\Pr$ to denote $\Pr^n$.

Example 1.3: Suppose we have two procedures F and G that compute integer-valued functions f and g. We only know that the functions f and g are polynomials of degree at most d. We wish to find out whether $f = g$ by just making calls to F and G (which we assume are "black boxes").

If $f = g$, then $f(x) - g(x) = 0$ for all x.

If $f \neq g$, then $f - g$ is a polynomial of degree at most d which is not identically zero. Hence $f(x) - g(x)$ holds for at most d values $x \in \mathbb{N}$. In particular, the set $\{1, \ldots, rd\}$ for any $r \in \mathbb{N}$ includes at least $(r-1)d$ elements x for which $f(x) - g(x) \neq 0$.

We take the following basic algorithm as a starting point:

1. Pick a random $x \in \{1, \ldots, rd\}$.
2. If $f(x) - g(x) \neq 0$, print "not equal".
3. Otherwise print "equal".

Based on the previous observations,

- if $f = g$, the algorithm always prints "equal"
- if $f \neq g$, the algorithm has at least probability $(r-1)d/(rd) = 1 - 1/r$ of printing "not equal."

We say the algorithm has one-sided probability of error at most $1/r$.

Let's now make k independent trials as follows:

1. Pick mutually independent $x_1, \ldots, x_k$ uniformly from $\{1, \ldots, rd\}$.
2. If $f(x_i) - g(x_i) \neq 0$ for at least one i, print "not equal."
3. Otherwise print "equal."

If $f = g$ we again always get "equal." If $f \neq g$ and we got "equal," we had at k independent trials in a row an event with probability at most $1/r$. This can happen with probability at most $(1/r)^k$.

Hence, by increasing the number of iterations we can make the error probability approach zero at an exponential rate. □

Law of Total Probability [M&U Thm. 1.6]

Let $\{ E_i \mid i \in I \}$ be a finite or countably infinite set of disjoint events such that $\cup_{i \in I} E_i = \Omega$. Directly from definitions we get

$$\Pr(B) = \sum_{i \in I} \Pr(B \cap E_i) = \sum_{i \in I} \Pr(B \mid E_i) \Pr(E_i).$$

One application for this is the principle of deferred decisions.

Suppose we want to prove a statement of the form $\Pr(x \in B) \leq \epsilon$.

We split x into two suitable components $x = (x_1, x_2)$. We think of x_1 as being fixed "first" and x_2 "only later."

We then prove that whatever the choice of x_1, the probability of choosing x_2 such that $(x_1, x_2) \in B$ holds is at most ϵ. The desired result follows by applying the law of total probability with

$$\begin{aligned} I &= \text{range of } x_1 \\ E_i &= \{ (x_1, x_2) \mid x_1 = i \}. \end{aligned}$$

Example 1.4 [M&U Thm. 1.4]**:** We are given three $n \times n$ matrices A, B and C. We want to check whether $AB = C$ holds but don't want to calculate the whole matrix product AB.

We proceed as in the previous example:

1. Pick a random $\boldsymbol{r} \in \{0,1\}^n$.
2. If $AB\boldsymbol{r} \neq C\boldsymbol{r}$, print "not equal."
3. Otherwise print "equal."

Let $D = AB - C$. We claim that if D is not the zero matrix, then $D\boldsymbol{r} \neq 0$ holds with probability at least $1/2$.

Write $D = (d_{ij})$. We assume that $D \neq 0$; let $d_{pq} \neq 0$.

If $D\boldsymbol{r} = \mathbf{0}$, we have in particular

$$\sum_{j=1}^{n} d_{pj} r_j = 0,$$

from which we can solve

$$r_q = -d_{pq}^{-1} \sum_{j \neq q} d_{pj} r_j.$$

Now imagine that we first picked $\boldsymbol{r}' = (r_1, \ldots, r_{q-1}, r_{q+1}, \ldots, r_n)$, and then consider choices for the missing component r_q. Because the choice of $\boldsymbol{r}'$ fixes some value v for the expression

$$-d_{pq}^{-1} \sum_{j \neq q} d_{pj} r_j,$$

the probability of $r_q = v$ is at most $1/2$ (as $r_q \in \{0, 1\}$).

By the principle of deferred decisions, we see that

$$\Pr(D\boldsymbol{r} = 0) \leq \frac{1}{2}.$$

□

Bayes' Law [M&U Thm. 1.7]

Again, directly from the defitions we get

$$\Pr(E_j \mid B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} = \frac{\Pr(B \mid E_j)\Pr(E_j)}{\sum_i \Pr(B \mid E_i)\Pr(E_i)}$$

where again E_j are assumed disjoint.

A usual interpretation for the formula is based on updating our beliefs when new data arrive:

- The events E_j are mutually exclusive hypotheses (with the rough interpretation $E_j =$ "theory number j is true" for some mutually contradictory competing theories).
- The even B describes some observation, measurement data etc.
- $\Pr(E_j)$ is the *prior* probability that indicates our degree of belief in hypothesis E_j before we have seen any data.
- $\Pr(B \mid E_j)$ measures how well the hypothesis E_j "explains" data B.
- $\Pr(E_j|B)$ is the *posterior* probability that indicates our degree of belief in hypothesis E_j after data B were observed.

Example 1.5: We are given three coins. Two of them are balanced, while one of them will give heads with probability 2/3; we don't know which one.

We assing the coins arbitrary numbers 1, 2 and 3 and toss them once. Suppose we get the results (1: heads, 2: heads, 3: tails).

What is the probability that coin 1 is the unbalanced one?

The Bayes' Law gives the answer 2/5. □

Remark The denominator in Bayes' Law is the same for all E_j. If we only wish to compare the posterior probabilities, we can ignore the constant factor $\Pr(B)$ and write

$$\Pr(E_j \mid B) \propto \Pr(B \mid E_j) \Pr(E_j).$$

However, in many machine learning applications computing $\Pr(B)$ cannot be avoided and is actually a crucial problem for efficient algorithms.

Randomised minimum cut [M&U Section 1.4]

Let $G = (V, E)$ be a connected undirected multigraph. (Unlike in a normal graph, a multigraph may have multiple edges between two vertices.)

A set of edges $C \subseteq E$ is a cut of the (multi)graph if $(V, E - C)$ is not connected. Minimum cut (or min cut) is the cut that contains the least number of edges.

We use an operation called edge contraction. Contracting an edge (u, v) replaces the vertices u and v with a new vertex. The edge (u, v) (all the copies, if there were multiple) is removed. The other edges are retained, and the new vertex replaces u and v when they occur as an endpoint of an egde.

If now C was a cut in the original graph and $(u, v) \notin C$, then C is a cut also after the contraction. On the other hand, contrations do not introduce any new cuts.

Consider the following algorithm:

1. Pick a random edge $(u, v) \in E$ so that each edge has the same probability of being picked.
2. Contract (u, v).
3. If at least three vertices remain, return to Step 1.
4. Otherwise print the remaining edges. (We assume that the original edges retain their "identity" even if their end points are contracted.)

Let C be a min cut. We know that if no edge in C is picked for contraction, then the algorithm gives the right output.

What's the probability of this desired outcome?

Let E_i denote the event that in iteration i the edge picked for contraction is not in C. Let $F_i = \cap_{j=1}^{i} E_i$. We need a lower bound for the probability $\Pr(F_{n-2})$ where $n = |V|$.

Let $k = |C|$ be the size of the min cut. Then in particular each vertex has degree at least k, so the graph has at least $kn/2$ edges. Therefore,

$$\Pr(E_1) = \frac{|E| - |C|}{|E|} \geq 1 - \frac{k}{nk/2} = 1 - \frac{2}{n}.$$

More generally, if the first $i-1$ iterations have avoided picking from C, then C is still a min cut. However, the number of vertices has beed reduced, so we get

$$\Pr(E_i \mid F_{i-1}) \geq 1 - \frac{2}{n-i+1}.$$

We get

$$\begin{aligned}
\Pr(F_{n-2}) &= \Pr(E_{n-2} \cap F_{n-3}) \\
&= \Pr(E_{n-2} \mid F_{n-3}) \Pr(F_{n-3}) \\
&= \ldots \\
&= \Pr(E_{n-2} \mid F_{n-3}) \Pr(E_{n-3} \mid F_{n-4}) \ldots \Pr(E_2 \mid F_1) \Pr(F_1) \\
&\geq \prod_{i=1}^{n-2} \left(1 - \frac{2}{n-i+1}\right) \\
&= \left(\frac{n-2}{n}\right)\left(\frac{n-3}{n-1}\right)\cdots\left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right) \\
&= \frac{2}{n(n-1)}.
\end{aligned}$$

The algorithm always outputs a cut, and with probability at least $2/(n(n-1))$ a min cut.

We repeat the algorithm m times and choose the smallest of the m cuts.

The probability that we fail to get a min cut is at most

$$\left(1 - \frac{2}{n(n-1)}\right)^m \leq \exp\left(-\frac{2m}{n(n-1)}\right)$$

where we used the bound $1 - x \leq e^{-x}$.

For example choosing $m = n(n-1)\ln n$ makes the error probability at most $1/n^2$. $\square$

2. Random variables

Consider a probability space $(\Omega, \mathcal{F}, \Pr)$. A real-valued function $X : \Omega \to \mathbb{R}$ is a random variable if $\{ s \in \Omega \mid X(s) \leq a \} \in \mathcal{F}$ for all $a \in \mathbb{R}$.

A random variable is discrete if its range is finite or countably infinite. Later we'll also consider continuous random variables, but for now we assume our random variables to be discrete.

Usually the probability $\Pr(\{ s \in \Omega \mid X(s) = a \})$ is denoted simply by $\Pr(X = a)$, and so on. The distribution of the random variable is defined by giving the values $\Pr(X = a)$ for all $a \in \mathbb{R}$. The distribution contains all the information we usually want to know.

A sequence $(X_1, \ldots, X_k)$ of random variables is mutually independent, if for all $I \subseteq \{1, \ldots, k\}$ and for all $x_1, \ldots, x_k \in \mathbb{R}$ we have

$$\Pr(\cap_{i \in I}(X_i = x_i)) = \prod_{i \in I} \Pr(X_i = x_i).$$

Let V be the range of a random variable X. If the sum $\sum_{x \in V} |x| \Pr(X = x)$ converges, we define the expectation of X as

$$\mathbf{E}[X] = \sum_{x \in V} x \Pr(X = x).$$

Otherwise the expectation does not exist, which is often denoted by $\mathbf{E}[X] = \infty$.

The expectation is linear [M&U Thm. 2.1]: for **all** $a, b \in \mathbb{R}$ and random variables X, Y we have

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

Linearity does not automatically extend to infinite sums. Whether the equality

$$\mathbf{E}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} \mathbf{E}\left[X_i\right]$$

holds is non-trivial. One sufficient condition is that all the expectations $\mathbf{E}[|X_i|]$ are defined and $\sum_{i=1}^{\infty} \mathbf{E}[|X_i|]$ converges.

If additionally X and Y are **independent**, we have

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

Jensen's Inequality [M&U luku 2.1.2]

From definitions we can easily see

$$\mathbf{E}[X^2] \geq (\mathbf{E}[X])^2.$$

(In general this is a strict inequality as X is not independent of itself.) This is a special case of Jensen's inequality.

A function $f\colon [a,b] \to \mathbb{R}$ is convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

for all $a \leq x_1, x_2 \leq b$ and $0 \leq \lambda \leq 1$.

A sufficient condition for convexity is that f is twice differentiable and $f''(x)$ is non-negative.

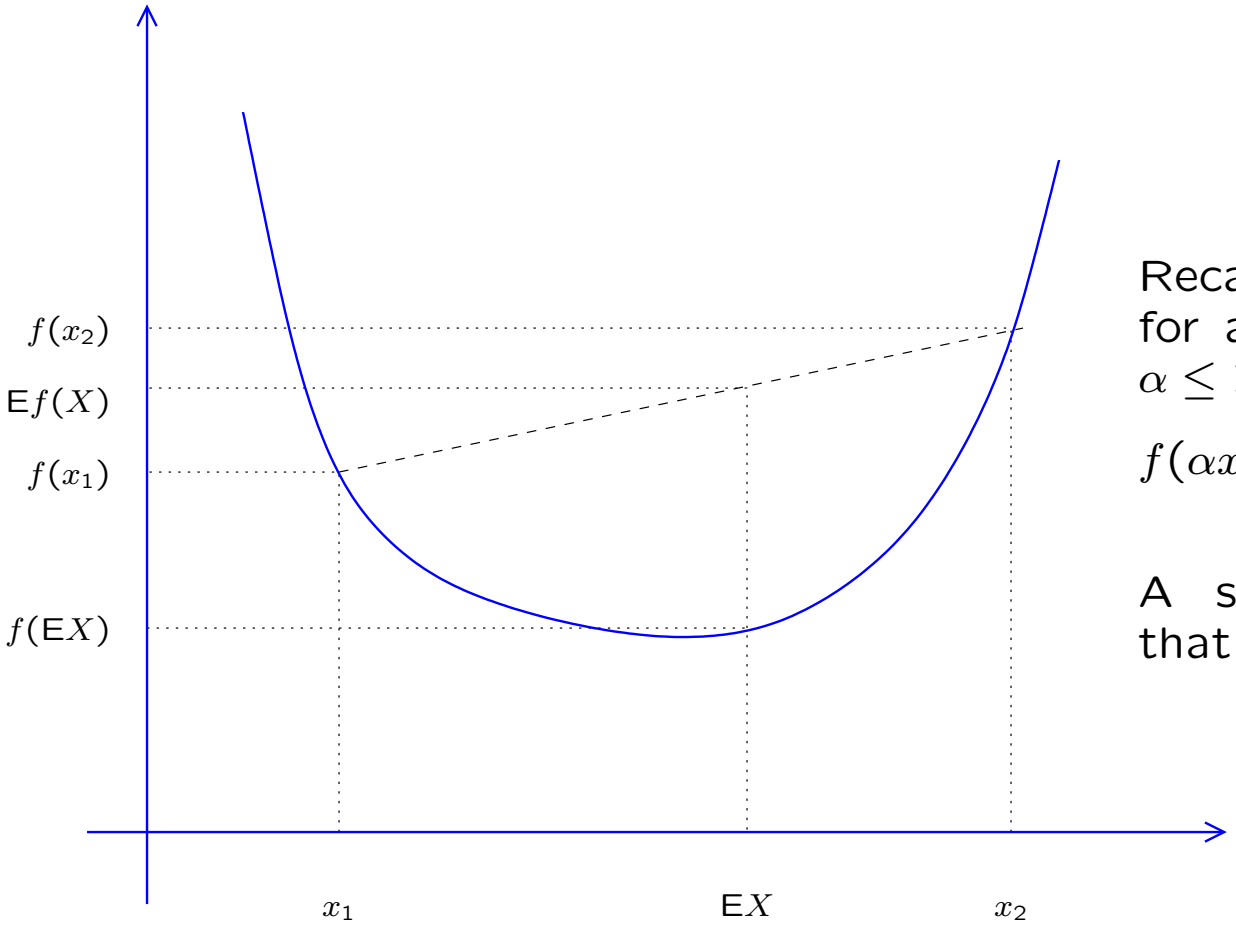
Theorem 2.1 [M&U Jensen]**:** If f is convex then

$$\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$$

for all random variables X. □

The special case at the top of the page is obtained by $f(x) = x^2$.

Jensen in a picture



Recall that f is convex if for all x_1 and x_2 and $0 \leq \alpha \leq 1$ we have

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2).$$

A sufficient condition is that $f''(x) \geq 0$ for all x.

Binomial distribution [M&U Section 2.2]

A random variable Y has Bernoulli distribution with parameter p if

$$\Pr(Y = 1) = p \quad \text{and} \quad \Pr(Y = 0) = 1 - p.$$

Then clearly $\mathbf{E}[Y] = p$.

A random variable X has binomial distribution with parameters n and p if X is the sum of n independent random variables each of which has Bernoulli distribution with parameter p. We denote this by $X \sim \text{Bin}(n, p)$. By linearity of expectation we have

$$\mathbf{E}[X] = np.$$

The distribution can be written as

$$\Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}, \qquad j = 0, \ldots, n.$$

Conditional expectation [M&U Section 2.3]

When Y and Z are random variables, the range of Y is V, and $z \in \mathbb{R}$, we write.

$$\mathbf{E}[Y \mid Z = z] = \sum_{y \in V} y \Pr(Y = y \mid Z = z).$$

Example 2.2: Let X_1 and X_2 be the results of two independent throws of a six-sided die, and $X = X_1 + X_2$. Then $\mathbf{E}[X \mid X_1 = 3] = 6\frac{1}{2}$ and

$$\mathbf{E}[X_1 \mid X = 4] = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} = 2.$$

□

For all X and Y we have

$$\mathbf{E}[X] = \sum_{y \in V} \mathbf{E}[X \mid Y = y] \Pr(Y = y)$$

assuming the expectations are defined.

The conditional expectation $\mathbf{E}[Y \mid Z]$ is a **random variable** defined as follows:

Let Y and Z be random variables over sample space Ω (that is, functions $\Omega \to \mathbb{R}$). Now $\mathbf{E}[Y \mid Z] : \Omega \to \mathbb{R}$ is the random variable for which

$$\mathbf{E}[Y \mid Z](\omega) = \mathbf{E}[Y \mid Z = Z(\omega)]$$

for all $\omega \in \Omega$.

Example 2.3: Let again $X = X_1 + X_2$ where X_1 and X_2 are independent die rolls. Now

$$\mathbf{E}[X \mid X_1] = X_1 + 3\frac{1}{2}.$$

□

Conditional expectation follows the basic rules of normal expectations: $\mathbf{E}[X_1 + X_2 \mid Z] = \mathbf{E}[X_1 \mid Z] + \mathbf{E}[X_2 \mid Z]$ etc. Additionally, we have

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y \mid Z]].$$

Example 2.4: Branching processes [M&U pp. 28–29].

Consider a situations where a process executes a certain procedure. This can in turn create more similar processes.

Let's assume that the number of new processes a process creates during its life time has distribution $\mathrm{Bin}(n, p)$. If we start with one process, how many do we get in expectation?

Let Y_i be the number of processes in "generation" i. Thus, $Y_0 = 1$ and $Y_1 \sim \mathrm{Bin}(n, p)$. Fix now some i and denote by Z_k the number of children of process number k in generation i. Therefore $Z_k \sim \mathrm{Bin}(n, p)$.

Consider the conditional expectations:

$$\begin{aligned}\mathbf{E}[Y_i \mid Y_{i-1} = y_{i-1}] &= \mathbf{E}\left[\sum_{k=1}^{y_{i-1}} Z_k \mid Y_{i-1} = y_{i-1}\right] \\ &= \mathbf{E}\left[\sum_{k=1}^{y_{i-1}} Z_k\right] \\ &= y_{i-1}np\end{aligned}$$

since Z_k and Y_{i-1} are independent. Therefore $\mathbf{E}[Y_i \mid Y_{i-1}] = npY_{i-1}$, so

$$\mathbf{E}[Y_i] = \mathbf{E}[\mathbf{E}[Y_i \mid Y_{i-1}]] = \mathbf{E}[npY_{i-1}] = np\mathbf{E}[Y_{i-1}].$$

As $Y_0 = 1$, induction yields $\mathbf{E}[Y_i] = (np)^i$. The expected total number of processes is

$$\mathbf{E}\left[\sum_{i\geq 0} Y_i\right] = \sum_{i\geq 0}(np)^i$$

which is finite if and only if $np < 1$. □

Geometric distribution [M&U Section 2.4]

A random variable X has a geometric distribution with parameter p, denoted by $X \sim \text{Geom}(p)$, if

$$\Pr(X = n) = (1-p)^{n-1}p, \qquad n = 1, 2, \ldots.$$

That is, X is the number of trials needed to get the first success in a sequence of independent trials each with probability p of success.

The geometric distribution is memoryless:

$$\Pr(X = n + k \mid X > k) = \Pr(X = n).$$

The expectation of a geometric random variable is

$$\mathbf{E}[X] = \frac{1}{p}.$$

We show this in two different ways.

Method 1: Use the equation

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} \Pr(X \geq i),$$

that holds for any X that gets only non-negative integer values.

For $X \sim \text{Geom}(p)$ we get

$$\Pr(X \geq i) = \sum_{n=i}^{\infty} (1-p)^{n-1} p = (1-p)^{i-1}.$$

Therefore

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{p}.$$

Method 2: We use the memoryless property. Let $X = \min\{\, i \mid Y_i = 1 \,\}$, where Y_i, $i = 1, 2, \ldots$ are independent Bernoulli(p) random variables.

By a well-known basic property,

$$\mathbf{E}[X] = \mathbf{E}[X \mid Y_1 = 0]\,\Pr(Y_1 = 0) + \mathbf{E}[X \mid Y_1 = 1]\,\Pr(Y_1 = 1).$$

Now $\Pr(Y_1 = 1) = p$, and $X = 1$ whenever $Y_1 = 1$. On the other hand, $Y_1 = 0$ means the same as $X > 1$. By the memoryless property,

$$\Pr(X = n + 1 \mid X > 1) = \Pr(X = n)$$

which by writing $Z = X + 1$ becomes

$$\Pr(X = m \mid X > 1) = \Pr(X = m - 1) = \Pr(Z = m), \qquad m \geq 2.$$

Therefore $\mathbf{E}[X \mid X > 1] = \mathbf{E}[Z] = \mathbf{E}[X] + 1$. We have

$$\mathbf{E}[X] = (1 - p)(\mathbf{E}[X] + 1) + p,$$

from which we can solve $\mathbf{E}[X] = 1/p$. □

Example 2.5: Coupon collector's problem [M&U Section 2.4.1]

A cereal box always contains one coupon. There are n different coupons. How many boxes of cereals do we need to buy to collect the whole set?

Let X be the random variable denoting the number of boxes needed for a full set. Let X_i be the number of boxes we bought while we already had exactly $i-1$ different coupons. Therefore

$$X = \sum_{i=1}^{n} X_i.$$

When $i-1$ coupons have been found, the probability that the next box contains a new one is $p_i = (n-i+1)/n$. Therefore, $X_i \sim \text{Geom}(p_i)$.

We get

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{i=1}^{n} \mathbf{E}[X_i] \\
&= \sum_{i=1}^{n} \frac{1}{p_i} \\
&= \sum_{i=1}^{n} \frac{n}{n-i+1} \\
&= n \sum_{j=1}^{n} \frac{1}{j} \\
&= nH(n),
\end{aligned}$$

where $H(n) = \sum_{i=1}^{n}(1/i)$. Since it is known [M&U Lemma 2.10] that

$$\ln n \leq H(n) \leq \ln n + 1,$$

we get

$$\mathbf{E}[X] = n \ln n + \Theta(n).$$

Example 2.6: Quicksort [M&U Section 2.5]

Consider a randomized version of the algorithm:

Quicksort($S[1..n]$)
If $n \leq 1$, then return S.
Pick a random $i \in \{1, \ldots, n\}$. Let $x = S[i]$.
Partition S into two sublists:
L contains elements less than x
H contains elements greater than x.
Return [Quicksort(L), x, Quicksort(H)].

The element x is called the pivot.

Worst case: The pivot is always the smallest or largest element of the list. We need $n(n-1)/2 = \Theta(n^2)$ comparisons.

Average case: Let X be the number of comparisons made by Quicksort.

Let the elements of S in ascending order be $y_1, \ldots, y_n$. Write $X_{ij} = 1$ if during the procedure the elements y_i and y_j are compared, otherwise $X_{ij} = 0$. Since no pair of elements is compared twice, we have

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}.$$

Fix some $i < j$. By studying the algorithm we can see that $X_{ij} = 1$ holds if and only if either y_i or y_j is the first pivot picked from the set $Y^{ij} = \{ y_i, y_{i+1}, \ldots, y_{j-1}, y_j \}$. Since all pivots are equally likely, we get

$$\mathbf{E}[X_{ij}] = \Pr(X_{ij} = 1) = \frac{2}{j - i + 1}.$$

We can now calculate

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} \\
&= \sum_{i=1}^{n-1} \sum_{k=2}^{n-i+1} \frac{2}{k} \\
&= \sum_{k=2}^{n} \sum_{i=1}^{n+1-k} \frac{2}{k} \\
&= \sum_{k=2}^{n} (n+1-k)\frac{2}{k} \\
&= (n+1) \sum_{k=2}^{n} \frac{2}{k} - 2(n-1) \\
&= (2n+2)H(n) - 4n.
\end{aligned}$$

Therefore, the expected number of comparisons is $\mathbf{E}[X] = 2n \ln n + \Theta(n)$.

For comparison, consider a simplified deterministic Quicksort where the pivot is always the first element of the list: $x = S[1]$.

If now the input is originally in random order, with all permutations equally likely, the expected number of comparisons is again $2n \ln n + \Theta(n)$. This can be seen with a similar argument as above. Now y_i and y_j will be compared if either one of them is in the input before other elements of Y^{ij}.

Remark In this second version the expectations is over inputs, not over any randomness in the algorithm (since there is none). Whether the underlying assumption about the distribution of the inputs is correct is often debatable. Of course in theory we could make sure by randomly permuting the input before sorting. □

3. Moments and deviations

The expectation by itself does not give a very good picture of the distribution of a random varible. The next step is typically to calculate the variance.

Variance and other quantities describing the "width" of the distribution are also useful for proving "tail bounds" (upper bounds for the probability of getting very large or very small values). Often in computer science (and also in statistics) these may be topics of primary interest.

The first technique for estimating tails is based on Markov's Inequality [M&U Thm. 3.1]: if X is a non-negative random variable, then

$$\Pr(X \geq a) \leq \frac{\mathbf{E}[X]}{a}.$$

Proof:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_x x \Pr(X = x) \\
&= \sum_{x<a} x \Pr(X = x) + \sum_{x \geq a} x \Pr(X = x) \\
&\geq 0 + a \sum_{x \geq a} \Pr(X = x)
\end{aligned}$$

where summations are over the range of X. □

Example 3.1: We toss a symmetrical coin n times. We want an upper bound for the probability of getting at least $3n/4$ heads.

If X is the number of heads, then $X \geq 0$ and $\mathbf{E}[X] = n/2$. Therefore,

$$\Pr(X \geq 3n/4) \leq \frac{n/2}{3n/4} = \frac{2}{3}.$$

This is a very crude estimate that did not even try to take into account any information about the shape of the distribution. Indeed, because of symmetry it's clear that the probability in question cannot be larger than $1/2$. □

Moments and variance [M&U Section 3.2]

The kth moment of a random variable X is $\mathbf{E}[X^k]$.

The variance of X is

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

and standard deviation

$$\sigma[X] = \sqrt{\mathbf{Var}[X]}.$$

The covariance of X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

From definitions and the linearity of expectation we get

$$\begin{aligned} \mathbf{Var}[X] &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ \mathbf{Var}[X + Y] &= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}[X, Y]. \end{aligned}$$

If X and Y are **independent**, we have

$$\begin{aligned} \mathbf{E}[XY] &= \mathbf{E}[X]\mathbf{E}[Y] \\ \mathbf{Cov}(X,Y) &= 0 \\ \mathbf{Var}[X+Y] &= \mathbf{Var}[X] + \mathbf{Var}[Y] \end{aligned}$$

By induction, this can be generalized for sums and products of more than two random variables.

Example 3.2: If $X_i \sim \text{Bernoulli}(p)$, a direct calculation gives

$$\mathbf{Var}[X_i] = p(1-p).$$

Therefore, if X is the sum of n independent Bernoulli(p) random variables, that is $X \sim \text{Bin}(n,p)$, we have

$$\mathbf{Var}[X] = np(1-p).$$

□

Chebyshev's inequality [M&U Section 3.3]

Theorem 3.3 [M&U Thm 3.6]: For any $a > 0$ we have

$$\Pr(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2}.$$

Proof: Write the probability in question as

$$\Pr(|X - \mathbf{E}[X]| \geq a) = \Pr((X - \mathbf{E}[X])^2 \geq a^2).$$

Applying Markov's Inequality to the non-negative random variable $Y = (X - \mathbf{E}[X])^2$ gives us

$$\Pr(Y \geq a^2) \leq \frac{\mathbf{E}[Y]}{a^2} = \frac{\mathbf{Var}[X]}{a^2}.$$

□

Example 3.4: Consider the same question we already analysed using Markov's Inequality: What is the probability of getting at least $3n/4$ heads when a symmetric coin is tosses n times?

Since X is binomially distributed, we have $\mathbf{E}[X] = n/2$ and $\mathbf{Var}[X] = n\frac{1}{2}(1-\frac{1}{2}) = n/4$. Therefore,

$$\Pr(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}) \leq \frac{\mathbf{Var}[X]}{(n/4)^2} = \frac{4}{n}.$$

By symmetry,

$$\Pr(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}) = 2\Pr(X - \frac{n}{2} \geq \frac{n}{4}),$$

so

$$\Pr(X \geq \frac{3n}{4}) \leq \frac{2}{n}.$$

□

Actually even this is extremely loose for large n. We get much better estimates by using Chernoff bounds that will be introduced soon.

Example 3.5: Coupon Collector's Problem (Example 2.5 continued)

We obtained $nH(n)$ as the expectation of the number X of cereal boxes we need to buy. Markov's Inequality then yields

$$\Pr(X \geq 2nH(n)) \leq \frac{1}{2}.$$

To apply Chebyshev, we also need the variance $\mathbf{Var}[X]$. Remember $X = \sum_{i=1}^{n} X_i$ where $X_i \sim \text{Geom}(p_i)$ and $p_i = (n - i + 1)/n$. The variance of $X \sim \text{Geom}(p)$ is known [M&U Lemma 3.8] to be

$$\mathbf{Var}[X] = \frac{1-p}{p^2}.$$

The random variables X_i are mutually independent, so

$$\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i].$$

By estimating $\mathbf{Var}[X_i] \leq 1/p_i^2$ we get

$$\sum_{i=1}^{n} \mathbf{Var}[X_i] \leq \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 \leq n^2 \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2 n^2}{6}.$$

Therefore, Chebyshev's Inequality gives us

$$\Pr(|X - nH(n)| \geq nH(n)) \leq \frac{\pi^2 n^2/6}{(nH(n))^2} = O\left(\frac{1}{(\log n)^2}\right).$$

This is not a very tight bound, either. The probability that the first $n(c + \ln n)$ fail to contain a given specific coupon is

$$\left(1 - \frac{1}{n}\right)^{n(c+\ln n)} \leq \exp(-(c + \ln n)).$$

Therefore, the probability that some coupon has failed to appear in the first $n(c + \ln n)$ boxes is by union bound at most $n \exp(-(c + \ln n)) = \mathrm{e}^{-c}$. Substituting $c = \ln n$ yields

$$\Pr(X \geq 2n \ln n) \leq \frac{1}{n}.$$

□

Randomized algorithm for the median [M&U Section 3.4]

Let S be a set of numbers for which we want to determine the median. For simplicity we consider the case where $n = |S|$ is odd, so the median is the element at position $\lceil n/2 \rceil$ in the ordering of the elements of S.

The median can easily be determined in time $O(n \log n)$ by sorting. There is also a (somewhat complicated) deterministic algorithm that runs in time $O(n)$. We give here a simple randomized algorithm with running time $O(n)$.

The idea is to use randomization to pick a lower bound d and upper bound u such that with high probability,

1. the median is between d and u and

2. there are not too many elements of S between d and u.

Ignoring for the moment how exactly we choose d and u, the algorithm is then

1. Choose d and u.
2. Create the set $C = \{ x \in S \mid d \leq x \leq u \}$ and calculate $\ell_d = |\{ x \in S \mid x < d \}|$ and $\ell_u = |\{ x \in S \mid u < x \}|$.
3. If $\ell_d > n/2$ or $\ell_u > n/2$, then **fail**.
4. If $|C| > 4n^{3/4}$, then **fail**.
5. Otherwise sort C and return its element number $\lfloor n/2 \rfloor - \ell_d + 1$.

If d and u are chosen in time $O(n)$, then clearly the whole algorithm runs in time $O(n)$.

If the algorithm does not fail, it gives the right answer. By repeating until it succeeds we therefore get a Las Vegas algorithm that always gives the right answer but may sometimes run for a long time.

The interesting part of the analysis is to determine d and u such that the failure probability is small.

(From now on we ignore rounding.)

We propose choosing d and u as follows:

1. Choose a (multi)set $R \subseteq S$ by choosing independently $n^{3/4}$ elements uniformly (with replacement) from S.

2. Sort R.

3. Now d is the element number $\frac{1}{2}n^{3/4} - n^{1/2}$ and u the element number $\frac{1}{2}n^{3/4} + n^{1/2}$ in the ordering of R.

Intuitively, the median of R, that is the element number $\frac{1}{2}n^{3/4}$, is also an estimate for the median of whole S. The first "fail" branch is the case where this estimate is badly off.

Between d and u there are $2n^{1/2}$ elements of R.

Therefore, if sampling has been uniform, then between d and u there are $2n^{1/2}(n/n^{3/4}) = 2n^{3/4}$ elements of S. The second "fail" branch is the case where the sample is not sufficiently uniform.

The numbers $n^{3/4}$, $n^{1/2}$ etc. come from known bounds for sampling accuracy. (In other words, they have been chosen so that the following proof goes through.)

We'll now derive an upper bound for the failure probability. Let m be the actual median of S, and $k = |R| = n^{3/4}$. Consider the following three events:

$$\begin{aligned} \mathcal{E}_1 &: |\{\, r \in R \mid r \leq m \,\}| < \frac{k}{2} - n^{1/2} \\ \mathcal{E}_2 &: |\{\, r \in R \mid r \geq m \,\}| < \frac{k}{2} - n^{1/2} \\ \mathcal{E}_3 &: |C| > 4k. \end{aligned}$$

The event $\mathcal{E}_3$ obviously represents the second "fail" case.

The events $\mathcal{E}_1$ and $\mathcal{E}_2$ correspond to $m < d$ ja $m > u$. Hence, together they cover the first "fail" case.

To estimate $\Pr(\mathcal{E}_1)$ write $Y_1 = |\{\, r \in R \mid r \leq m \,\}|$. Thus, $Y_1 = \sum_{i=1}^{k} X_i$ where

$$X_i = \begin{cases} 1 & \text{if sample point number } i \text{ is less than or equal to } m \\ 0 & \text{otherwise.} \end{cases}$$

There are $(n-1)/2+1$ elements in S that are less than or equal to m. Therefore, $Y_1 \sim \mathsf{Bin}(k,p)$ where $p = 1/2 + 1/(2n)$. Hence, $\mathbf{E}[Y_1] \geq k/2$, and

$$\mathbf{Var}[Y_1] = k\left(\frac{1}{2}+\frac{1}{2n}\right)\left(\frac{1}{2}-\frac{1}{2n}\right) < \frac{k}{4}.$$

We apply Chebyshev's Inequality:

$$\Pr(\mathcal{E}_1) \leq \Pr(|Y_1 - \mathbf{E}[Y_1]| > n^{1/2}) \leq \frac{\mathbf{Var}[Y_1]}{n} \leq \frac{1}{4}n^{-1/4}.$$

Similarly we see that

$$\Pr(\mathcal{E}_2) \leq \frac{1}{4} n^{-1/4}.$$

For the event $\mathcal{E}_3$ we consider two subevents:

$$\begin{aligned} \mathcal{E}_{3,1} &: |\{\, c \in C \mid c > m \,\}| \geq 2k \\ \mathcal{E}_{3,2} &: |\{\, c \in C \mid c < m \,\}| \geq 2k. \end{aligned}$$

If $|C| > 4k$, then at least one of these has occurred. The subevents are symmetrical, so let's consider $\mathcal{E}_{3,1}$. Now the position of u in S is at least $n/2 + 2k$. Hence, u and any larger elements in R are among the $n/2 - 2k$ largest elements of S. By definition of u, there are $k/2 - n^{1/2}$ such elements.

Define

$$X_i = \begin{cases} 1 & \text{if smaple point number } i \text{ is among the } n/2 - 2k \text{ largest elements in } S \\ 0 & \text{otherwise} \end{cases}$$

and $X = \sum_{i=1}^{k} X_i$. Again, X has binomial distribution,

$$\mathbf{E}[X] = \frac{k}{2} - 2n^{1/2}$$

and

$$\mathbf{Var}[X] = k\left(\frac{1}{2} - 2n^{-1/4}\right)\left(\frac{1}{2} + 2n^{-1/4}\right) < \frac{k}{4}.$$

Therefore,

$$\Pr(\mathcal{E}_{3,1}) \leq \Pr(|X - \mathbf{E}[X]| \geq n^{1/2}) \leq \frac{\mathbf{Var}[X]}{n} < \frac{1}{4}n^{-1/4}.$$

Altogether, the probability of failure is at most

$$\Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) + \Pr(\mathcal{E}_{3,1}) + \Pr(\mathcal{E}_{3,1}) < n^{-1/4}.$$

□

4. Chernoff bounds

"Chernoff bounds" is a general name for several inequalities that estimate how tightly the value of a random variable is concentrated around its expectation.

Basic example: If $X \sim \text{Bin}(n,p)$, then for all $0 < \delta \leq 1$ we have

$$\Pr\left(\frac{X-np}{np} \geq \delta\right) \leq \exp\left(-\frac{1}{3}np\delta^2\right).$$

For example, this implies that with probability $1/2$ we have

$$X \leq np + \sqrt{3np\ln 2}.$$

This bounds can be made (a) more general and (b) tighter.

In this section we review some bounds of this variety, including their proofs and applications.

Moment Generating Function [M&U Section 4.1]

The moment generating function of a random variable X is defined as

$$M_X(t) = \mathbf{E}[\mathrm{e}^{tX}],$$

assuming this expectation is finite.

By differentiating the moment generating function n times in the origin we get the nth moment.

Theorem 4.1: If $M_X(t)$ is defined in some neighbourhood $t \in (-\delta, \delta)$ of 0, we have

$$\mathbf{E}[X^n] = M_X^{(n)}(0)$$

for all $n = 1, 2, \ldots$.

Proof: We defined

$$M_X(t) = \sum_x \Pr(X = x) \exp(tx).$$

Under the given assumptions, we can differentiate termwise:

$$M_X^{(n)}(t) = \sum_x \Pr(X = x) x^n \exp(tx).$$

Substituting $t = 0$ gives the result. □

Example 4.2: When $X \sim \text{Geom}(p)$, we have

$$\begin{aligned}
\mathbf{E}[\mathrm{e}^{tX}] &= \sum_{k=1}^{\infty}(1-p)^{k-1}p\mathrm{e}^{tk} \\
&= \frac{p}{1-p}\sum_{k=1}^{\infty}((1-p)\mathrm{e}^{t})^{k} \\
&= \frac{p}{1-p}\left(\frac{1}{1-(1-p)\mathrm{e}^{t}}-1\right)
\end{aligned}$$

from which we get the derivatives

$$\begin{aligned}
M'_X(t) &= \frac{p\mathrm{e}^{t}}{(1-(1-p)\mathrm{e}^{t})^{2}} \\
M''_X(t) &= \frac{2p(1-p)\mathrm{e}^{2t}}{(1-(1-p)\mathrm{e}^{t})^{3}} + \frac{p\mathrm{e}^{t}}{(1-(1-p)\mathrm{e}^{t})^{2}}.
\end{aligned}$$

By substituting $t = 0$ we get the familiar results $\mathbf{E}[X] = 1/p$ and $\mathbf{E}[X^2] = (2-p)/p^2$. □

It can be proved (but we will not do so on this course) that giving the moment generating function (or alternatively giving all the moments) specifies the distribution uniquely.

Theorem 4.3: If X and Y are random variables such that for some $\delta > 0$ we have $M_X(t) = M_Y(t)$ for all $-\delta < t < \delta$, the X and Y have the same distribution. $\square$

This can be used for example to determine the distribution of the product of two independent random variables together with the following.

Theorem 4.4: If X and Y are independent, we have

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Proof: Now also e^{tX} and e^{tY} are independent, so

$$\mathbf{E}[\mathrm{e}^{t(X+Y)}] = \mathbf{E}[\mathrm{e}^{tX}\mathrm{e}^{tY}] = \mathbf{E}[\mathrm{e}^{tX}]\mathbf{E}[\mathrm{e}^{tY}].$$

$\square$

Deriving Chernoff bounds [M&U Section 4.2.1]

The basic technique is to apply Markov's Inequality to the random variable e^{tX} for a suitable $t \in \mathbb{R}$. Thus,

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

for any $t > 0$, so in particular

$$\Pr(X \geq a) \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

For a negative t the direction of the inequality changes, so

$$\Pr(X \leq a) \leq \min_{t<0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

To make use of this observation, we need an estimate for the moment generating function $\mathbf{E}[e^{tX}]$ and a good choice for t.

Often we introduce bounds that are a bit loose to make the formulas more intelligible.

In the most widely used variant we take $X = \sum_{i=1}^{n} X_i$ where $X_i \sim$ Bernoulli(p_i) are independent. We call the random variables X_i Poisson trials. If the distributions are identical, with $p_i = p$ for all i, we call them Bernoulli trials.

Write $\mu = \mathbf{E}[X] = \sum_{i=1}^{n} p_i$. We estimate the probabilities $\Pr(X \geq (1+\delta)\mu)$ and $\Pr(X \leq (1-\delta)\mu)$.

First let us consider the moment generating functions for the individual trials

$$M_{X_i}(t) = p_i e^{t \cdot 1} + (1 - p_i) e^{t \cdot 0} = 1 + p_i(e^t - 1) \leq \exp(p_i(e^t - 1))$$

where we applied the inequality $1 + z \leq e^z$. This implies

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t) \leq \exp\left(\sum_{i=1}^{n} p_i(e^t - 1)\right) = \exp\left((e^t - 1)\mu\right).$$

We now derive bounds for the probability that X gets very large or very small values.

We first prove a bound that's (fairly) tight but somewhat difficult to use. From this we can derive variants that are easier to use but less tight.

Theorem 4.5 [M&U Thm 4.4.1]**:** For all $\delta > 0$ we have

$$\Pr(X \geq (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}.$$

Proof: As noted above, for $t > 0$ Markov's Inequality yields

$$\Pr(X \geq (1+\delta)\mu) = \Pr(e^{tX} \geq e^{t(1+\delta)\mu}) \leq \frac{\mathbf{E}[e^{tX}]}{\exp(t(1+\delta)\mu)}.$$

We choose $t = \ln(1+\delta)$, which gives us

$$\mathbf{E}[e^{tX}] \leq \exp((e^t - 1)\mu) = e^{\delta\mu}$$

and

$$\exp(t(1+\delta)\mu) = (1+\delta)^{(1+\delta)\mu}.$$

□

The following simplification is often useful:

Theorem 4.6 [M&U Thm 4.4.2]**:** For $0 < \delta \leq 1$ we have

$$\Pr(X \geq (1+\delta)\mu) \leq \exp(-\mu\delta^2/3).$$

Proof: It is sufficient to show

$$\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \leq e^{-\delta^2/3}$$

or equivalently (by taking log of both sides) $f(\delta) \leq 0$ where

$$f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{1}{3}\delta^2.$$

Take now derivatives:

$$\begin{aligned} f(\delta) &= \delta - (1+\delta)\ln(1+\delta) + \frac{1}{3}\delta^2 \\ f'(\delta) &= -\ln(1+\delta) + \frac{2}{3}\delta \\ f''(\delta) &= -\frac{1}{1+\delta} + \frac{2}{3}. \end{aligned}$$

Thus, $f''(\delta) < 0$ for $0 \leq \delta < 1/2$, so $f'(\delta)$ is decreasing. On the other hand, $f''(\delta) > 0$ for $1/2 < \delta < 1$, so $f'(\delta)$ is increasing.

Since $f'(0) = 0$ and $f'(1) = 2/3 - \ln 2 \approx 2/3 - 0{,}69 < 0$, we get $f'(\delta) \leq 0$ for all $0 \leq \delta \leq 1$.

Since $f(0) = 0$, we get $f(\delta) \leq 0$ for all $0 < \delta < 1$. □

Another simplification is the following:

Theorem 4.7 [M&U Thm 4.4.3]**:** For $R \geq 6\mu$ we have

$$\Pr(X \geq R) \leq 2^{-R}.$$

Proof: Write $R = (1+\delta)\mu$, so $\delta = R/\mu - 1 \geq 5$. We get

$$\begin{aligned}
\left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} &\leq \left(\frac{\mathrm{e}}{1+\delta}\right)^{(1+\delta)\mu} \\
&\leq \left(\frac{\mathrm{e}}{6}\right)^{R} \\
&\leq 2^{-R}.
\end{aligned}$$

□

Next consider the probability that X is very small.

Theorem 4.8 [M&U Thm 4.5.1]**:** For all $0 < \delta < 1$ we have

$$\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.$$

Proof: As earlier, for all $t < 0$ we have

$$\Pr(X \leq (1-\delta)\mu) \leq \frac{\mathbf{E}[\mathrm{e}^{tX}]}{\mathrm{e}^{t(1-\delta)\mu}} \leq \frac{\exp((e^t-1)\mu)}{\exp(t(1-\delta)\mu)}.$$

Substituting $t = \ln(1-\delta)$ gives the desired bound. □

We can simplify this as earlier.

Theorem 4.9 [M&U Thm 4.5.2]**:** For all $0 < \delta < 1$ we have

$$\Pr(X \leq (1-\delta)\mu) \leq \exp(-\mu\delta^2/2).$$

Proof: similar to the case for "$(1+\delta)$"; details omitted. □

We can use the union bound to get a combined bound.

Corollary 4.10 [M&U Cor 4.6]**:** For all $0 < \delta < 1$ we have

$$\Pr(|X-\mu| \leq \delta\mu) \leq 2\exp(-\mu\delta^2/3).$$

□

Coin flips [M&U Section 4.2.2]

We flip a fair coin n times. Thus, $\mu = n/2$. What kind of a bound can we have with probability $2/n$ (that is, the probability of violation gets vanishingly small for large n)?

We want $\exp(-(n/2)\delta^2/3) = 1/n$, so we take $\delta = \sqrt{(6 \ln n)/n}$. By plugging this into the bound we get

$$\Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{1}{2}\sqrt{6n \ln n}\right) \leq \frac{2}{n}.$$

Therefore, with a very high probability the deviations are at most $O(\sqrt{n \log n})$.

Compare this with the earlier Chebyshev bound

$$\Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{4}{n}.$$

By using Chernoff to estimate the same probability we get

$$\Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq 2e^{-n/24}$$

which is exponentially tighter.

Application: parameter estimation [M&U Section 4.2.3]

We pick repeated independent samples from a fixed distribution that is known to be Bernoulli(p) for some unknown p. We wish to estimate p based on the sample.

Let $X = \sum_{i=1}^{n} X_i$ be the result of n trials and $\tilde{p} = X/n$. Clearly $\mathbf{E}[\tilde{p}] = \mu/n = p$. What about error probabilities?

We call an interval $[\tilde{p} - \delta, \tilde{p} + \delta]$ a $(1-\gamma)$ confidence interval for parameter p if

$$\Pr(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \geq 1 - \gamma.$$

Interpretation: After seeing a trial sequence with relative frequency $\tilde{p}$ of ones, we have "confidence" $1 - \gamma$ for the true parameter value p belonging to the interval $[\tilde{p} - \delta, \tilde{p} + \delta]$. If p is not in the interval, then we have just observed a very unlikely deviation (probability less than γ).

Notice The parameter p is a constant, it does not have any disribution (unless we assign a prior to it which is a quite different way of thinking about this).

If $p \notin [\tilde{p} - \delta, \tilde{p} + \delta]$, then one of the following events has occurred:

$p < \tilde{p} - \delta$: therefore, $X = n\tilde{p} > n(p + \delta) = \mu(1 + \delta/p)$.

$p > \tilde{p} + \delta$: therefore, $X = n\tilde{p} < n(p - \delta) = \mu(1 - \delta/p)$.

Chernoff bounds give us

$$\Pr(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) \leq \mathrm{e}^{-\mu(\delta/p)^2/2} + \mathrm{e}^{-\mu(\delta/p)^2/3} = \mathrm{e}^{-n\delta^2/(2p)} + \mathrm{e}^{-n\delta^2/(3p)}.$$

Since p is not known, we use the conservative upper bound $p \leq 1$. Based on that we can choose

$$\gamma = \mathrm{e}^{-n\delta^2/2} + \mathrm{e}^{-n\delta^2/3}$$

(or conversely solve δ from here if γ has been chosen).

Bounds for some special cases [M&U Section 4.3]

Consider the case where X_i is has two symmetrical values.

Theorem 4.11 [M&U Thm 4.7]: If $\Pr(X_i = 1) = \Pr(X_i = -1) = 1/2$, then for all $a > 0$ we have

$$\Pr(X \geq a) \leq \exp\left(-\frac{a^2}{2n}\right).$$

Proof: For all $t > 0$ we have

$$\mathbf{E}[e^{tX_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t}.$$

We apply

$$e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!}.$$

This yields

$$\begin{aligned}
\mathbf{E}[e^{tX_i}] &= \frac{1}{2}\left(1 + t + \frac{t^2}{2} + \frac{t^3}{3!} + \frac{t^4}{4!} + \ldots\right) + \frac{1}{2}\left(1 - t + \frac{t^2}{2} - \frac{t^3}{3!} + \frac{t^4}{4!} - \ldots\right) \\
&= 1 + \frac{t^2}{2} + \frac{t^4}{4!} + \ldots \\
&= \sum_{j=0}^{\infty} \frac{t^{2j}}{(2j)!} \\
&\leq \sum_{j=0}^{\infty} \frac{1}{j!}\left(\frac{t^2}{2}\right)^j \\
&= \exp\left(\frac{t^2}{2}\right).
\end{aligned}$$

Therefore,

$$\mathbf{E}[e^{tX}] = \prod_{i=1}^{n} \mathbf{E}[e^{tX_i}] \leq \exp\left(\frac{t^2 n}{2}\right),$$

so

$$\Pr(X \geq a) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}} \leq \exp\left(\frac{t^2 n}{2} - ta\right).$$

Choosing $t = a/n$ gives the desired result

$$\Pr(X \geq a) \leq \exp\left(-\frac{a^2}{2n}\right).$$

□

Corollary 4.12: If $\Pr(X_i = 1) = \Pr(X_i = -1) = 1/2$, then for all $a > 0$ we have

$$\Pr(|X| \geq a) \leq 2\exp\left(-\frac{a^2}{2n}\right).$$

□

Corollary 4.13 [M&U Cor 4.8]: Let Y_i be mutually independent with $\Pr(Y_i = 1) = \Pr(Y_i = 0) = 1/2$. Write $Y = \sum_{i=1}^{n} Y_i$ and $\mu = \mathbf{E}[Y] = n/2$. Now for all $a > 0$ we have

$$\Pr(Y \geq \mu + a) \leq \exp\left(-\frac{2a^2}{n}\right)$$

and for all $\delta > 0$ we have

$$\Pr(Y \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right).$$

Proof: Let X_i be as before and $Y_i = \frac{1}{2}(X_i + 1)$. In particular, $Y = \frac{1}{2}X + \mu$.

From the previous theorem we get

$$\Pr(Y \geq \mu + a) = \Pr(X \geq 2a) \leq \exp\left(-\frac{4a^2}{2n}\right).$$

For the second part, choose $a = \delta\mu$, so

$$\Pr(Y \geq (1+\delta)\mu) = \Pr(X \geq 2\delta\mu) \leq \exp\left(-\frac{4\delta^2\mu^2}{2n}\right) = \exp\left(-\frac{\delta^2\mu}{2}\right).$$

□

Similarly we can prove

Corollary 4.14 [M&U Cor 4.9]**:** Let Y_i be mutually independent and $\Pr(Y_i = 1) = \Pr(Y_i = 0) = 1/2$. Write $Y = \sum_{i=1}^{n} Y_i$ and $\mu = \mathbf{E}[Y] = n/2$. Now for all $0 < a < \mu$ we have

$$\Pr(Y \leq \mu - a) \leq \exp\left(-\frac{2a^2}{n}\right)$$

and for all $\delta > 0$ we have

$$\Pr(Y \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right).$$

□

Application: set balancing [M&U Section 4.4]

Suppose we have a set of m people and n properties. We wish to partition the people into two sets A and $\overline{A}$ such that for $i = 1, \ldots, n$ we have

$$|\{\, p \in A \mid p \text{ has property } i \,\}| \approx \left|\left\{\, p \in \overline{A} \mid p \text{ has property } i \,\right\}\right|.$$

Let's define an array $A = (a_{ij}) \in \{\, 0, 1 \,\}^{n \times m}$ where $a_{ij} = 1$ if person j has property i.

We represent a partition $(A, \overline{A})$ as a vector $\boldsymbol{b} \in \{\, -1, 1 \,\}^m$ where $b_j = 1$ if person j is in the set A.

With these notations, we thus wish to minimise the quantity

$$\|A\boldsymbol{b}\|_\infty = \max_i |c_i|$$

where $c_i = \sum_j a_{ij} b_j$.

How good a result do we get by choosing $\boldsymbol{b}$ at random so that each b_j is 1 with probability 1/2 independently of each other?

We claim that

$$\Pr(\|A\boldsymbol{b}\|_\infty \geq \sqrt{4m \ln n}) \leq \frac{2}{n}.$$

We prove this by showing for each individual row $i \in \{1, \ldots, n\}$ that the event $|c_i| \geq \sqrt{4m \ln n}$ has probability at most $2/n^2$.

Write $k = \sum_j a_{ij}$. If $k \leq \sqrt{4m \ln n}$, the claim clearly holds.

Otherwise $a_{ij}b_j$ get values 1 and -1 symmetrically and independently for those j for which $a_{ij} \neq 0$. Therefore,

$$\Pr\left(\left|\sum_j a_{ij}b_j\right| > \sqrt{4m \ln n}\right) \leq 2 \exp\left(-\frac{4m \ln n}{2k}\right) \leq \frac{2}{n^2}$$

since $k \leq m$. □

Example: packet routing [M&U Section 4.5]

Consider a graph of N nodes, some of which may be connected with an edge. We assume the edges to be directed, but the particular topologies we consider are symmetrical: there's an edge (v, v') if and only if there's an edge (v', v).

The task is to transmit a set of packets through the network. Each packet has a start node and a destination node (address). The route of a packet is a path in the graph from the start node to the destination.

During one time step,

- each packet may travel at most one edge
- at most one packet may travel along any single edge.

We assume sufficient buffer memory in the nodes so packets can wait for an edge to become available.

To determine how the network works, we must fix

- how to choose the route of a packet when the start node and address are known
- if several packets wish to use the same edge, how are they priorized (queueing).

For the results we are going to present, the queueing strategy is not important, as long as we never let an edge sit idle if there are packets for it.

Possible congestion in the network of course depends on between which nodes the packets are addressed. Here we consider permutation routing: each node has exactly one packet starting from it, and one addressed to it.

Routing in hypercube [M&U Section 4.5.1]

The routing problem is interesting mainly in sparse graphs (number of edges much less than $N(N-1)$). As an example we consider the hypercube topology. In an n-dimensional hypercube, or n-cube, there are $N = 2^n$ nodes, and we identify the nodes with elements of the set $\{0,1\}^n$. In a hypercube, the nodes $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ are connected if there is exactly one index i such that $a_i \neq b_i$.

Thus, there are $N \log_2 N$ edges in the hypercube, and the diameter (longest distance between two nodes) is $\log_2 N$.

The starting point for routing in an n-cube is the bit-fixing algorithm.

Consider a packet starting from node $a = (a_1, \ldots, a_n)$ with destination $b = (b_1, \ldots, b_n)$. For $i = 1, \ldots, n+1$, define

$$v_i = (b_1, \ldots, b_{i-1}, a_i, \ldots, a_n).$$

The packet is now routed via the nodes $a = v_1, v_2, \ldots, v_{n+1} = b$. (The actual path is obtained by removing from this the repetitions that occur when $a_i = b_i$.)

Hence, we "fix" the address of the packet bit by bit, from left to right.

The bit fixing algorithm has good average case performance when the addresses are picked at random. However, it can be shown that in some cases it leads to congestion and takes $\Omega(N^{1/2})$ steps to deliver all the packets.

To avoid the worst case of bit-fixing we consider randomized two-phase routing:

Phase I: Pick for each packet a random node as an intermediate point. Route the packets to their intermediate points by bit-fixing.

Phase II: Route the packets from the intermediate points to their actual destinations.

We will show that with probability $1 - O(N^{-1})$ this two-phase routing delivers all packets in time $O(\log N)$. Since $\log_2 N$ is the diameter of the graph, this is optimal (up to a constant factor).

How we handle queues will clearly have on effect on when a given packet crosses a given edge on its route.

To simplify the analysis, let $T(M)$ be the time packet M takes to reach its destination. Each of these $T(M)$ steps is consumed by one of the following actions:

1. packet M crosses an edge on its route, or

2. packet M is in queue as some other packet crosses an edge M would need.

Let $X(e)$ be the number of packets that have edge e on their route. Based on the above, we can make

Observation: If the route of packet M consists of edges $e_1, \ldots, e_m$, then

$$T(M) \leq \sum_{i=1}^{m} X(e_i).$$

The preceding observation allows us to ignore the queue behaviour and concentrate on the paths. For a path P consisting of edges $e_1, \ldots, e_m$, define

$$T(P) = \sum_{i=1}^{m} X(e_i).$$

Based on our observation, the time taken by the routing is always upper bounded by $\max_{P \in \mathcal{R}} T(P)$ where $\mathcal{R}$ is the set of all the paths used in routing.

This applies to any routing scenario. In particular, let T_1 and X_1 be the quantities T and X when we consider only Phase I of our algorithm. We will show that with a high probability we have $T_1(P) \leq 30n$ for all possible paths P.

Fix now some paths $P = (v_0, \ldots, v_m)$ which is a possible route in bit-fixing.

We want a high-probability upper bound for $T_1(P) = \sum_{i=1}^{m} X_1(e_i)$. As the random variables $X_1(e_i)$ are not independent, we cannot directly apply Chernoff bounds.

To resolve the problem, we first estimate the probability that at least $6n$ different packets cross at least one edge that belongs to P.

After this we show that with high probability no individual packet will use too many edges on P.

Let v_{i-1} be the ith node on path P, and j the bit on which v_{i-1} and v_i differ. We say a packet k is active in node v_{i-1}, if

1. packet k is routed through v_{i-1} and

2. when packet k reaches v_{i-1}, its jth bit has not been fixed yet, but bits $1, \ldots, j-1$ either have been fixed or were correct to start with.

Notice that we do not require that the packet actually goes from v_{i-1} to v_i. However, depending on bit j of the address, it has the potential of doing that.

For $k = 1, \ldots, N$, let $H_k = 1$ if packet k is active in at least one node on P, and $H_k = 0$ otherwise. Let $H = \sum_{k=1}^{N} H_k$.

Notice that random variables H_k are mutually independent.

Let

$$\begin{aligned} v_{i-1} &= (b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_n) \\ v_i &= (b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_n). \end{aligned}$$

By condition 2, a packet that is active in v_{i-1} started in a node of form $(*, \ldots, *, a_j, \ldots, a_n)$ where $*$ can be 0 or 1. Therefore, there are 2^{j-1} possible start nodes.

By condition 1, if a packet is active in v_{i-1}, its destination must be of form $(b_1, \ldots, b_{j-1}, *, \ldots, *)$. Therefore, if we consider some fixed node among the possible start nodes, the packet starting from the node will become active with probability 2^{-j+1}.

Therefore, the expected number of active packets in v_{i-1} is 1, so

$$\mathbf{E}[H] \leq m \cdot 1 \leq n.$$

Since the random variables H_k are mutually independent, we may apply Chernoff bounds (Theorem 4.7 [M&U Thm 4.4.3]):

$$\Pr(H \geq 6n) \leq 2^{-6n}.$$

We choose $B = \{ H \geq 6n \}$ in the estimate

$$\begin{aligned} \Pr(A) &= \Pr(A \mid B) \Pr(B) + \Pr(A \mid \overline{B}) \Pr(\overline{B}) \\ &\leq \Pr(B) + \Pr(A \mid \overline{B}). \end{aligned}$$

Therefore,

$$\Pr(T_1(P)) \geq 30n) \leq 2^{-6n} + \Pr(T_1(P) \geq 30n \mid H < 6n).$$

We will next estimate the latter conditional probability.

Assume that packet k is active in v_{i-1}.

For k to actually cross the edge (v_{i-1}, v_i), we require its jth address bit to be b_j. This has probability $1/2$.

However, we also require that the packet does not need to fix any earlier bits $1, \ldots, j-1$. Thus, the actual probability for a packet that is active in v_{i-1} to cross the edge (v_{i-1}, v_i) is at most $1/2$.

More generally, if the packet is on path P in node v_{l-1}, $l > i$, then the probability that it next goes to v_l is at most $1/2$.

On the other hand, if the packet enters v_{l-1} but does not go to v_l in the next step, it won't ever return to path P. This is because in this case one of the address bits $1, \ldots, l$ of the packet must differ from the corresponding bit of the destination node of path P. As these earlier bits won't be touched again by the bit-fixing algorithm, the route of the packet remains separate from P.

Assume that there are a total of h active packets for the nodes of P. What is the probability that together they make a total of at least $30n$ steps along path P?

Consider as an individual trial a situation where a given active packet is in some given node on P. With probability at most $1/2$ we get success, meaning that the packet proceeds along an edge on P. At least with probability $1/2$ we get failure, so the packet leaves path P and never returns. When a failure occurs, we move to considering the next active packet.

Hence, each success contributes one transition along P, but each failure removes one packet from consideration. To get $30n$ transitions, the first $30n + h$ trials may have at most h failures.

The desired conditional probability

$$\Pr(T_1(P) \geq 30n \mid H \leq 6n)$$

is therefore the probability that in the repeated trials we get at most $6n$ failures in $36n$ iterations.

Since each success probability is at most $1/2$, we easily see that

$$\Pr(T_1(P) \geq 30n \mid H \leq 6n) \leq \Pr(Z \leq 6n),$$

where $Z \sim \text{Bin}(36n, 1/2)$. By applying the Chernoff bound of Theorem 4.9 [M&U Thm 4.5.2] we get

$$\begin{aligned} \Pr(T_1(P) \geq 30n \mid H \leq 6n) &\leq \Pr(Z \leq (1-2/3)18n) \\ &\leq \exp(-18n(2/3)^2/2) = e^{-4n} \leq 2^{-3n-1}. \end{aligned}$$

Hence,

$$\Pr(T_1(P)) \geq 30n) \leq 2^{-6n} + \Pr(T_1(P) \geq 30n \mid H < 6n) \leq 2^{-3n}.$$

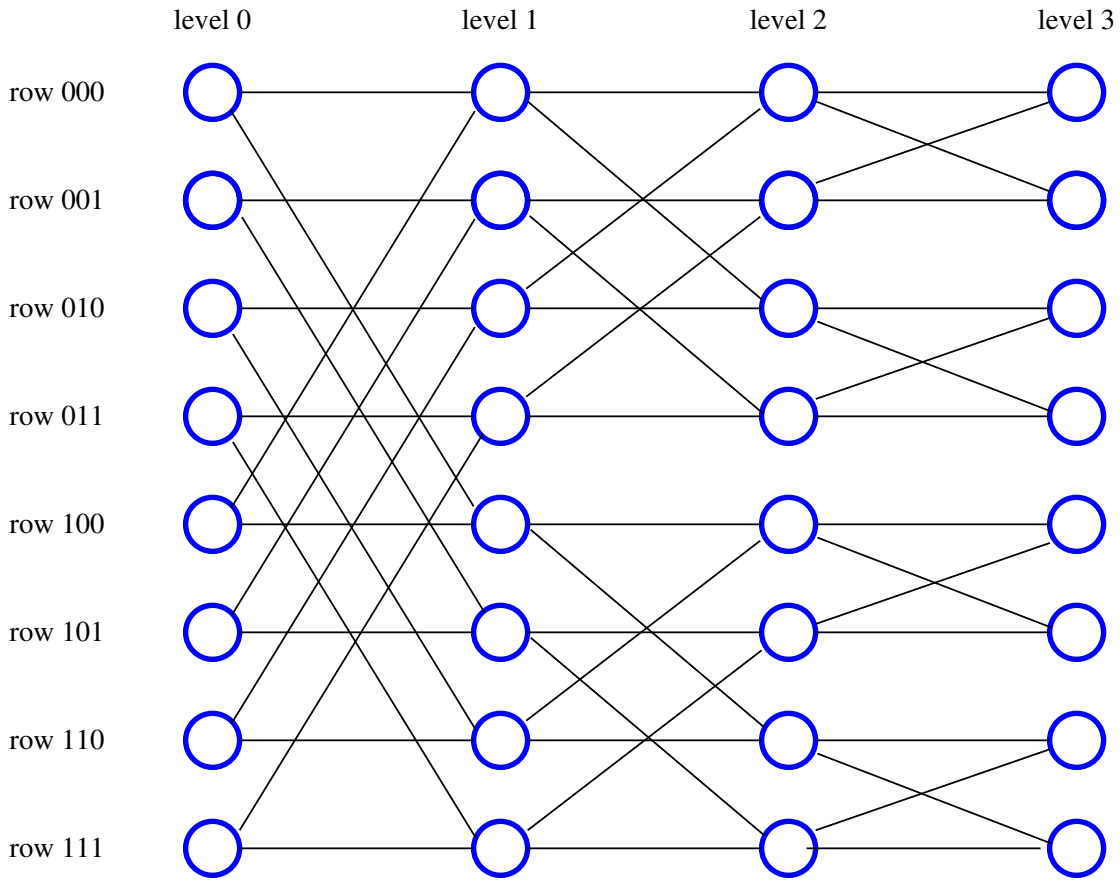
Since there are $N^2 = 2^{2n}$ possible paths, the probability for having $T_1(P) \geq 30n$ for at least one path P is at most $2^{2n}2^{-3n} = 2^{-n}$. Therefore, if Phase II is not started until Phase I is finished, the time for Phase I is $O(\log N)$ with probability $1 - O(N^{-1})$.

The analysis for Phase II is exactly the same. The only difference that the packets actually go along the paths in reverse direction.

Finally, we remark that Phase II can be started even before Phase I finishes. The preceding analysis can easily be extended to show that with probability $1 - O(N^{-1})$ no path has its edges used more than a total of $60n$ times.

Routing in butterfly network [M&U Section 4.5.2]

Butterfly network. In the wrapped version we merge the first and last node on each row.



The butterfly network has $N = n2^n$ nodes for some n. The address of a node has form (x, r), where $0 \leq x \leq 2^n - 1$ is the row and $0 \leq r \leq n - 1$ the column.

In the wrapped butterfly network, nodes (x, r) and (y, s) are connected if $s = (r + 1) \bmod n$ and

1. $x = y$ ("direct" edge) or

2. x and y differ in in the $(s + 1)$th bit position ("flip" edge).

The butterfly network becomes a hypercube if we collapse each row into one "supernode."

Unlike the hypercube, the butterfly network has constant degree and $O(N)$ edges. Hence, if we can get similar routing times as in hypercube, this is in some sense more efficient.

Again we take bit-fixing as the starting point. We use it only to "fix" the row part of the address; getting then to the right column is obvious.

1. Let the start and destination nodes be (x, r) and (y, r), where $x = (a_1, \ldots, a_n)$ and $y = (b_1, \ldots, b_n)$.

2. Repeat for $i = 0, \ldots, n$:

(a) $j := ((r + i) \bmod n) + 1$

(b) If $a_j = b_j$, move to level $j \bmod n$ along the direct edge, otherwise along the flip edge.

The randomized version has three phases. A packet from start node (x, r) to destination (y, s) is routed as follows:

Phase I: pick a random $w \in \{ 0, \ldots, 2^n - 1 \}$ and route to (w, r) by bit-fixing

Phase II: route to (w, s) via direct edges

Phase III: route to (y, s) by bit-fixing.

We will show that with high probability this solves a permutation routing problem in $O(n)$ steps.

Unlike in hypercube, we need to take care with the queue priorities. The rules are as follows:

1. If a packet is in Phase i and has during this phase traversed t edges, it has priority $(i-1)n+t$.

2. If more than one packet wants to use an edge, the lowest priority wins. In case of ties, the packet that arrived first is sent first.

Since the paths in each phase have at most n edges, this means in particular that later phases do not delay the earlier ones. Therefore, we may assume that the phases are executed separately and then add the time requirements.

Consider Phase II first.

Let X_w be the number of packets whose intermediate point is picked from row w. Then $X_w \sim \mathsf{Bin}(n2^n, 2^{-n})$, and Theorem 4.5 [M&U Thm 4.4.1] yields

$$\Pr(X_w \geq 4n) \leq \left(\frac{e^3}{4^4}\right)^n \leq 3^{-2n}.$$

There are 2^n such rows w. The probability that at least one of them gets over $4n$ packets is at most $2^n 3^{-2n} = O(N^{-1})$.

Since Phase II uses only direct edges, all nodes can send packets at least as fast as they arrive. Therefore the queue length does not increase in any node.

In Phase II all packets follow the same paths. This implies that if packet k arrives at node v with priority i, then packets that arrive later cannot pass it in the queue.

Therefore, the total queueing time of a packet is the sum of the queue lengths that nodes have when the packet arrives. Since the queue lengths don't increase, this is at most the same as the total length when Phase II starts, namely X_w.

Therefore, with probability $1 - O(N^{-1})$ no packet spends more time than $4n$ in queue, so the total time is at most $5n$.

For an edge $e = (v_1, v_2)$, let $P(e)$ be the three-edge set including e and the two incoming edges of v_1.

To analyse Phase I, we say that a sequence of edges $e_1, \ldots, e_n$ is a possible delay sequence if for all i we have $e_i \in P(e_{i+1})$. In other words, it's a directed path that may also include "pauses" $e_{i+1} = e_i$.

A possible delay sequence is a delay sequence if among the edges in $P(e_{i+1})$, the edge e_i is among the last ones along which a packet is transmitted with priority at most i.

Consider now an execution of Phase I and there a delay sequence $(e_1, \ldots, e_n)$.

Let t_i be the number of packets that traverse the edge e_i with priority i. Let T_i be the time at which e_i delivers the last packet with priority at most i. In particular, Phase I for edge e_n ends at time T_n.

The definitions imply that at time T_i, the edge e_{i+1} has already

- transmitted all packets with priority i and
- taken into its queue all packets it will transmit with priority $i+1$.

Thus,

$$T_{i+1} \leq T_i + t_{i+1},$$

and because $T_1 = t_1$, we get by induction

$$T_n \leq \sum_{i=1}^{n} t_i.$$

Suppose now the time taken by Phase I is T, and in particular e is one of the edges that transmitted its last packet at time T.

We can recursively define a delay sequence that ends in e:

- $e_n = e$ and
- e_{i-1} is the edge in $P(e_i)$ that last transmitted a packet with priority at most $i - 1$.

Based on the previous slide,

$$\sum_{i=1}^{n} t_i \geq T.$$

Thus, if we have an upper bound for $\sum_{i=1}^{n} t_i$ for all delay sequences, this is also an upper bound for time taken by Phase I.

Given a possible delay sequence $e_1, \ldots, e_n$, let t_i be the number of packets that traverse edge e_i with priority i, and $T = \sum_{i=1}^{n} t_i$.

Based on the above, if $T \leq 40n$ for all $(e_1, \ldots, e_n)$, then Phase I takes at most $40n$ steps. We show that this holds with high probability.

Consider first some fixed $(e_1, \ldots, e_n)$ that is a possible delay sequence. The packets traversing $e_i = (v, v')$ with priority i have started from nodes at exactly distance i from v. There are 2^i such nodes. When a packet in Phase I starts towards the intermediate point from one of these nodes, it will go through e_i with probability 2^{-i-1}. Therefore,

$$\mathbf{E}[t_i] = 2^i 2^{-i-1} = \frac{1}{2} \quad \text{ja} \quad \mathbf{E}[T] = \frac{n}{2}.$$

For $j = 1, \ldots, N$, let $H_j = 1$ if the packet that started from node j contributes to the sum T at least once. Otherwise $H_j = 0$.

The random variables H_j are mutually independent,

$$H = \sum_{j=1}^{N} H_j \leq T \quad \text{and} \quad \mathbf{E}[H] \leq \mathbf{E}[T] = \frac{n}{2}.$$

The Chernoff bound of Theorem 4.7 [M&U Thm 4.4.3] yields

$$\Pr(H \geq 5n) \leq 2^{-5n}.$$

Consider now how much larger T can be compared to H. In other words, how many times can a single packet be counted.

Let u be a packet that gets counted in term t_i. We consider two cases:

- $e_{i+1} = e_i$: Because the movement of packet j with priority $i+1$ takes place in the next column, j is not counted in t_{i+i}. Similarly we see that it won't be counted in any of the terms t_j, $j > i$.

- $e_{i+1} \neq e_i$: The probability that u also traverses e_{i+1} is at most $1/2$. If u does not traverse e_{i+1}, it won't enter this path later, either.

Therefore, when u has been counted for the first time, getting counted for additional times requires success in a trial with probability $1/2$.

The rest goes as with the hypercube.

For $5n$ packets entering the path to produce a total of at least $40n$ edge traversals, we need at most $5n$ failures in $40n$ attempts, when failure probability is at least $1/2$. Chernoff bounds give

$$\Pr(T \geq 40n \mid H \leq 5n) \leq \exp(-20n(3/4)^2/2) \leq 2^{-5n}.$$

We conclude that

$$\Pr(T \geq 40n) \leq \Pr(T \geq 40n \mid H \leq 5n) + \Pr(H \geq 5n) \leq 2^{-5n+1}.$$

This holds for a fixed possible delay sequence. There are at most $2N3^{n-1} \leq n2^n3^n$ possible delay sequences. Hence, the probability that some delay sequence has $T \geq 40n$ is at most

$$n2^n3^n2^{-5n+1} = O(N^{-1}).$$

Therefore, with probability $1 - O(N^{-1})$ all delay sequences in Phase I have $T \leq 40n$, impying that the whole phase takes no longer than $40n$ steps.

Phase III is completely symmetrical with Phase I, and as we noted, our priority rule guarantees that the later phases won't interfer with the ealier ones.

Since also Phase II goes in time $O(n)$ with probability $1 - O(N^{-1})$, so does the whole algorithm. $\square$

5. Balls and Bins

We consider placing m balls independently of each other into n bins where each bin has the same probability. In particular, we are interested in the limit where $n \to \infty$ such that the ratio m/n is constant.

Questions:

- How many balls end up in a given bin?
- What is the largest number of balls in a bin?

Applications: data structures (hashing), load balancing.

We simplify calculations by using Poisson approximation.

Birthday Paradox [M&U Section 5.1]

In a group of 30 people, what's the probability that at least two persons have the same birthday? Thus, we consider $m = 30$ and $n = 365$. (Obviously in reality this problem does not satisfy the distribution assumptions of the balls and bins model.)

If $k-1$ birthdays have already been chosen, the probability that the kth one does match any of these is $1-(k-1)/365$. Therefore, the probability of getting 30 different birthdays is

$$\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)\ldots\left(1-\frac{29}{365}\right) \approx 0.2937.$$

Hence, there's about a 70% probability of at least one match.

More generally, with m people and n possible birthdays, the probability of different birthdays is

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{3}{n}\right)\ldots\left(1-\frac{m-1}{n}\right)=\prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right).$$

Since $\lim_{n\to\infty}(1-x/n)^n=\mathrm{e}^{-x}$, we can approximate

$$1-\frac{j}{n}\approx\mathrm{e}^{-j/n}$$

which makes the above approximately

$$\prod_{j=1}^{m-1}\mathrm{e}^{-j/n}=\exp\left(-\sum_{j=1}^{m-1}\frac{j}{n}\right)=\exp\left(-\frac{m(m-1)}{2n}\right)\approx\exp\left(-\frac{m^2}{2n}\right).$$

For example, if we ask how many people are needed to have at least a probability 1/2 that the birthdays are not all different, we get the equation

$$\frac{m^2}{2n}=\ln 2$$

giving $m=\sqrt{2n\ln 2}$. For example for $n=365$ this approximation gives $m=22.49$, which is reasonably close.

We can make this more precise by replacing the approximations by suitable upper and lower bounds. Next we see one fairly crude estimate.

Let E_j be the event that person j does not share birthdays with any of the persons $1, \ldots, j-1$. The probability that k persons do not all have different birthdays is

$$\Pr(\overline{E_1} \cup \ldots \cup \overline{E_k}) \leq \sum_{j=1}^{k} \Pr(\overline{E_j}) \leq \sum_{j=1}^{k} \frac{j-1}{n} = \frac{k(k-1)}{2n}.$$

For $k \leq \sqrt{n}$ this is at most $1/2$, so $\lfloor \sqrt{n} \rfloor$ people have different birthdays with probability at least $1/2$.

On the other hand, assume we have at least $2\lceil\sqrt{n}\rceil$ people. If all their birthdays are different, then both of the following events have occurred:

1. persons $1, \ldots, \lceil\sqrt{n}\rceil$ have different birthdays and

2. persons $\lceil\sqrt{n}\rceil + 1, \ldots, 2\lceil\sqrt{n}\rceil$ have different birthdays from any of persons $1, \ldots, \lceil\sqrt{n}\rceil$.

On the condition that the first event occurred, the probability that the second one occurs is

$$\left(1 - \frac{\lceil\sqrt{n}\rceil}{n}\right)^{\lceil\sqrt{n}\rceil} \leq \left(1 - \frac{1}{\sqrt{n}}\right)^{\sqrt{n}} < \frac{1}{\mathrm{e}}.$$

Hence, the probability for all the birthdays being different is at most $1/\mathrm{e} \approx 0.368$.

Maximum load [M&U Section 5.2.1]

Consider now the maximum load, that is, the largest number of balls in any bin.

Based on the previous, having $m = \Omega(\sqrt{n})$ balls is sufficient for the maximum load to be probably at least 2.

We derive an upper bound for the special case $m = n$: with high probability no bin gets more than $3 \ln n / \ln \ln n$ balls.

The bound $O(\ln n / \ln \ln n)$ is actually tight, but the constant 3 is not the best possible.

The probability for bin 1 receiving at least M balls is at most

$$\binom{n}{M}\left(\frac{1}{n}\right)^M.$$

We use the bounds

$$\binom{n}{M}\left(\frac{1}{n}\right)^M \leq \frac{1}{M!} \leq \left(\frac{\mathsf{e}}{M}\right)^M.$$

The first part follows directly from the definition of the binomial coefficient, the second part from the fact that

$$\frac{M^M}{M!} \leq \sum_{i=0}^{\infty} \frac{M^i}{i!} = \mathsf{e}^M.$$

Assume $M \geq 3 \ln n / \ln \ln n$. The probability that at least one bin receives more than M balls is upper bounded by

$$\begin{aligned} n \left(\frac{e}{M}\right)^M &\leq n \left(\frac{e \ln \ln n}{3 \ln n}\right)^{3 \ln n / \ln \ln n} \\ &\leq n \left(\frac{\ln \ln n}{\ln n}\right)^{3 \ln n / \ln \ln n} \\ &= \exp(\ln n) \exp(\ln \ln \ln n - \ln \ln n)^{3 \ln n / \ln \ln n} \\ &= \exp(-2 \ln n + 3(\ln n)(\ln \ln \ln n)/ \ln \ln n) \\ &\leq \frac{1}{n} \end{aligned}$$

for large n. ($\square$)

Bucket Sort [M&U Section 5.2.2]

We have to sort $n = 2^m$ elements that are drawn uniformly from $\{0, \ldots, 2^k - 1\}$, where $k \geq m$.

Algorithm:

1. Create n buckets (for example, linked lists). Element $a \in \{0, \ldots, 2^k - 1\}$ goes into bucket j, where j is obtained by considering the binary representation of a and taking only the first m bits.

2. Sort each bucket (for example, by insertion sort).

Let X_j be the number of elements that go to bucket j. This random variable depends on the random input. The algorithm itself is deterministic.

The time complexity of the algorithm is at most

$$an + b\sum_{j=1}^{n} X_j^2$$

for some constants a, b. The random variables X_j have identical distributions, so the expectation of this is

$$an + bn\mathbf{E}[X_1^2].$$

Since the distribution of X_1 is $\mathrm{Bin}(n, 1/n)$, we know that

$$\mathbf{E}[X_1^2] = \mathbf{Var}[X_1] + \mathbf{E}[X_1]^2 = n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + 1 = 2 - \frac{1}{n} < 2.$$

Hence, the average running time is at most

$$an + 2bn = O(n).$$

Poisson distribution [M&U Section 5.3]

We will here use the Poisson distribution mainly as a tool for making calculations easier. Nevertheless, let us briefly consider its basic application.

Consider a service that has a very large number of customers that are independent and identical (from the service's point of view). There is some fixed probability that during a fixed time period we are interested in, a given customer will need service. This probability is the same for all customers.

Let X be the number of service requests during the time period. If the expected number of requests per time unit is μ, then X is a discrete Poisson random variable with parameter μ. We denote this by $X \sim \text{Poisson}(\mu)$, and

$$\Pr(X = j) = \frac{\mathrm{e}^{-\mu}\mu^j}{j!}.$$

The expectation is $\mathbf{E}[X] = \mu$ (as desired).

From balls and bins point of view, balls are customers and there is a number of services, each represented by a bin. For example, ball a going into bin b might mean that process a needs to access memory location b during the time interval under consideration. Let us see how this is connected to the Poisson distribution.

Consider first the bins that remain empty. The probability for a given bin to remain empty is

$$\left(1 - \frac{1}{n}\right)^m \approx e^{-m/n}.$$

If X is the number of empty bins, then

$$\mathbf{E}[X] = n\left(1 - \frac{1}{n}\right)^m \approx ne^{-m/n}.$$

More generally, the probability that the bin receives exactly r balls is

$$\binom{m}{r}\left(\frac{1}{n}\right)^r\left(1 - \frac{1}{n}\right)^{m-r} = \frac{1}{r!} \cdot \frac{m(m-1)\ldots(m-r+1)}{n^r}\left(1 - \frac{1}{n}\right)^{m-r}.$$

For $r \ll m$ we can estimate

$$m(m-1)\dots(m-r+1) \approx m^r.$$

For large m and n, we have $(1-1/n)^m \approx \mathrm{e}^{-m/n}$. Therefore,

$$\binom{m}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{m-r} \approx \frac{1}{r!}\cdot\frac{m^r}{n^r}\cdot\mathrm{e}^{-m/n} = \frac{\mathrm{e}^{-\mu}\mu^r}{r!}$$

where $\mu = m/n$. Hence, when m and n are large and r small in comparison, the probability for getting r balls into a given bin is roughly as in Poisson(m/n) distribution.

We can also see that the expected number m/n of balls in a bin is the same as the expectation of the approximating Poisson distribution.

We now make this more precise and general.

Theorem 5.1: Let $X_n \sim \mathrm{Bin}(n, p_n)$ where $\lim_{n\to\infty} np_n = \lambda$ is a constant. For any fixed k we have

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}.$$

Thus, rare events can be approximately modelled by a Poisson distribution.

Proof: To simplify notation, write $p = p_n$ and keep in mind that p is a function of n. We want to estimate the quantity

$$\Pr(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We use bounds that hold for $|x| \le 1$ and can be obtained by simple calculus:

$$\mathrm{e}^{-x}(1-x^2) \le 1 - x \le \mathrm{e}^{-x}.$$

We start with the upper bound:

$$\begin{aligned}\Pr(X_n = k) &\leq \frac{n^k}{k!} \cdot p^k \cdot \frac{(1-p)^n}{(1-p)^k} \\ &\leq \frac{(np)^k}{k!} \cdot \frac{\mathrm{e}^{-pn}}{1-pk}\end{aligned}$$

where we used $1 - p \leq \mathrm{e}^{-p}$ and $(1-p)^k \geq 1 - pk$.

Similarly, using $1 - p \geq \mathrm{e}^{-p}(1-p^2)$ gives a lower bound:

$$\begin{aligned}\Pr(X_n = k) &\geq \frac{(n-k+1)^k}{k!} \cdot p^k (1-p)^n \\ &\geq \frac{((n-k+1)p)^k}{k!} (\mathrm{e}^{-p}(1-p^2))^n \\ &\geq \frac{\mathrm{e}^{-pn}((n-k+1)p)^k}{k!} (1-np^2).\end{aligned}$$

Therefore,

$$\frac{\mathrm{e}^{-pn}((n-k+1)p)^k}{k!}(1-np^2) \leq \Pr(X_n = k) \leq \frac{\mathrm{e}^{-pn}(np)^k}{k!}\frac{1}{1-pk}.$$

Consider now the limit $n \to \infty$, keeping in mind that $np \to \lambda$ and therefore $p \to 0$. The limit for the lower bound is

$$\lim_{n\to\infty} \frac{\mathrm{e}^{-pn}((n-k+1)p)^k}{k!}(1-np^2) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}$$

and for the upper bound

$$\lim_{n\to\infty} \frac{\mathrm{e}^{-pn}(np)^k}{k!}\frac{1}{1-pk} = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}.$$

Since $\Pr(X_n = k)$ is between these two bounds, the claim follows. □

Consider now the moment generating function for Poisson distribution.

Lemma 5.2 [M&U Lemma 5.3]: When $X \sim \text{Poisson}(\mu)$, we have

$$M_X(t) = \exp(\mu(e^t - 1)).$$

Proof:

$$\mathbf{E}[\mathrm{e}^{tX}] = \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\mathrm{e}^{tk} = \mathrm{e}^{-\mu}\sum_{k=0}^{\infty}\frac{(\mu\mathrm{e}^t)^k}{k!} = \mathrm{e}^{-\mu}\exp(\mu\mathrm{e}^t).$$

□

Corollary 5.3: When $X \sim \text{Poisson}(\mu)$, we have $\mathbf{E}[X] = \mu$ and $\mathbf{Var}[X] = \mu$. □

Corollary 5.4: When $X \sim \text{Poisson}(\mu)$ and $Y \sim \text{Poisson}(\lambda)$ are independent, we have $X + Y \sim \text{Poisson}(\mu + \lambda)$.

Proof: $M_{X+Y}(t) = M_X(t)M_Y(t) = \exp((\mu + \lambda)(\mathrm{e}^t - 1))$. □

We can use this to obtain Chernoff-type bounds.

Theorem 5.5 [M&U Thm 5.4]: Let $X \sim \text{Poisson}(\mu)$. Then

1. for $x > \mu$ we have
$$\Pr(X \geq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$$

2. for $x < \mu$ we have
$$\Pr(X \leq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}.$$

We can write this in a more familiar form

$$\begin{aligned} \Pr(X \geq (1+\delta)\mu) &\leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \\ \Pr(X \leq (1-\delta)\mu) &\leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}. \end{aligned}$$

Proof: As we did earlier, notice that for positive t we have

$$\Pr(X \geq x) = \Pr(e^{tX} \geq e^{tx}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{tx}}.$$

Therefore,

$$\Pr(X \geq x) \leq \exp(\mu(e^t - 1) - tx).$$

The desired bound follows by choosing $t = \ln(x/\mu)$.

The case $\Pr(X \leq x)$ is similar. □

Poisson approximation [M&U Section 5.4]

Fix the number of bins n. Let $X_i^{(m)}$ be the number of balls received by bin i when there are m balls. On the other hand, let $Y_i^{(m)}$, $i = 1, \ldots, n$ be independent random variables with Poisson(m/n) distribution.

We wish to simplify our analyses by approximating the X variables by the Y variables. We have already observed that for large n and m, each *individual* $X_i^{(m)}$ has approximately the same distribution as $Y_i^{(m)}$. This is not enough, because $X_i^{(m)}$ are *not* mutually independent.

We are going to show that events that are rare in the Poisson model (that is, when considering $(Y_i^{(m)})_i$) are rare also in the exact model (that is, when considerings $(X_i^{(m)})_i$). This is usually the direction we are interested in.

Example: We saw earlier that the event $\max_i X_i^{(m)} > 3 \ln n / \ln \ln n$ is "rare" for $m = n$. We will soon see how to analyse this in the Poisson model.

Fundamentally the dependencies among $(X_i^{(m)})_i$ are caused by the fact that always

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

This of course is not true for $(Y_i^{(m)})_i$. In some sense, this is the only difference between the exact and Poisson models.

Theorem 5.6 [M&U Thm 5.6]**:** The distribution of the random variables $(Y_1^{(m)}, \ldots, Y_n^{(m)})$ subject to the condition $\sum_{i=1}^{n} Y_i^{(m)} = k$ is the same as the distribution of the random variables $(X_1^{(k)}, \ldots, X_n^{(k)})$.

Proof: Pick arbitrary $k \in \mathbb{N}$ and $(k_1, \ldots, k_n) \in \mathbb{N}^n$. We want to show

$$\begin{aligned} &\Pr((X_1^{(k)}, \ldots, X_n^{(k)}) = (k_1, \ldots, k_n)) \\ &\quad = \Pr\left((Y_1^{(m)}, \ldots, Y_n^{(m)}) = (k_1, \ldots, k_n) \mid \sum_{i=1}^{n} Y_i^{(m)} = k \right). \end{aligned}$$

It is sufficient to consider the case $k = \sum_{i=1}^{n} k_i$. Otherwise both sides are clearly zero.

The number of ways to partition k balls into n classes so that class i contains k_i balls is

$$\binom{k}{k_1 \ \ldots \ k_n} = \frac{k!}{k_1! \ldots k_n!}.$$

After a partitioning has been fixed, the probability that for all i, all balls assigned to class i actually go into bin i is $(1/n)^k$.

Therefore,

$$\Pr((X_1^{(k)}, \ldots, X_n^{(k)}) = (k_1, \ldots, k_n)) = \frac{k!}{(k_1)! \ldots (k_n)! n^k}.$$

The random variables $Y_i^{(m)}$ are independent and Poisson(m/n) distributed. Therefore, their sum has distribution Poisson(m) (Corollary 5.4). Hence,

$$\begin{aligned}
&\Pr\left((Y_1^{(m)},\ldots,Y_n^{(m)}) = (k_1,\ldots,k_n) \mid \sum_{i=1}^{n} Y_i^{(m)} = k\right) \\
&= \frac{\prod_{i=1}^{n} \Pr(Y_i^{(m)} = k_i)}{\Pr(\sum_{i=1}^{n} Y_i^{(m)} = k)} \\
&= \frac{\prod_{i=1}^{n} \mathrm{e}^{-m/n}(m/n)^{k_i}/k_i!}{\mathrm{e}^{-m}m^k/k!} \\
&= \frac{k!}{(k_1)!\ldots(k_n)!n^k}.
\end{aligned}$$

□

Since we have a reasonably large probability of getting $\sum_i Y_i^{(m)} = m$, we obtain

Theorem 5.7: Let $f : \mathbb{N}^n \to [0, \infty)$ be arbitrary. Then

$$\mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \leq \mathrm{e}\sqrt{m}\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})].$$

Before the proof we note an important implication:

Corollary 5.8: If the probability of an even in the Poisson model is at most p, then its probability in the exact model is at most $\mathrm{e}\sqrt{m}p$.

Proof: Choose the indicator function of the event as f. □

In particular, if we want to upper bound some probability in the exact model as $O(1/m^r)$, it's sufficient to show a upper bound $O(1/m^{r+1/2})$ in the Poisson approximation.

Proof of Theorem 5.7: We start from the estimate

$$\begin{aligned}\mathbf{E}[f(Y_1^{(m)},\ldots,Y_n^{(m)})] &= \sum_{k=0}^{\infty}\mathbf{E}[f(Y_1^{(m)},\ldots,Y_n^{(m)}) \mid \sum_{i=1}^{n} Y_i^{(m)} = k]\,\mathsf{Pr}(\sum_{i=1}^{n} Y_i^{(m)} = k)\\ &\geq \mathbf{E}[f(Y_1^{(m)},\ldots,Y_n^{(m)}) \mid \sum_{i=1}^{n} Y_i^{(m)} = m]\,\mathsf{Pr}(\sum_{i=1}^{n} Y_i^{(m)} = m).\end{aligned}$$

By using Theorem 5.6 and the fact that $\sum_{i=1}^{n} Y_i^{(m)} \sim \mathsf{Poisson}(m)$ we get

$$\mathbf{E}[f(Y_1^{(m)},\ldots,Y_n^{(m)})] \geq \mathbf{E}[f(X_1^{(m)},\ldots,X_n^{(m)})]\frac{\mathsf{e}^{-m}m^m}{m!}.$$

The claim follows using the estimate

$$m! \leq \mathsf{e}\sqrt{m}\left(\frac{m}{\mathsf{e}}\right)^m$$

which will be proved next. □

Lemma 5.9:

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

Proof: First we write

$$\ln n! = \sum_{i=1}^{n} \ln i.$$

Since ln is concave, for $i \geq 2$ we have

$$\int_{i-1}^{i} \ln x \, dx \geq \frac{1}{2}\left(\ln(i-1) + \ln i\right).$$

Therefore,

$$\int_{1}^{n} \ln x \, dx \geq \sum_{i=2}^{n} \frac{1}{2}\left(\ln(i-1) + \ln i\right) = \sum_{i=1}^{n} \ln i - \frac{1}{2}\ln n$$

implying (because $\int \ln x \, dx = x \ln x - x$)

$$n \ln n - n + 1 \geq \ln n! - \frac{1}{2}\ln n.$$

The claim follows by taking exponentials. □

In a common special case we get a tighter bound.

Theorem 5.10: Let f be non-negative and such that $\mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})]$ is monotone in m. Then

$$\mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \leq 2\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})].$$

Proof: Omitted. □

Corollary 5.11: Consider an event, the probability of which is monotone in m. If the probability of the event is at most p in the Poisson model, then its probability is at most $2p$ in the exact model. □

Example: The probability of the event

$$\max_i X_i^{(m)} \geq r$$

is clearly increasing and the probability of the event

$$\max_i X_i^{(m)} \leq r$$

decreasing as function of m. Thus, we can apply the above bound. □

As an example, consider the maximum load problem.

Let $m = n$ and $M = \ln n / \ln \ln n$. We saw earlier that the probability of at least one bin receiving more than $3M$ balls is $O(1/n)$. We show that on the other hand, with probability $1 - O(1/n)$ some bin receives at least M balls.

Thus, the parameter in the Poisson model is $m/n = 1$, and

$$\Pr(Y_i^{(m)} \geq M) \geq \Pr(Y_i^{(m)} = M) = \frac{1}{\mathrm{e}M!}.$$

Since $Y_i^{(m)}$ are mutually independent,

$$\Pr(Y_i^{(m)} < M \text{ for all } i) \leq \left(1 - \frac{1}{\mathrm{e}M!}\right)^n \leq \exp\left(-\frac{n}{\mathrm{e}M!}\right).$$

If now $\exp(-n/(\mathrm{e}M!)) = O(n^{-2})$, the claim follows.

Thus, it suffices to show $\exp(-n/(\mathrm{e}M!)) \le n^{-2}$. Write this as $M! \le n/2\mathrm{e}\ln n$, and further as

$$\ln M! \le \ln n - \ln\ln n - \ln(2\mathrm{e}).$$

By Lemma 5.9,

$$M! \le \mathrm{e}\sqrt{M}\left(\frac{M}{\mathrm{e}}\right)^M \le M\left(\frac{M}{\mathrm{e}}\right)^M.$$

Therefore, for large n and using $\ln\ln n = o(\ln n/\ln\ln n))$ we get

$$\begin{aligned}
\ln M! &\le M\ln M - M + \ln M \\
&= \frac{\ln n}{\ln\ln n}(\ln\ln n - \ln\ln\ln n) - \frac{\ln n}{\ln\ln n} + (\ln\ln n - \ln\ln\ln n) \\
&\le \ln n - \frac{\ln n}{\ln\ln n} \\
&\le \ln n - \ln\ln n - \ln(2\mathrm{e})
\end{aligned}$$

as desired. (□)

Example: more coupon collecting

Think about coupon collecting in the balls and bins framework. Cereal boxes are balls and different coupons are bins. We ask how many balls are needed to get at least one ball in every bin.

We saw earlier that

- the expected number of balls is $n \ln n + \Theta(n)$
- probability that $n \ln n + cn$ balls is not enough is at most e^{-c}.

We can now get a much more accurate estimate.

Theorem 5.12: Let X be the number of balls required to have at least one ball in every bin. For any constant c we have

$$\lim_{n\to\infty} \Pr(X > n \ln n + cn) = 1 - \exp(-e^{-c}).$$

By considering for example $c = -4$ and $c = 4$ we see that $\Pr(|X - n \ln n| \leq 4n) \approx 0.98$ (for large n).

Proof: Consider first the Poisson model. Choose $m = n \ln n + cn$, so the parameter for each Poisson random variable will be $m/n = \ln n + c$. The probability of a given bin remaing empty is

$$e^{-m/n}\frac{(m/n)^0}{0!} = e^{-\ln n - c} = \frac{e^{-c}}{n}.$$

Since the bins in the Poisson model are independent, the probability that all are non-empty is

$$\left(1 - \frac{e^{-c}}{n}\right)^n \to \exp(-e^{-c})$$

kun $n \to \infty$.

Thus, the claimed result holds in the Poisson model. Since we want a more accurate bound than in previous examples, conversion to the exact model takes a bit more work.

Let $\mathcal{E}$ be the event (still in the Poisson model) that no bin is empty, and X the number of balls we used. (Thus, $X \sim \text{Poisson}(m)$.) We just observed that

$$\lim_{n\to\infty} \Pr(\mathcal{E}) = \exp(-e^{-c}).$$

To get the desired bound in the exact model, to apply Theorem 5.6 we need

$$\lim_{n\to\infty} \Pr(\mathcal{E} \mid X = m) = \exp(-e^{-c}).$$

Therefore, it remains to show that

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E} \mid X = m) + o(1)$$

where $o(1) \to 0$ when $n \to \infty$.

Let A be the event

$$|X - m| \leq \sqrt{2m \ln m}.$$

Remember that $m = \mathbf{E}[X]$, so in some sense A consists of "usual" cases.

We will derive two estimates:

$$\begin{aligned} \Pr(\overline{A}) &= o(1) \\ \Pr(\mathcal{E} \mid A) &= \Pr(\mathcal{E} \mid X = m) + o(1) \end{aligned}$$

These will imply the desired result:

$$\begin{aligned} \Pr(\mathcal{E}) &= \Pr(\mathcal{E} \mid A) \Pr(A) + \Pr(\mathcal{E} \mid \overline{A}) \Pr(\overline{A}) \\ &= \Pr(\mathcal{E} \mid A)(1 - o(1)) + o(1) \\ &= \Pr(\mathcal{E} \mid X = m) + o(1). \end{aligned}$$

Because $X \sim \text{Poisson}(m)$, we can estimate the probability of $\overline{A}$ using a Chernoff bound (Theorem 5.5)

$$\Pr(X \geq x) \leq \exp(x - m - x \ln(x/m)).$$

We choose $x = m + \sqrt{2m \ln m}$ and estimate $\ln(1+z) \geq z - z^2/2$ (for $z \geq 0$):

$$\begin{aligned}
&\Pr(X \geq m + \sqrt{2m \ln m}) \\
&\quad \leq \exp\left(\sqrt{2m \ln m} - (m + \sqrt{2m \ln m}) \ln\left(1 + \sqrt{(2 \ln m)/m}\right)\right) \\
&\quad \leq \exp\left(\sqrt{2m \ln m} - \left(m + \sqrt{2m \ln m}\right)\left(\sqrt{(2 \ln m)/m} - (\ln m)/m\right)\right) \\
&\quad = \exp\left(-\ln m + \sqrt{2m \ln m}(\ln m)/m\right) \\
&\quad = o(1).
\end{aligned}$$

Similarly we see that $\Pr(X \leq m - \sqrt{2m \ln m}) = o(1)$, so

$$\Pr(\overline{A}) = \Pr(X \geq m + \sqrt{2m \ln m}) + \Pr(X \leq m - \sqrt{2m \ln m}) = o(1).$$

We still need to show that

$$\Pr(\mathcal{E} \mid A) = \Pr(\mathcal{E} \mid X = m) + o(1)$$

where again A is the event

$$m - \sqrt{2m \ln m} \leq X \leq m + \sqrt{2m \ln m}.$$

Write

$$\begin{aligned} a &= \Pr(\mathcal{E} \mid X = m - \sqrt{2m \ln m}) \\ b &= \Pr(\mathcal{E} \mid X = m + \sqrt{2m \ln m}). \end{aligned}$$

Clearly $\Pr(\mathcal{E} \mid X = k)$ is increasing in k, so

$$a \leq \Pr(\mathcal{E} \mid A) \leq b \quad \text{and} \quad a \leq \Pr(\mathcal{E} \mid X = m) \leq b.$$

Therefore

$$|\Pr(\mathcal{E} \mid X = m) - \Pr(\mathcal{E} \mid A)| \leq b - a.$$

We will show that $b - a = o(1)$.

The difference

$$b - a = \mathsf{Pr}(\mathcal{E} \mid X = m + \sqrt{2m \ln m}) - \mathsf{Pr}(\mathcal{E} \mid X = m - \sqrt{2m \ln m})$$

is the probability that the first $m - \sqrt{2m \ln m}$ balls leave at least one bin empty, but the next $2\sqrt{2m \ln m}$ fill all the gaps.

The probability of hitting a given empty bin at least once with $2\sqrt{2m \ln m}$ balls is at most $2\sqrt{2m \ln m}/n = o(1)$. The probability of hitting all the empty bins is certainly no larger than this. □

Set membership problems and hashing [M&U Section 5.5]

We consider accessing a set $S \subseteq U$ where the universe U is very large compared to available memory space. We want to be able efficiently decide whether an element $x \in U$ given as input belongs to S. We do not here consider updating S.

Let $S = \{ s_1, \ldots, s_m \}$. Known solution methods include

binary search: assumes U is ordered, time complexity $\Theta(\log m)$

hashing with overflow chains: for a "good" hash function, time complexity $O(1)$ *on the average* .

Finding good hash functions is not trivial. Even for a good hash function if the storage area has size m, the longest overflow chain, and therefore the worst-case time complexity, is $O(\log m / \log \log m)$ with high probability (pp. 147–148).

Having constant time complexity (with high probability) would of course be very desirable, but the memory requirement is also important. Here we consider trading off between memory and time requirements.

We consider approximate randomized algorithms that can with a small probability give false positives, meaning the algorithm answers yes even though $x \notin S$.

Consider a hash function $f : U \to \{0, \ldots, n-1\}$ where n is the size of the hash table. The simplest solution is to forget about overlow chains and just store in each bin of the hash table a single bit which tells whether the set contains at least one elements hashing to that address. Then false positives may be caused by collisions, when $x \notin S$ but $f(x) = f(y)$ for some $y \in S$.

We call $f(x)$ the fingerprint of x. Thus, collisions may happen when two elements have the same fingerprint.

For any given f we can always find a case where it performs poorly, so we analyse choosing f randomly.

Assume for simplicity that the function $f\colon U \to \{0, \ldots, n-1\}$ is chosen at random so that for each $x \in U$ independently of the other ones, $f(x)$ is chosen uniformly from $\{0, \ldots, n-1\}$.

This assumption as such is **not** reasonable in practice. Just to represent a random f would take $|U| \log n$ bits. We omit here the question

- what are realistic assumptions about a random hash function
- how can the assumptions be satisfied in an efficient manner.

(See for example universal hashing).

In the fingerprinting solution we thus store the fingerprints of all the elements of S. If $n = 2^b$, this takes mb bits.

If $x \notin S$, then our assumptions give $f(x) = f(s_i)$ with probability $1/n$ for all i independently of each other. Therefore, the probability of a false positive is

$$1 - \left(1 - \frac{1}{2^b}\right)^m \geq 1 - \exp\left(-\frac{m}{2^b}\right).$$

To make this less than a given value c, we require

$$b \geq \log_2 \frac{m}{\ln(1/(1-c))} = \Omega(\log m)$$

bits per fingerprint. On the other hand, by choosing for example $b = 2\log_2 m$ we get an upper bound for false positives

$$1 - \left(1 - \frac{1}{m^2}\right)^m \leq \frac{1}{m}.$$

Bloom filter [M&U luku 5.5.3]

We now pick k mutually independent hash functions $h_1, \ldots, h_k$ that are as above functions $U \to \{\, 0, \ldots, n-1 \,\}$. We use a single n bit array $A[0 \ldots n]$ as storage area.

When s is inserted to S, we set to one all the bits $A[h_i(s)]$, $i = 1, \ldots, k$.

Accordingly, the query "$x \in S$?" is answered *"yes"* if $A[h_i(x)] = 1$ for all $i = 1, \ldots, k$.

Clearly there are no false negatives (if $x \in S$ the answer is always "yes").

To get a false positive we must have set all the i bits $A[h_i(x)]$ to one because of the elements in S. We analyse the probability of this.

The case $k = 1$ is our original hash table solution. Increasing k has two opposite effects:

- more checks for each query tends to decrease the probability of false a positive
- more ones in the array tend to increase the probability of false a positive.

When storing m elements, the probability of a given bit to remain zero is

$$\left(1 - \frac{1}{n}\right)^{km} \approx \exp\left(-\frac{km}{n}\right).$$

Consider first for simplicity the case of exactly pn zeros in the array, where $p = \exp(-km/n)$. The probability of a false positive is then

$$(1 - p)^k = (1 - \exp(-km/n))^k.$$

We consider this as a function of k.

Write

$$f(k) = (1 - \exp(-km/n))^k$$

and

$$g(k) = \ln f(k) = k \ln (1 - \exp(-km/n)) .$$

Taking derivatives gives

$$\begin{aligned} g'(k) &= \ln (1 - \exp(-km/n)) + k \cdot \frac{m}{n} \cdot \frac{\exp(-km/n)}{1 - \exp(-km/n)} \\ &= \ln(1-p) - (\ln p)\frac{p}{1-p}. \end{aligned}$$

From this we see easily that $g(k)$ and hence $f(k)$ get their smallest values for $p = 1/2$, that is, $k = (\ln 2) \cdot (n/m)$.

Possible interpretation: For $p = 1/2$ the content of the array has maximum entropy, meaning the information content is maximised.

By plugging in $k = (\ln 2) \cdot (n/m)$ we get

$$f(k) = \left(\frac{1}{2}\right)^k = \exp\left(-(\ln 2)^2 \frac{n}{m}\right) \approx 0{,}6185^{n/m}.$$

Therefore, for a constant error probability $f(k) = c$ it is enough to have a constant number

$$\frac{n}{m} = \frac{\ln(1/c)}{(\ln 2)^2}$$

of bits per element to be inserted.

Now go back to the assumption about the number of bits set to one. The expected proportion of one-bits is

$$p' = \left(1 - \frac{1}{n}\right)^{km},$$

and we have a usual balls and bins situation.

Thus, even though the bits are not independent, we can apply the Poisson approximation (Corollary 5.8, [M&U Cor 5.9]) and there the Chernoff bound (Corollary 4.10, [M&U Cor 4.6]). Denoting by X the number of one-bits, we get

$$\Pr(\left|X - np'\right| \geq \varepsilon n) \leq 2e\sqrt{n}\exp(-n\varepsilon^2/(3p')).$$

With high probability, the assumption about the number of one-bits is therefore approximately correct.

Symmetry breaking by hashing [M&U Section 5.5.4]

Suppose we have n customers who want to use some resource. Give the customers identifiers $s_1, \ldots, s_n$. We serve the customers in the order of the values $h(s_i)$ of a random hash function h.

Consider again hash values with b bits. The probability that given two customers get the same hash value is $(1/2)^b$. Since there are $n(n-1)/2$ pairs of customers, the probability that at least one pair shares the same value is at most

$$\frac{n(n-1)}{2^{b+1}}.$$

For $b \geq 3 \log_2 n$, the probability of collision is at most $1/n$.

Random graphs [M&U Section 5.6]

We consider two different models for choosing a random undirected graph $G = (V, E)$ where $|V| = n$.

$G_{n,p}$ For each pair of vertices (u, v) independently of each other we set $(u, v) \in E$ with probability p. Since there are $n(n-1)/2$ pairs of vertices, the probability of getting a given graph that has m edges is

$$p^m (1-p)^{n(n-1)/2-m}.$$

$G_{n,N}$ All graphs that have N edges have the same probability, and this probability is

$$\binom{n(n-1)/2}{N}^{-1}.$$

The relationship between $G_{n,p}$ and $G_{n,N}$ where $N = pn(n-1)/2$ is similar to the the relationship between a Poisson approximation and the exact balls and bins model.

When considering random graphs as inputs for computational problems, we can often find threshold values. We'll get back to this later.

Example If $p \gg n^{-2/3}$, then with a very high probability $G_{n,p}$ includes a clique of 4 vertices. If $p \ll n^{-2/3}$, then with very high probability $G_{n,p}$ does not include a clique of 4 vertices. □

This entails that for inappropriate parameter choices, random graphs are often uninteresting as inputs for graph algorithms. For example, if p is very large, clique finding problems may become trivial because the graph is full of large cliques.

On the other hand, if the parameters have been chosen near the threshold values, random graphs may be interesting difficult test cases for algorithms.

Hamiltonian cycles in random graphs [M&U Section 5.6.2]

We consider an algorithm that gradually builds up a path starting from a vertex. The basic operations for building a path P are

Extension: When $P = (v_1, \ldots, v_k)$ and $(v_k, u) \in E$ where $u \notin \{ v_1, \ldots, v_k \}$, let

$$P := (v_1, \ldots, v_k, u).$$

Rotation: When $P = (v_1, \ldots, v_k)$ and $(v_k, u) \in E$ where $u = v_i \in \{ v_1, \ldots, v_k \}$, let

$$P := (v_1, \ldots, v_{i-1}, v_i, v_k, v_{k-1}, \ldots, v_{i+2}, v_{i+1}).$$

If a rotation found a Hamiltonian cycle ($k = n$ and $u = v_1$), the search is over.

Otherwise if we are in a dead end, the algorithm gives a negative answer.

The first version works as follows:

Initialize P as an empty path and pick any vertex v_1 as the head.
Repeat until a Hamiltonian cycle is found or the head runs out of neighbors:
Write $P = (v_1, \ldots, v_k)$, where v_k is the head.
Let (v_k, u) be the first egde in the adjacency list of v_k.
Remove (v_k, u) from the adjacency lists of v_k and u.
If $u \notin \{ v_1, \ldots, v_k \}$, make an extension.
The new head is u.
If $u = v_i \in \{ v_1, \ldots, v_k \}$, make a rotation
(and if a Hamiltonian cycle was found, return it).
The new head is v_{i+1}.
If no Hamiltonian cycle was found, return *"no"*.

This however is very difficult to analyse because of dependencies that arise between the adjacency lists of different vertices.

For the final version of the algorithm we change the data structure for representing the graph.

- The adjacency list of v is split into parts used-edges(v) ja unused-edges(v). Initially used-edges(v) is empty and unused-edges(v) contains all the edges incident to v.
- Edge (v, u) may be in the adjacency list of v, even if (u, v) is not in the adjacency list of u.

A rotation is allowed also for a used edge. We also include an operation that reverses the orientation of the path.

The purpose of all this is that combined with a random input graph, at any given time each vertex $v \in V$ has the same probability of becoming the head, regardless of previous events.

Additionally, we assume that the input graph is created as follows:

- For all v each edge (u, v) is in the adjacency list of v with probability q independent of the other choices.
- The order of edges in adjacency lists is random.

Again, remember that edges (u, v) and (v, u) are treated independently.

We will later consider how this relates to the model $G_{n,p}$.

We get the final version of the algorithm.

Initialize P as empty path and pick an arbitrary vertex v_1 as head.
Repeat until a Hamiltonian cycle is found
or all the neighbors of the head are used:
Write $P = (v_1, \ldots, v_k)$ where v_k is the head.
Choose one of the following with probabilities
$1/n$, $|\text{used-edges}(v_k)|/n$ and $1 - 1/n - |\text{used-edges}(v_k)|/n$, respectively:
(i) Reverse the path. The new head is v_1.
(ii) Choose a random (v_k, u) from used-edges(v_k).
If $u \in \{ v_1, \ldots, v_{k-2} \}$, make a rotation. Otherwise do nothing.
(iii) Pick the first edge (v_k, u) from unused-edges(v_k).
If $u \notin \{ v_1, \ldots, v_k \}$, extend. Otherwise, make a rotation.
Update used-edges and unused-edges.
If no Hamiltonian cycle was found, return *"no"*.

Let V_t be the vertex that is the head after t steps.

Lemma 5.13: If unused-edges(V_t) has at least one edge, then every $v \in V$ has the same probability $1/n$ of becoming the next head V_{t+1}.

Proof: Let $P = (v_1, \ldots, v_k)$ be the path after t steps.

The vertex v_1 can become the head only by reversing the path, which has probability $1/n$.

If (v_k, v_i) is in the list used-edges(v_k), then v_{i+1} becomes the head with probability

$$\frac{|\text{used-edges}(v_k)|}{n} \cdot \frac{1}{|\text{used-edges}(v_k)|} = \frac{1}{n}.$$

If (v_k, u) is not in the list used-edges(v_k), then by the principle of deferred decisions, the probability for it being the first edge in unused-edges(v_k) is

$$\frac{1}{n - |\text{used-edges}(v_k)| - 1}.$$

This is because we assume there to be at least one edge left, and by the construction any edge has the same probability of being in the adjacency list and there in any given position. Hence, in this case the probability for having $V_{t+1} = v_k$ is

$$\left(1 - \frac{1}{n} - \frac{|\text{used-edges}(v_k)|}{n}\right) \cdot \frac{1}{n - |\text{used-edges}(v_k)| - 1} = \frac{1}{n}.$$

The case where $u \notin \{ v_1, \ldots, v_k \}$ is similar. Notice that in this case (v_k, u) cannot be a used edge. □

We observe that this is a variant of the coupon collecting problem. In each step there is a probability $1/n$ for including a new vertex into to path, and the algorithm finished when all the vertices are there.

Theorem 5.14: Assume that initially each edge (u, v) appears in unused-edges(u) independently with probability $q \geq 20 \ln n/n$. With probability $1 - O(1/n)$ the algorithm finds a Hamiltonian cycle in $O(n \log n)$ iterations.

Notice From this clearly follows that with high probability graphs generates as here do have a Hamiltonian cycle.

Proof: We divide failures into two classes.

- $\mathcal{E}_1$: $3n \ln n$ iterations were executed without running out of unused edges at head, but a cycle was not found
- $\mathcal{E}_2$: some head during the first $3n \ln n$ iterations run out of unused edges.

Consider the event $\mathcal{E}_1$. By Lemma 5.13, we may assume that at each iteration, each vertex will become the new head with probability $1/n$.

The probability that at least one vertex never becomes the head during the first $2n \ln n$ iterations is at most

$$n\left(1 - \frac{1}{n}\right)^{2n \ln n} \leq n\mathrm{e}^{-2 \ln n} = \frac{1}{n}.$$

When all vertices have been at the head at least once, and therefore are included in the path, each iteration has probability $1/n$ of closing the cycle. Hence, the probability that $n \ln n$ iterations pass without this happening is

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

Therefore $\Pr(\mathcal{E}_1) \leq 2/n$.

The event $\mathcal{E}_2$ is further divided into two subevents.

$\mathcal{E}_{2a}$: during the first $3n \ln n$ iterations, at least $9 \ln n$ edges were removed from at least one unused-edges list

$\mathcal{E}_{2b}$: in some unused-edges list there originally were at most $10 \ln n$ edges.

To analyse event $\mathcal{E}_{2a}$ fix a vertex v. Let X be the number of times v becomes the head during the first $3n \ln n$ iterations. Now $X \sim \mathrm{Bin}(3n \ln n, 1/n)$, so (Theorem 4.5; [M&U Thm 4.4.1])

$$\Pr(X \geq 9 \ln n) \leq \left(\frac{e^2}{27}\right)^{3 \ln n} \leq \frac{1}{n^2}.$$

Since unused-edges(v) may lose at most one edge each time v is the head, this is also an upper bound for having more than $9 \ln n$ edges removed.

By using the union bound for n vertices v we get $\Pr(\mathcal{E}_{2a}) \leq 1/n$.

For the event $\mathcal{E}_{2b}$, let Y be the original number of edges in the unused-edges list of some fixed vertex. Because

$$\mathbf{E}[Y] = (n-1)q \geq \frac{20(n-1)\ln n}{n} \geq 19 \ln n$$

for large n, the Chernoffin bound of Theorem 4.9 [M&U Thm 4.5.2] yields

$$\Pr(Y \leq 10 \ln n) \leq \exp(-19(\ln n)(9/19)^2/2) \leq \frac{1}{n^2}.$$

Again the union bound gives $\Pr(\mathcal{E}_{2b}) \leq 1/n$.

Hence the probability of failure is at most

$$\Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_{2a}) + \Pr(\mathcal{E}_{2b}) \leq \frac{4}{n}.$$

□

Corollary 5.15: A random graph in model $G_{n,p}$ where $p \geq (40 \ln n)/n$ can be represented in such a way that with probability $1 - O(1/n)$ our algorithm finds a Hamiltonian cycle in $O(n \ln n)$ iterations.

Proof: We need to show how the edges of $G_{n,p}$ are divided into unused-edges lists. Let q be such that $p = 2q - q^2$.

For any edge (u, v) that was chosen for $G_{n,p}$ we do the following:

- With probability $q(1-q)/(2q-q^2)$ insert (u, v) into unused-edges(u), but not (v, u) into unused-edges(v).
- With probability $q(1-q)/(2q-q^2)$ insert (v, u) into unused-edges(v), but not (u, v) into unused-edges(u).
- With probability $q^2/(2q-q^2)$ insert both (v, u) into unused-edges(v) and (u, v) into unused-edges(u).

The probability for a given (u, v) to be included in unused-edges(u) is

$$p\left(\frac{q(1-q)}{2q-q^2}+\frac{q^2}{2q-q^2}\right)=q$$

as we wanted. Additionally, for any (u, v) the probability of having the egde both in unused-edges(u) and unused-edges(v) is

$$p\cdot\frac{q^2}{2q-q^2}=q^2$$

so the edges are independent. □

6. Probabilistic method

We want to prove the existence of a combinatorial object (such as graph) that satisfies some non-trivial conditions. We do this by proving that a random object chosen with a suitable distribution has a strictly positive probability of satisfying the condition.

This also gives a randomized algorithm for constructing such objects. Sometimes they can be efficiently derandomized.

The counting argument [M&U Section 6.1]

We start with graphs coloring. Let K_n be the complete n-vertex graph (all $n(n-1)/2$ possible edges are there).

Theorem 6.1: If $\binom{n}{k}2^{-k(k-1)/2+1} < 1$, then the edges of K_n can be colored with two colors so that there is no monochromatic k-clique (that is, any subgraph isomorphic to K_k contains edges of two different colors).

The condition of the theorem is implied for example by $n \leq 2^{k/2}$ and $k \geq 3$. Then

$$\begin{aligned}\binom{n}{k}2^{-k(k-1)/2+1} &\leq \frac{n^k}{k!}2^{-k(k-1)/2+1} \\ &\leq \frac{2^{k/2+1}}{k!} \\ &< 1.\end{aligned}$$

Proof: Color the edges of K_n with two colors so that each edge gets either color with probability $1/2$ independently of the others.

We fix an arbitrary numbering on the k-cliques of K_n. (There are $\binom{n}{k}$ of them). Let A_i be the event that clique number i is monochromatic. For this to happen, after arbitrarily coloring the first edge in the clique, the remaining $k(k-1)/2-1$ edges must get the matching color. Therefore

$$\Pr(A_i) = 2^{-k(k-1)/2+1}.$$

Since the number of k-cliques is $\binom{n}{k}$, we get

$$\Pr(\cup_i A_i) \leq \binom{n}{k} 2^{-k(k-1)/2+1}.$$

The condition in the statement of the theorem has been picked so that the right-hand side is strictly less than 1. $\square$

Consider now algorithms for constructing a coloring with no monochromatic k-cliques.

By comparing the comment made after the theorem to the proof, we see that for $n \leq 2^{k/2}$ and $k \geq 3$ a random coloring has the desired property with probability at least $1 - 2^{k/2+1}/k!$. For example, $k = 20$ and $n = 1000$ give success probability at least

$$1 - \frac{2^{20/2+1}}{20!} \geq 1 - 8{,}5 \cdot 10^{-16}.$$

This directly gives a Monte Carlo algorithm for the problem.

To get a Las Vegas algoritmi, we'd also need to check that the solution actually is correct. For constant k this can be done by exhaustive search in time $\binom{n}{k} = O(n^k)$. In the general case it is not clear how to do this efficiently.

The expectation argument [M&U Section 6.2]

This method is based on the following simple observation.

Theorem 6.2: If $\mathbf{E}[X] = \mu$, then $\Pr(X \geq \mu) > 0$ and $\Pr(X \leq \mu) > 0$.

Proof: If $\Pr(X \geq \mu) = 0$, then

$$\mathbf{E}[X] = \sum_x x \Pr(X = x) = \sum_{x<\mu} x \Pr(X = x) < \sum_{x<\mu} \mu \Pr(X = x) = \mu.$$

The case $X \leq \mu$ is similar. $\square$

In particular, if X is a random variable $\Omega \to \mathbb{R}$ and $\mu = \mathbf{E}[X]$, then for some $\omega_+ \in \Omega$ and $\omega_- \in \Omega$ we have $X(\omega_-) \leq \mu$ and $X(\omega_+) \geq \mu$.

As an example we consider finding large cuts in a graph.

A cut in an edge-weighted graph is a partition of its vertices into two subsets A and B, and the value of the cut is the total weight of all edges between A and B.

Finding the maximum cut (i.e. cut with maximum value) is NP-hard. Here we give a simple lower bound for the value of the maximum cut when all the edges have weight 1.

Theorem 6.3: If a graph $G = (V, E)$ has m edges each with weight 1, the value of the maximum cut is at least $m/2$.

Proof: Assign each vertex independently of the others into A or B. Let $C(A, B)$ be the value of the cut (A, B) (which is a random variable).

Write $E = \{ e_1, \ldots, e_m \}$ and define

$$X_i = \begin{cases} 1 & \text{if } e_i \text{ is between } A \text{ and } B \\ 0 & \text{otherwise.} \end{cases}$$

When one end point of the edge has been assigned, the probability for the other end to be assigned to the other part is

$$\mathbf{E}[X_i] = \Pr(X_i = 1) = \frac{1}{2}.$$

Therefore

$$\mathbf{E}[C(A, B)] = \mathbf{E}\left[\sum_{i=1}^{m} X_i\right] = \frac{m}{2},$$

so at least one cut (A, B) must have $C(A, B) \geq m/2$. □

Consider now the obvious Las Vegas -algorithm to find a cut with value at least $m/2$.

Repeat until success:
1. Assign sets A and B randomly.
2. If there are at least $m/2$ edges between A and B
   then print (A, B)
   else failed.

The test on line 2 takes time $O(m)$. Next we estimate the probability of success

$$p = \Pr\left(C(A,B) \geq \frac{m}{2}\right).$$

Always $C(A,B) \leq m$, so

$$\begin{aligned}
\frac{m}{2} &= \mathbf{E}[C(A,B)] \\
&= \sum_{i \leq m/2-1} i \Pr(C(A,B)=i) + \sum_{i \geq m/2} i \Pr(C(A,B)=i) \\
&\leq (1-p)\left(\frac{m}{2}-1\right) + pm.
\end{aligned}$$

Therefore

$$p \geq \frac{1}{m/2+1}.$$

The expected number of iteration to get a valid solution is at most $m/2+1$, so we have a polynomial time Las Vegas algorithm.

We will soon return to the question of a deterministic algorithm for the problem.

As the second example we consider maximum satisfiability (MAXSAT).

The input is a set of clauses, which are disjunctions of literals. A literal is a boolean variable or a negated boolean variable. We write clauses like for example $x_1 \vee \overline{x_3} \vee x_8$.

In the boolean satisfiability problem (SAT) the task is to decide whether there is a value assignment for the variables that satisfies all the clauses. This is known to be NP-complete.

In maximum satisfiability we try to find a value assignment that satisfies as many clauses as possible, but not necessarily all of them. Clearly an exact solution is NP-hard.

We start with a lower bound for the number of satisfied clauses.

Theorem 6.4: Let a MAXSAT instance consist of m clauses, with clause number i having k_i literals. Write $k = \min_i k_i$. There is a variable assignment that satisfies at least

$$\sum_{i=1}^{m}(1 - 2^{-k_i}) \geq m(1 - 2^{-k})$$

clauses.

Proof: Assign the variables at random. Clause i remains unsatisfied with probability at most $(1/2)^{k_i}$. Therefore, the expected number of satisfied clauses is at least

$$\sum_{i=1}^{m}(1 - 2^{-k_i}).$$

In particular, there is at least one assignment for which this value is reached. □

Method of conditional expectations [M&U Section 6.3]

This is a way of derandomising algorithms. Consider the maximum cut problem we saw recently.

Fix an ordering for the vertices and number them accordingly: $V = \{ v_1, \ldots, v_n \}$. We make assign the vertices into A or B in this order. The notation $\mathbf{E}[C(A,B) \mid x_1, \ldots, x_k]$ stands for the expected value of the cut, when the vertices $v_1, \ldots, v_k$ have been assigned, but vertices $v_{k+1}, \ldots, v_n$ remain to be assigned randomly. Here x_i is either "$v_i \in A$" or "$v_i \in B$".

We saw earlier that $\mathbf{E}[C(A,B)] \geq m/2$. By symmetry, $\mathbf{E}[C(A,B) \mid x_1] \geq m/2$ regardless of whether we assigned v_1 to A or B.

We will now show that for any given assignments $x_1, \ldots, x_k$ we can choose $v_{k+1} \in A$ or $v_{k+1} \in B$ to guarantee

$$\mathbf{E}[C(A,B) \mid x_1, \ldots, x_k, x_{k+1}] \geq \mathbf{E}[C(A,B) \mid x_1, \ldots, x_k].$$

By induction we get

$$\mathbf{E}[C(A,B) \mid x_1, \ldots, x_n] \geq m/2.$$

Because here all the vertices are assigned, the desired result follows.

By the definition of $\mathbf{E}[C(A,B) \mid \cdot]$,

$$\begin{aligned}\mathbf{E}[C(A,B) \mid x_1,\ldots,x_k] &= \frac{1}{2}\mathbf{E}[C(A,B) \mid x_1,\ldots,x_k, v_{k+1} \in A] \\ &\quad + \frac{1}{2}\mathbf{E}[C(A,B) \mid x_1,\ldots,x_k, v_{k+1} \in B].\end{aligned}$$

Therefore,

$$\max_{X \in \{A,B\}} \mathbf{E}[C(A,B) \mid x_1,\ldots,x_k, v_{k+1} \in X] \geq \mathbf{E}[C(A,B) \mid x_1,\ldots,x_k].$$

If we pick $X = A$ or $X = B$ depending on which makes $\mathbf{E}[C(A,B) \mid x_1,\ldots,x_k, v_{k+1} \in X]$ larger, we get the desired result.

The conditional expectation $\mathbf{E}[C(A,B) \mid x_1, \ldots, x_k, v_{k+1} \in X]$ can be determined by adding the contributions of individual edges:

- If both end points of an edge are in $\{ v_1, \ldots, v_{k+1} \}$, they have been assigned and we know whether the edge contributes 0 or 1.
- Otherwise at least one edge is in the set $\{ v_{k+2}, \ldots, v_n \}$ and unassigned. The edge will be in the cut with probability $1/2$, which is also its contribution to the expectation.

Based on this, we can calculate $\mathbf{E}[C(A,B) \mid x_1, \ldots, x_k, v_{k+1} \in X]$ in linear time both for $X = A$ and $X = B$, and assign v_{k+1} based on that.

Looking into this more closely we see that the only edges that do not have the same contribution in $\mathbf{E}[C(A,B) \mid x_1, \ldots, x_k, v_{k+1} \in A]$ and $\mathbf{E}[C(A,B) \mid x_1, \ldots, x_k, v_{k+1} \in B]$ are those whose one end point is v_{k+1} and the other one is in $\{ v_1, \ldots, v_k \}$.

We get a greedy algorithm:

1. Initialize $A := \emptyset$ and $B := \emptyset$.
1. Arbitrarily assign $A := A \cup \{ v_1 \}$ or $B := B \cup \{ v_1 \}$.
2. Repeat for $k = 1, \ldots, n-1$:
   (a) Let $N_A = \{ u \in A \mid (v_k, u) \in E \}$ and $N_B = \{ u \in B \mid (v_k, u) \in E \}$.
   (b) If $|N_A| \leq |N_B|$ then $A := A \cup \{ v_{k+1} \}$
   else $B := B \cup \{ v_{k+1} \}$.

Based on the above, this gives a cut with value at least $m/2$.

Sample and modify [M&U Section 6.4]

In this technique we first choose a structure at random and then modify it to get the desired properties.

Consider as a first example independent sets. In a graph $G = (V, E)$, we say that a set of vertices $U \subseteq V$ is independent if there are no edges between its vertices, that is, $(U \times U) \cap E = \emptyset$. Finding the maximum independent set is a known NP-hard problem.

Theorem 6.5: If a graph has n vertices and m edges, it has an independent set with at least $n^2/4m$ vertices.

Proof: Let $d = 2m/n$ be the average degree of vertices. We make the following two randomized steps:

1. For each vertex independently, remove the vertex and its incident edges with probability $1 - 1/d$.
2. For each remaining edge independently, remove the edge and one of its end points.

Thus, we first take a random sample of the graph and then modify it.

If there is an edge between vertices u and v, then at least one of the vertices is removed at latest in Step 2. Therefore, after the two steps we are left with a set of vertices that constitutes an independent set.

Let X be the number of vertices remaining after Step 1. Therefore, $\mathbf{E}[X] = n/d$.

Let Y be the number of edges left after Step 1. The edge is left, if both its end points are. Therefore,

$$\mathbf{E}[Y] = m\left(\frac{1}{d}\right)^2 = \frac{nd}{2}\frac{1}{d^2} = \frac{n}{2d}.$$

After Step 2 we have at least $X - Y$ vertices left, and

$$\mathbf{E}[X - Y] = \frac{n}{d} - \frac{n}{2d} = \frac{n}{2d} = \frac{n^2}{4m}.$$

□

As a second example, consider the girth of a graph, that is the length of the shortest cycle. We show that even a fairly dense graph can have a fairly large girth.

Theorem 6.6: For $k \geq 3$ and n sufficiently large, there exists a graph with n vertices, at least $\frac{1}{4}n^{1+1/k}$ edges, and girth at least k.

Proof: First choose a random graph in model $G_{n,p}$ with $p = n^{1/k-1}$. After that remove an arbitrary edge from every cycle of length at most $k-1$. The girth of the remaining graph is thus at least k.

For the number X of edges originally chosen into the random graph we have

$$\mathbf{E}[X] = p\binom{n}{2} = \frac{1}{2}\left(1 - \frac{1}{n}\right) n^{1/k+1}.$$

Let Y be the number of cycles of length at most $k-1$. A given cycle of length i has probability p^i of being present, and there are $\binom{n}{i}(i-1)!/2$ such cycles. Therefore,

$$\mathbf{E}[Y] = \sum_{i=3}^{k-1} \binom{n}{i} \frac{(i-1)!}{2} p^i \leq \sum_{i=3}^{k-1} n^i p^i = \sum_{i=3}^{k-1} n^{i/k} < k n^{(k-1)/k}.$$

We are left with at least $X - Y$ edges, and for large n

$$\mathbf{E}[X - Y] \geq \frac{1}{2}\left(1 - \frac{1}{n}\right) n^{1/k+1} - k n^{(k-1)/k} \geq \frac{1}{4} n^{1/k+1}.$$

□

The second moment method [M&U Section 6.5]

From Chebyshev's Inequality we obtain

Theorem 6.7: If all possible values of X are non-negative integers, then

$$\Pr(X = 0) \leq \frac{\mathbf{Var}[X]}{(\mathbf{E}[X])^2}.$$

Proof:

$$\Pr(X = 0) \leq \Pr(|X - \mathbf{E}[X]| \geq \mathbf{E}[X]) \leq \frac{\mathbf{Var}[X]}{(\mathbf{E}[X])^2}.$$

□

We use this to analyse threshold values. In a random graph $G_{n,p}$, the probabilities of certain events transition very quicly from 0 to 1 when p crosses some threshold value (which is a function of n).

Theorem 6.8: Let $G = G_{n,p}$, where $p = f(n)$. Let A be the event that G contains a 4-clique.

- If $f(n) = o(n^{-2/3})$ (that is, $\lim_{n\to\infty} pn^{2/3} = 0$), then $\Pr(A) = o(1)$.
- If $f(n) = \omega(n^{-2/3})$ (that is, $\lim_{n\to\infty} pn^{2/3} = \infty$), then $\Pr(A) = 1 - o(1)$.

Proof: Let $C_1, \ldots, C_M$ be all the four-vertex sets in G, where $M = \binom{n}{4}$. For $i = 1, \ldots, M$, define

$$X_i = \begin{cases} 1 & \text{if } C_i \text{ is a clique} \\ 0 & \text{otherwise,} \end{cases}$$

and $X = \sum_{i=1}^{M} X_i$. Since $\Pr(X_i = 1) = p^6$, we have

$$\mathbf{E}[X] = \binom{n}{4} p^6 = \Theta(n^4 p^6) = \Theta((pn^{2/3})^6).$$

In the case $f = o(n^{-2/3})$ we thus have $\mathbf{E}[X] = o(1)$. As X only gets non-negative integer values, $\mathbf{E}[X] \geq \Pr(X \geq 1)$. Therefore,

$$\Pr(X \geq 1) \leq \mathbf{E}[X] = o(1).$$

In the case $f = \omega(n^{-2/3})$ we similarly get $\lim_{n\to\infty} \mathbf{E}[X] = \infty$. To apply Theorem 6.7 we still need to show that

$$\frac{\mathbf{Var}[X]}{(\mathbf{E}[X])^2} = o(1).$$

We start with an auxiliary result: for any $Y = \sum_i Y_i$ we have

$$\begin{aligned}
\mathbf{Var}[Y] &= \mathbf{E}[Y^2] - (\mathbf{E}[Y])^2 \\
&= \mathbf{E}\left[\sum_i Y_i^2 + 2\sum_{i<j} Y_i Y_j\right] - \sum_i (\mathbf{E}[Y_i])^2 - 2\sum_{i<j} \mathbf{E}[Y_i]\mathbf{E}[Y_j] \\
&= \sum_i \mathbf{Var}[Y_i] + 2\sum_{i<j} \mathbf{Cov}(Y_i, Y_j).
\end{aligned}$$

In particular, if Y_i is 0-1 valued, we get

$$\mathbf{Var}[Y_i] = \mathbf{E}[Y_i^2] - (\mathbf{E}[Y_i])^2 = \mathbf{E}[Y_i] - (\mathbf{E}[Y_i])^2 \leq \mathbf{E}[Y_i].$$

Hence,

$$\mathbf{Var}[Y] \leq \mathbf{E}[Y] + 2\sum_{i<j} \mathbf{Cov}(Y_i, Y_j).$$

To apply this to X, we need the covariances $\mathbf{Cov}(X_i, X_j)$.

If $C_i \cap C_j = \emptyset$, then X_i and X_j are independent and $\mathbf{Cov}(X_i, X_j) = 0$. This is true also if $|C_i \cap C_j| = 1$.

If $|C_i \cap C_j| = 2$, the corresponding cliques share one egde, so the total number of distinct edges is $6 + 6 - 1$. Then

$$\mathbf{Cov}(X_i, X_j) = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\mathbf{E}[X_j] \leq \mathbf{E}[X_i X_j] = p^{11}.$$

There are

$$\binom{n}{2}\binom{n-2}{2}\binom{n-4}{2} = \Theta(n^6)$$

such pairs (i, j). If $|C_i \cap C_j| = 3$, the corresponding cliques share three edges and

$$\mathbf{Cov}(X_i, X_j) \leq \mathbf{E}[X_i X_j] = p^9.$$

There are

$$n(n-1)\binom{n-2}{3} = \Theta(n^5).$$

such pairs (i, j).

Altogether, for $p = \omega(n^{-2/3})$ we get

$$\begin{aligned}\mathbf{Var}[X] &\leq \mathbf{E}[X] + \sum_{i \neq j} \mathbf{Cov}(X_i, X_j) \\ &= \Theta(n^4 p^6) + \Theta(n^6 p^{11}) + \Theta(n^5 p^9) \\ &= o(n^8 p^{12}),\end{aligned}$$

since, for example,

$$\frac{n^4 p^6}{n^8 p^{12}} = n^{-4} p^{-6} = o(n^{-4}(n^{-2/3})^{-6}) = o(1).$$

Because

$$(\mathbf{E}[X])^2 = \left(\binom{n}{4} p^6\right)^2 = \Theta(n^8 p^{12}),$$

we get

$$\mathbf{Var}[X] = o((\mathbf{E}[X])^2)$$

as desired. □

The Conditional Expectation Inequality [M&U Section 6.6]

If X is the sum of 0-1 valued random variables, as in the previous example, the following bound may be easier to apply.

Theorem 6.9: Let $X = \sum_i X_i$, where each X_i is 0-1 valued. Then

$$\Pr(X > 0) \geq \sum_i \frac{\Pr(X_i = 1)}{\mathbf{E}[X \mid X_i = 1]}.$$

Notice There is no independence assumption on the X_i.

Proof: Define $Y = 1/X$ if $X > 0$, and $Y = 0$ if $X = 0$. Thus, $\Pr(X > 0) = \mathbf{E}[XY]$.

We can estimate this as

$$\begin{aligned} E[XY] &= \sum_i \mathbf{E}[X_i Y] \\ &= \sum_i (\mathbf{E}[X_i Y \mid X_i = 1] \Pr(X_i = 1) + \mathbf{E}[X_i Y \mid X_i = 0] \Pr(X_i = 0)) \\ &= \sum_i \mathbf{E}[1/X \mid X_i = 1] \Pr(X_i = 1) \\ &\geq \sum_i \frac{\Pr(X_i = 1)}{\mathbf{E}[X \mid X_i = 1]}, \end{aligned}$$

where the last step is from Jensen's Inequality. □

As an example we give an alternative proof for the case $p = \omega(n^{-2/3})$ in Theorem 6.8

Let the random variables X_i, $i = 1, \ldots, M$, again be the indicator variables of the cliques, so $\Pr(X_j = 1) = p^6$. We have

$$\mathbf{E}[X \mid X_j = 1] = \sum_{i=1}^{M} \mathbf{E}[X_i \mid X_j = 1] = \sum_{i=1}^{M} \Pr(X_i = 1 \mid X_j = 1).$$

For a given j there are $\binom{n-4}{4}$ indices i such that $|C_i \cap C_j| = 0$, and $4\binom{n-4}{3}$ indices i such that $|C_i \cap C_j| = 1$. For all these we have

$$\Pr(X_i = 1 \mid X_j = 1) = \Pr(X_i = 1) = p^6.$$

Altogether there are $\binom{4}{2}\binom{n-4}{2} = 6\binom{n-4}{2}$ indices i such that $|C_i \cap C_j| = 2$ and $Pr(X_i = 1 \mid X_j = 1) = p^5$.

In total, $\binom{4}{3}\binom{n-4}{1} = 4(n-4)$ indices i have $|C_i \cap C_j| = 3$ and $Pr(X_i = 1 \mid X_j = 1) = p^3$.

By including the case $i = j$ we obtain

$$\begin{aligned}\mathbf{E}[X \mid X_j = 1] &= \sum_{i=1}^{M} \mathbf{E}[X_i \mid X_j = 1] \\ &= \binom{n-4}{4}p^6 + 4\binom{n-4}{3}p^6 + 6\binom{n-4}{2}p^5 + 4(n-4)p^3 + 1.\end{aligned}$$

When $p = \omega(n^{-2/3})$, the term $\binom{n-4}{4}p^6$ dominates, so

$$\mathbf{E}[X \mid X_j = 1] \sim \binom{n-4}{4}p^6 \sim \frac{1}{4!}n^4p^6.$$

Since $M = \binom{n}{4} \sim n^4/4!$, we get

$$\sum_{j=1}^{M} \frac{\Pr(X_j = 1)}{\mathbf{E}[X \mid X_j = 1]} \sim \frac{Mp^6}{\binom{n-4}{4}p^6} = 1.$$

□

Lovász Local Lemma [M&U Section 6.7]

We consider a set of undesired events $E_1, \ldots, E_n$. We want to show that if each of them individually has low probability, then also the intersection of their complements, that is the set of desired events, is non-empty.

If the E_i are mutually independent, there is no problem. Then also the complements are mutually independent, and

$$\Pr\left(\bigcap_{i=1}^{n} \overline{E_i}\right) = \prod_{i=1}^{n} \Pr(\overline{E_i}) > 0$$

assuming $\Pr(\overline{E_i}) > 0$ for all i.

In the following we weaken the independence assumption to allow "local" dependencies. We say that an event A is mutually independent of $E_1, \ldots, E_n$, if for all $I \subseteq \{1, \ldots, n\}$ we have

$$\Pr\left(A \mid \bigcap_{i \in I} E_i\right) = \Pr(A).$$

Consider any events $E_1, \ldots, E_n$. A dependency graph for these events is a graph $G = (V, E)$ such that $V = \{1, \ldots, n\}$ and for all i the event E_i is mutually independent of $\{E_j \mid (i,j) \notin E\}$.

In particular, a dependency graph always contains the edge (i, i) for all i.

Example 6.10 [M&U Section 6.7.2]**:** Let $\phi = \phi_1 \wedge \ldots \wedge \phi_m$ be a CNF formula where each ϕ_i is a clause. Assign mutually independent random values to the variables in the formula. Let E_i be the event "clause ϕ_i is satisfied." Define a graph $G = (V, E)$ where $V = \{1, \ldots, m\}$ and $(i, j) \in E$ if ϕ_i and ϕ_j have at least one common variable. Now G is a dependency graph for the E_i. □

Theorem 6.11 (Lovász Local Lemma [M&U Thm 6.11]): Let $E_1, \ldots, E_n$ be a set of events satisfying the following conditions:

1. $\Pr(E_i) \leq p$, where p is a constant
2. the degree of the dependency graph of the E_i is at most d and
3. $4dp \leq 1$.

Then

$$\Pr\left(\bigcap_{i=1}^{n} \overline{E_i}\right) > 0.$$

Note 1: This is just a special case of the original lemma (so-called symmetrical case).

Note 2: Compare with the union bound. If $\Pr(E_i) \leq p$ for all i and $np < 1$, then $\Pr\left(\cap_{i=1}^{n} \overline{E_i}\right) > 0$.

Note 3: Since $d \geq 1$, condition 3 implies $p \leq 1/4$.

Example 6.12: Consider satisfiability of Boolean CNF formulas as earlier. We assume that ϕ is a k-CNF formula, meaning that each clause ϕ_i has exactly k literals. Then

$$\Pr(\phi_i = 0) = \left(\frac{1}{2}\right)^k.$$

Assume futher that no variable appears in more than T clauses, where $T = 2^k/4k$. Then in the dependency graph, the number of edges related to one clause is at most

$$d = kT = 2^{k-2}.$$

By choosing $p = 2^{-k}$, we get $4pd \leq 1$, and the assumptions of the local lemma hold. Therefore,

$$\Pr\left(\bigcap \overline{E_i}\right) > 0,$$

so there exists a value assignment that satisfies all the clauses. □

Proof of the local lemma: We do an induction over a parameter s to show that if $|S| \leq s$, then for all $k \notin S$ we have

$$\Pr\left(E_k \mid \bigcap_{j \in S} \overline{E_j}\right) \leq 2p.$$

This yields the desired result:

$$\begin{aligned}
\Pr\left(\bigcap_{i=1}^{n} \overline{E_i}\right) &= \prod_{i=1}^{n} \Pr\left(\overline{E_i} \mid \bigcap_{j=1}^{i-1} \overline{E_j}\right) \\
&= \prod_{i=1}^{n} \left(1 - \Pr\left(E_i \mid \bigcap_{j=1}^{i-1} \overline{E_j}\right)\right) \\
&\geq \prod_{i=1}^{n} (1 - 2p) \\
&> 0.
\end{aligned}$$

The base case $s = 0$ is directly in the assumptions. Assume now that the claim holds for $|S| < s$. We first show

$$\Pr\left(\bigcap_{j\in S} \overline{E_j}\right) > 0,$$

which is required for the conditional probability to be defined.

If $s = 1$, we get directly $\Pr(\overline{E_j}) = 1 - \Pr(E_j) \geq 1 - p > 0$. If $s > 1$, we may assume $S = \{1, \ldots, s\}$, and as on previous page,

$$\begin{aligned}
\Pr\left(\bigcap_{i=1}^{s} \overline{E_i}\right) &= \prod_{i=1}^{s} \Pr\left(\overline{E_i} \mid \bigcap_{j=1}^{i-1} \overline{E_j}\right) \\
&= \prod_{i=1}^{s} \left(1 - \Pr\left(E_i \mid \bigcap_{j=1}^{i-1} \overline{E_j}\right)\right) \\
&\overset{\text{ind.ass.}}{\geq} \prod_{i=1}^{s} (1 - 2p) \\
&> 0.
\end{aligned}$$

Let the dependency graph be (V, E). Fix E_k and S, and define $S_1 = \{ j \in S \mid (k, j) \in E \}$ and $S_2 = \{ j \in S \mid (k, j) \notin E \}$.

If $S_1 = \emptyset$, then E_k is mutually independent of $\{ E_j \mid j \in S \}$, and the claim holds. Consider then the case $|S_2| < s$. Write

$$F_X = \bigcap_{j \in X} \overline{E_j},$$

for $X \in \{ S, S_1, S_2 \}$. In particular, $F_S = F_{S_1} \cap F_{S_2}$. We apply the basic property of conditional expectation

$$\Pr(A \mid B \cap C) = \frac{\Pr(A \cap B \mid C)}{\Pr(B \mid C)},$$

which yields

$$\Pr(E_k \mid F_S) = \frac{\Pr(E_k \cap F_{S_1} \mid F_{S_2})}{\Pr(F_{S_1} \mid F_{S_2})}.$$

We estimate separately the numerator and the denominator.

We estimate the numerator simply by

$$\Pr(E_k \cap F_{S_1} \mid F_{S_2}) \leq \Pr(E_k \mid F_{S_2}) = \Pr(E_k) \leq p,$$

since E_k is mutually independent of $\{ E_j \mid j \in S_2 \}$

For the denominator notice first that $|S_2| < s$, so we can apply the inductive assumption:

$$\Pr(E_i \mid F_{S_2}) = \Pr\left(E_i \mid \bigcap_{j \in S_2} \overline{E_j} \right) \leq 2p.$$

Therefore,

$$\begin{aligned}
\Pr(F_{S_1} \mid F_{S_2}) &= \Pr\left(\bigcap_{i \in S_1} \overline{E_i} \mid F_{S_2}\right) \\
&\geq 1 - \sum_{i \in S_1} \Pr\left(E_i \mid F_{S_2}\right) \\
&\geq 1 - \sum_{i \in S_1} 2p \\
&\geq 1 - 2dp \\
&\geq \frac{1}{2}.
\end{aligned}$$

We get

$$\Pr(E_k \mid F_S) = \frac{\Pr(E_k \cap F_{S_1} \mid F_{S_2})}{\Pr(F_{S_1} \mid F_{S_2})} \leq \frac{p}{1/2} = 2p$$

which was the claim of the induction. □

Disjoint paths [M&U Section 6.7.1]

Assume that n pairs of users want to simultaneously communicate over a network. We consider finding for each pair their own communication path that does not share edges with others.

Let F_i be the set of edges that pair i could use if there were no other users.

Theorem 6.13: Let m and k be such that $8nk/m \leq 1$. If

1. $|F_i| \geq m$ for all i and

2. for all $i \neq j$ and any path $P' \in F_i$ there are at most k paths $P'' \in F_j$ such that P' and P'' have at least one common edge

then it's possible to choose one path from each set F_i such that none of the chosen paths have common edges.

Proof: It is sufficient to consider the case $|F_i| = m$ for all i. Choose a random path from each F_i. Let $E_{i,j}$ be the event that the paths chosen from F_i and F_j have at least one common edge.

Whichever path P' we choose from F_i, there are m ways of choosing P'' from F_j and at most k of them have a common edge with P'.

Therefore, if we choose $p = k/m$ we have

$$\Pr(E_{i,j}) \leq p.$$

Since $E_{i,j}$ is mutually independent of $\{ E_{s,t} \mid \{ s,t \} \cap \{ i,j \} = \emptyset \}$, the dependency graph has degree $d < 2n$. Hence,

$$4dp < \frac{8nk}{m} \leq 1.$$

The local lemma now gives

$$\Pr\left(\bigcap_{i \neq j} \overline{E_{i,j}} \right) > 0$$

from which the claim follows. $\Box$

The general version of local lemma [M&U Section 6.9]

For completeness, we give without proof the general version of Lovász local lemma.

Theorem 6.14: Let $G = (V, E)$ be a dependency graph for events $\{ E_1, \ldots, E_n \}$. Assume we have values $0 \leq x_i \leq 1$, $i = 1, \ldots, n$, such that

$$\Pr(E_i) \leq x_i \prod_{(i,j) \in E} (1 - x_j) \qquad \text{for all } i.$$

Then

$$\Pr\left(\bigcap_{i=1}^{n} \overline{E_i} \right) \geq \prod_{i=1}^{n} (1 - x_i).$$

We omit the proof, which can be found in the textbook. Here we just show how this general version can be used to derive the symmetric case we considered earlier.

Assume that the conditions for the symmetric version of the local lemma (Theorem 6.11 [M&U Thm 6.11]) are satisfied:

1. $\Pr(E_i) \leq p$
2. the degree of the dependency graph is at most d and
3. $\mathrm{e}p(d+1) \leq 1$.

(Condition 3 is a bit weaker than in Theorem 6.11, so we get a slightly stronger result.)

We want to prove

$$\Pr\left(\bigcap_{i=1}^{n} \overline{E_i}\right) > 0.$$

We claim that choosing $x_i = 1/(d+1)$ for all i satisfies the conditions of the general lemma. With this choice we get

$$x_i \prod_{(i,j)\in E} (1 - x_j) \geq \frac{1}{d+1} \left(1 - \frac{1}{d+1}\right)^d$$

since there are at most d indices j such that $(i,j) \in E$. Denote the quantity on the right-hand side by q.

To apply the general version of the lemma, we want to show $q \geq \Pr(E_i)$. Our assumptions $\mathrm{e}p(d+1) \leq 1$ and $\Pr(E_i) \leq p$ imply $\Pr(E_i) \leq \mathrm{e}^{-1}(d+1)^{-1}$. Therefore it is sufficient to show $q(d+1) \geq \mathrm{e}^{-1}$.

We apply the inequality

$$\mathrm{e}^{-x}\left(1-\frac{x^2}{k}\right) \leq \left(1-\frac{x}{k}\right)^k.$$

Using this in the bound from previous page gives

$$\begin{aligned} q(d+1) &\geq \left(1-\frac{1}{d+1}\right)^d \\ &= \frac{d+1}{d}\left(1-\frac{1}{d+1}\right)^{d+1} \\ &\geq \frac{d+1}{d}\mathrm{e}^{-1}\left(1-\frac{1}{d+1}\right) \\ &= \mathrm{e}^{-1} \end{aligned}$$

as desired.

The conditions of the general version of the lemma are satisfied, so we can conlude that

$$\Pr\left(\bigcap_{i=1}^{n}\overline{E_i}\right) > 0.$$

7. Brief summary

Randomization is a general tool that can be applied to very different problems. In this course we covered some basic mathematical tools for analysing randomness. They are useful also for analysing deterministic algorithms in random environments.

Usually we start by finding out what happens in expectation. Linearity of expectation is a powerful tool.

To find out the probability of getting a result far away from the expectation, we use results such as Chernoff bounds, often together with the union bound.

Many of these tools, including Jensen's inequality, are also very important in machine learning.

About the exam

The exam cover the whole material of the course. Possible types of question include

- similar to what you've seen as homework
- exactly what you've seen as homework
- explain some concept from the course material
- prove a known theorem from the course material.

You don't need to memorize complicated formulas. If you need to use any, they will be provided. However, remember that you may be asked to prove for example Chernoff bounds.

What next?

In *Randomized algorithms II*, we start with Markov chains, which in some sense are randomized state machines.

They can be used for modelling various processes. In particular, they often have a stationary distribution that may be easy to analyse.

This leads also to the Markov Chain Monte Carlo method: to sample from a complicated distribution (say, pick a random independent set in a graph), construct a Markov Chain with the desired distribution as its stationary distribution. Analysing such sampling methods can be quite difficult.

We also consider Poisson processes, which can be used to model waiting times between certain types of random events.