Probability & Statistics 2

Robert Šámal

January 29, 2024

Contents

This is work in progress, likely full of typos and small (or larger) inprecisions. If you find something that looks wrong, feel free to reach out to me. Thanks to Kryštof Višňák for helping me with spotting typos in a previous version.

1 Markov Chains

1.1 Introduction, basic properties

Two examples to start with A machine can be in two states: working or broken. For simplicity, we assume that the state stays the same for the whole day. Then, during the night, the state changes at random according to the figure below: for instance, if the machine is working one day, it will work the next day with probability 0.99, with probability 0.01 it breaks over night. Crucially, we assume that this probability does not depend on the age of the machine, nor on the previous states.

A fly is moving in a corridor, that we consider as a collection of four spaces, labeled 0, 1, 2, 3. If the fly is in spaces 1 or 2, it stays at the same space with probability 0.4. Otherwise, it moves equally likely one step left or right. At positions 0 and 3 is a spider and the fly can never leave. Again, we assume that "the fly has no memory", so the probabilities do not depend on the past trajectory of the fly.

TODO: add figures

What are the common features of these examples? We consider a sequence of random variables, so called *random process*. We do not care about the numerical value of these variables, as we consider them as mere labels – so we will not ask about expected value of a position of the fly, for instance. We may assume that all the random variables have range contained in set S of labels. For simplicity we assume S to be finite or countable (and frequently we will assume that $S = \{1, \dots, s\}$ or $S = \mathbb{N}$). We also want to prescribe *transition probabilities* $p_{i,j}$ such that $P(X_{t+1} = j \mid X_t = i) = p_{i,j}$. However, there is more subtlety to this: we want to explicitly forbid the history (values of $X_0, \dots, X_{t-1}$ to have an influence on X_{t+1}. (See also section of stochastic processes.)

Definition 1 (Markov chain)**.** *Let S be any finite or countably infinite set. A sequence $(X_t)_{t=0}^{\infty}$ of random variables with range S is a (discrete time, discrete space, time-homogeneous)* Markov chain *if for every $t \geq 0$ and every $a_0, \dots, a_{t+1} \in S$ we have*

$$P(X_{t+1} = a_{t+1} \mid X_t = a_t \,\&\, \dots \,\&\, X_0 = a_0) = P(X_{t+1} = a_{t+1} \mid X_t = a_t) \qquad (1)$$

whenever the conditional probabilities are defined, that is when $P(X_t = a_t \,\&\, \dots \,\&\, X_0 = a_0) > 0$. We call the numbers $p_{i,j}(t) = P(X_{t+1} = j \mid X_t = i)$ the transition probabilities. We will only study cases where $p_{i,j}(t)$ does not depend on t and will omit the t in the notation.

Transition matrix is a matrix P such that $P_{i,j} = p_{i,j}$, that is the entry at i-th row and j-th column is the probability of transition from state i to state j. As a consequence of the definition, all entries in the transition matrix are nonnegative, and each row sums to 1. We can describe this succintly by writing $Pj = j$ with j denoting the column vector of all 1's.

Let P denote the transition matrix for the machine example and Q for the fly example. We have

$$P = \begin{pmatrix} 0.99 & 0.01 \\ 0.9 & 0.1 \end{pmatrix} \qquad Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Transition graph/diagram is a directed graph with vertex set S and arcs (directed edges) (i, j) *for every* $i, j \in S$ *such that* $p_{i,j} > 0$. We label arc (i, j) by $p_{i,j}$. In other words, the figures above (TODO) are transition graphs.

Describing the distribution. We will again use the basic tool to describe a random variable, namely a PMF (probability mass function), that is giving a probability of each state (element of S). A common notation is

$$\pi_i^{(t)} = P(X_t = i).$$

For any $t \geq 0$ we also consider $\pi^{(t)}$ as a row vector with coordinates $\pi_i^{(t)}$ for $i \in S$.

Transition of the distribution Suppose we know $\pi^{(0)}$, what can we say about $\pi^{(1)}$, and $\pi^{(t)}$ in general? By law of total probability we have

$$\begin{aligned} P(X_1 = j) &= \sum_{i=1}^{s} P(X_0 = i) \cdot P(X_1 = j \mid X_0 = i) && \text{So, in other notation} \\ \pi_j^{(1)} &= \sum_{i=1}^{s} \pi_i^{(0)} \cdot P_{i,j} && \text{and using matrix multiplication:} \\ \pi^{(1)} &= \pi^{(0)} P \end{aligned}$$

From this we easily get the following theorem:

Theorem 2. *For any Markov chain and any* $k \geq 0$ *we have*

$$\pi^{(k)} = \pi^{(0)} P^k$$

and, more generally, $\pi^{(t+k)} = \pi^{(t)} P^k$.

Proof. By induction. TODO □

k-step transition To look at the above theorem in a different way, we define the following:

$$\begin{aligned} r_{i,j}(k) &= P(\text{we get from } i \text{ to } j \text{ in } k \text{ steps}) \\ &= P(X_k = j \mid X_0 = i) \end{aligned}$$

As we will see, we also have $r_{i,j}(k) = P(X_{t+k} = j \mid X_t = i)$ for any $t > 0$, but this remains to be seen, there may be a dependency on t.

Theorem 3 (Chapman–Kolmogorov). *For any Markov chain and any $k, \ell \geq 0$ we have*

- $r_{i,j}(k) = (P^k)_{i,j}$
- $r_{i,j}(k+\ell) = \sum_{u=1}^{s} = r_{i,u}(k) r_{u,j}(\ell)$
- $r_{i,j}(k+1) = \sum_{u=1}^{s} = r_{i,u}(k) p_{u,j}$

1.2 TODAY

Theorem 4.

$$P(X_0 = a_0 \,\&\, X_1 = a_1 \,\&\, \dots \,\&\, X_t = a_t) = \pi_{a_0}^{(0)} p_{a_0,a_1} p_{a_2,a_3} \dots p_{a_{t-1},a_t}$$

Proof. Chain rule for conditional probability plus Markov property. TODO: write more details □

The next theorem shows a general version of the Markov property (1), that "future is not depending of the history, only of the present".

Theorem 5 (General Markov property). *For any Markov chain and any $t \geq 0$, any $i \in S$ and any*

- *H – event depending only on values of $X_0, \dots, X_{t-1}$,*
- *F – event depending only on values of $X_{t+1}, X_{t+2}, \dots$*

we have

$$P(F \mid H \,\&\, X_t = i) = P(F \mid X_t = i).$$

Proof. (skipped) □

1.3 Classification of states

Definition 6 (Accessible states). *For states i, j of a Markov chain we say that j* is accessible from *i, if starting at i we have nonzero probability of reaching j in the future. For short we write $j \in A(i)$ or $i \to j$. In formula:*

$$j \in A(i) \Leftrightarrow P((\exists t \geq 0) X_t = j | X_0 = i) > 0.$$

It is easy to observe (TODO) that $j \in A(i)$ is equivalent with existence of a directed walk from i to j in the transition graph.

Definition 7 (Communicating states). *We say that states i, j of a Markov chain* communicate *if $i \in A(j)$ and $j \in A(i)$. For short we write $i \leftrightarrow j$.*

Theorem 8. *For any Markov chain the relation $\leftrightarrow$ is an equivalence on the set of states.*

Proof. To show reflexivity, just observe that $i \to i$ holds for any state i, as we are allowed to choose time $t = 0$. Therefore, $i \leftrightarrow j$ as weel.

To show transitivity, assume $i \leftrightarrow j \leftrightarrow k$. Thus we have $i \to j \to k$. Therefore there is a directed walk in the transition digraph from i to j and another one from j to k. Thus, their concatenation is a walk from i to k, showing $i \to k$. By symmetric argument we also have $k \to i$, which finishes the proof. □

1.4 Reducing the MC to smaller ones

Let us consider a MC together with its decomposition to communicating classes.

TODO

- think long-term
- MC moves within one equivalence class for a while, when it leaves it, it never comes back. (Because ...)
- This process goes on, until we get to an equivalence class, that we cannot leave.

transient vs. recurrent states Let $f_{i,j}$ be the probability that we get to j, if we start from i; if $i = j$ we need to get there *again*. In formula,

$$f_{i,j} = P(\exists t \geq 1 : X_t = j \mid X_0 = i)$$

We call $f_{i,i}$ *recurrence probability* and states i such that $f_{i,i} = 1$ are called *recurrent* (sometimes also persistent). The other states (such that $f_{i,i} < 1$) are called *transient*.

To explore this idea a bit further, let $T_i = \min\{t \geq 1 : X_t = i\}$ (or $T_i = \infty$, if there is not such t). Recurrent states are such that T_i is finite.

Also let V_i be the number of visits to i, that is, $V_i := |\{t \geq 0 : X_t = i\}$. It is clear that if i is a recurrent state and $X_0 = i$ then $V_i = \infty$. (After each visit we have probability $f_{i,i} = 1$ that we visit again.) If i is transient, then after each visit to i we have probability $1 - f_{i,i} > 0$ that this is the last visit. Thus (by definition of the geometric distribution)

$$V_i | X_0 = i \sim Geo(1 - f_{i,i}).$$

TODO: is the next theorem trivi for finite MC?

Theorem 9. *Let C be a communicating class. Then either all states in C are recurrent or all are transient.*

Proof. Suppose $i \Leftrightarrow j$, in particular $r_{i,j}(t) > 0$ for some t. Assume that i is recurrent.

Then $f_{i,i} = 1$, thus we visit i infinitely often. In each of these visits we have probability $r = r_{i,j}(t)$ that we visit j in t units of time. Thus the probability that this never happens is $\lim_{n\to\infty}(1 - r)^n = 0$. So we will visit j, starting from i, in symbols $f_{i,j} = 1$.

Suppose $f_{j,i} < 1$. Then with positive probability we never visit i again, a contradiction, as we now we will visit i infinitely often. Thus $f_{j,i} = 1$. This implies, that $f_{j,j} = 1$ as well, as $f_{j,j} \leq f_{j,i} f_{i,j}$. So, j is recurrent as well. □

Suppose a Markov chain is irreducible, so there is a positive probability of moving from i to j for every pair of states. It is tempting to conclude, that all states must be recurrent. Indeed, this is true for finite Markov chains:

Theorem 10. *Let (X_t) be a finite irreducible Markov chain. Then all states are recurrent.*

Proof. In view of the last theorem, the alternative is that all states are transient. This means, no state will be visited infinitely often. So there is time M_1 such that for $t > M_1$ we have $X_1 \neq 1$. Similarly, we define M_i for every state i. But what is the value of X_t for $t > \max\{M_1, \dots, M_s\}$? □

However, this is not necessarily the case, if the Markov chain is infinite. TODO: simple example

In fact, a random walk in $\mathbb{Z}^3$ (or in higher dimensions) has also all states transient, while a random walk in $\mathbb{Z}$ or in $\mathbb{Z}^2$ has all states recurrent. TODO: more details?

1.5 Convergence to stationary distribution

Chapman-Kolmogorov theorem gives us a way how to describe the behaviour of a Markov chain in a short time: If we start with known $\pi^{(0)}$ (distribution if X_0, the state at time 0), we can compute $\pi^{(k)}$. Next, we turn to describing the long-term behaviour.

Convergence to stationary distribution Given a Markov chain with transition matrix $P \in \mathbb{R}^{n \times n}$, we say $\pi \in \mathbb{R}^n$ is a *stationary distribution* if $\pi P = \pi$. In other words, if $\pi^{(0)} = \pi$ then $\pi^{(t)} = \pi$ for all t, which explains the term. For some Markov chains we are guaranteed that the distribution $\pi^{(t)}$ will approach the stationary distribution no matter what is $\pi^{(0)}$.

Theorem 11. *If a Markov chain is finite, aperiodic and irreducible, then*

1. *there is a unique stationary distribution π, and*
2. $\lim_{n\to\infty}(P^n)_{i,j} = \pi_j$.

In other words, regardless of $\pi^{(0)}$ we know what $\pi^{(n)}$ will (approximately) be, if n is large enough.

TODO: def. of irreducible, aperiodic TODO: what for finite TODO: examples, when it fails: periodic states, two components of $\leftrightarrow$, infinite.

1.6 Probability of absorption, time to absorption

Yet another way to look at long-term behaviour of a Markov chain is to study *absorbing states*, states that can never be left. Formally, $a \in S$ is absorbing state if $p_{a,a} = 1$. Not every Markov chain has such state, but for those that do, two natural questions arise: how long (on average) will it take till we reach an absorbing state? And if there is more than one such state, what is the probability of reaching each of them? Both questions are easily answers, if one approaches it right: it is significantly easier to compute these times and probabilities for all states at the same time, than do to it just for one state.

In the following, assume $A \subseteq S$ is a nonempty set of absorbing states; also assume $0 \in A$. For every $i \in S$ we define μ_i to be the expected time to absorption starting from i, formally

$$\mu_i = \mathbb{E}(T \mid X_0 = i), \text{where } T = \min\{t : X_t \in A\}.$$

Further, we let a_i be the probability we end at state 0, starting from i.

$$a_i = P(\exists t : X_t = 0 \mid X_0 = I).$$

Here we tacitly assume that A contains more absorbing states than just 0, otherwise $a_i = 1$.

Theorem 12. *The probabilities a_i are the unique solution to the following system of equations:*

$$\begin{aligned} a_0 &= 1 \\ a_i &= 0 && \textit{for } 0 \neq i \in A \\ a_i &= \sum_{j \in S} p_{i,j} a_j && \textit{otherwise}. \end{aligned}$$

TODO: proof simple by law of total probability.

Theorem 13. *The expected times μ_i are the unique solution to the following system of equations:*

$$\begin{aligned} \mu_i &= 0 && \textit{for } i \in A \\ \mu_i &= 1 + \sum_{j \in S} p_{i,j} \mu_j && \textit{otherwise}. \end{aligned}$$

TODO: proof simple by law of total expectation.
Example: random walk on a path

1.7 Application: algorithm for 2-SAT, 3-SAT

A formula is in conjunctive normal form if it is a conjunction of a list of clauses, each of them is a disjunction of literals (a variable or its negation). If each of the clauses has at most k literals, we say the formula is a k-CNF. It is well known (and discussed in other classes) that 2-SAT has a polynomial-time solution, while 3-SAT is an NP-complete problem. Here, we show how we can apply our knowledge of Markov chains to get a randomized algorithm: algorithm, that is allowed to give a wrong answer, but only with a small probability.

2-SAT algorithm

- Input: 2-CNF φ with variables $x_1, \dots, x_n$
- Output: satisfying assignment or statement that none exists
- arbitrary initialize $x_1, \dots, x_n$
- If $\varphi(x_1, \dots, x_n)$ is true, return $(x_1, \dots, x_n)$. Otherwise, let C be an unsatisfied clause and change random variable in C.
- repeat previous step at most $2mn^2$-times

- say that no solution exists

Theorem 14. *The above algorithm gives wrong answer with probability at most 2^{-m}. The running time is $O(m \cdot n^4)$.*

Proof. For the running time estimate we just notice that φ has $O(n^2)$ clauses of two literals, ignoring possibility for faster search for unsatisfied clause. If φ is not satisfiable, the algorithm never finds a satisfying assignment and thus gives a correct answer. So suppose there is a solution and let $(x_1^*, \dots, x_n^*)$ be one of the (possibly many) solutions. Let D_t be the distance from solution at time t. Explicitly, D_t is the number of i such that $x_i \neq x_i^*$, where value of x_i is considered at time t.

The algorithm does not know about D_t, we only use it for the analysis. Clearly, if $D_t = 0$ at some time t, we have found the solution (and we see that φ is being satisfied, so the algorithm ends).[1] Otherwise let C be the unsatisfied clause. To simplify notation, assume $C = x_1 \vee x_2$. As C is unsatisfied, we have $x_1 = x_2 = F$. As x^* is a satisfying assignment, either x_1^* or x_2^* (or both) are true. Suppose $x_1^* = F$, $x_2^* = T$. Then the algorithm has probability $1/2$ of increasing D_t by one (if we change x_1) and probability $1/2$ of decreasing it. This is certainly independent of anything else, in particular of value of $D_0, \dots, D_{t-1}$.

The issue is that there are two other cases: $x_1^* = T$, $x_2^* = F$ is another good case, it works in the same way. However, if $x_1^* = x_2^* = T$ then this step of algorithm will certainly decrease from the solution: $D_{t+1} = D_t - 1$. While this looks like a good thing, we call this a bad case: we were hoping to use our knowledge of Markov chains to analyse the behaviour of (D_t).

To solve this problem, we (as people analyzing the algorithm) create an auxiliary sequence D'_t. Like D_t, this is a quantity the algorithm does not know about.

- We define $D'_0 = D_0$.
- When the choice of C makes a good case, we make sure that $D'_{t+1} - D'_t = D_{t+1} - D_t$.
- In the bad case we toss a coin to ensure that

  $$P(D'_{t+1} = D_t + x) = 1/2$$

  for both $x = +1$ and $x = -1$.
- If $D'_t = n$ then $D'_{t+1} = n - 1$.

It is easy to see that $0 \leq D_t \leq D'_t$ (actually $D'_t - D_t$ is an even number, so if it is not zero, it is at least 2). Mainly, D'_t is a Markov chain given by the following transition digraph TODO.

Let $T \geq 0$ be the first time such that $D_T = 0$ and, similarly, $T' \geq 0$ first time such that $D'_{T'} = 0$. Using Theorem 13 we find TODO that $\mathbb{E}(T') \leq n^2$. Clearly, $\mathbb{E}(T) \leq \mathbb{E}(T')$, so $\mathbb{E}(T) \leq n^2$. This starts to look useful. But we want to understand

[1] It would be satisfied also for a different satisfying assignment. Think whether it changes anything about the analysis.

how likely it is, that T is much larger. For this we use Markov inequality from the first semester. Using it, we get

$$P(T \geq 2n^2) \leq \frac{\mathbb{E}(T)}{2n^2} \leq \frac{n^2}{2n^2} = \frac{1}{2}.$$

To wrap it up: we divide the $m \cdot 2n^2$ steps into m blocks, each of size $2n^2$. By what we just did, in each block we fail with probability at most $1/2$: as failure means that the algorithm runs without finding a solution, and T is the time till we find the solution.[2]

Each steps are independent: or rather, the probability of failure is $\leq 1/2$ no matter how the previous block has ended. Thus, probability of failure in all m blocks is at most $(1/2)^m$. □

TODO: what if there is a clause with just one variable?

3-SAT algorithm

- Input: 3-CNF φ with variables $x_1, \ldots, x_n$
- Output: satisfying assignment or statement that none exists
- arbitrary initialize $x_1, \ldots, x_n$
- If $\varphi(x_1, \ldots, x_n)$ is true, return $(x_1, \ldots, x_n)$. Otherwise, let C be an unsatisfied clause and change random variable in C.
- repeat previous step at most ???-times
- say that no solution exists

This is the obvious attempt. If we run along with it, again using a Markov chain to analyze the distance from one particular solution, we have new issue: the Markov chain is skewed: in the typical case $D_{t+1} = D_t + 1$ with probability $2/3$, while $D_{t+1} = D_t - 1$ only with probability $1/3$. We can again use Theorem 13 but it gives $\mathbb{E}(T) \leq 2^n$ TODO. And this is no good, as in 2^n steps we can try all possible values of the n variables.

To solve this issue, we take into account the initialization phase. For every i we have $x_i = x_i^*$ with probability $1/2$, so $D_0 \sim Bin(n, 1/2)$. In particular, we have $P(D_0 \leq n/2) \geq 1/2$. In such case we have a decent chance of direct success: $P(D_k = 0 \mid D_0 = k) \geq 1/3^k$: in each step we have a probability $\geq 1/3$ that we choose correct variable to change (in an unsatisfied clause) and thus decrease the distance to solution.

[2] Or a bound on it, as there may be other solutins than x^*.

From this we get that

$$\begin{aligned} P(T \leq n/2) &= \sum_{k=0}^{n/2} P(D_0 = k) P(T \leq n/2 \mid D_0 = k) \\ &\geq \sum_{k=0}^{n/2} P(D_0 = k) \cdot 3^{-k} \\ &\geq P(D_0 \leq n/2) \cdot 3^{-k} \geq \frac{1}{2 \cdot 3^k}. \end{aligned}$$

This leads to a modified algorithm:

Faster 3-SAT algorithm

- Input: 3-CNF φ with variables $x_1, \ldots, x_n$
- Output: satisfying assignment or statement that none exists
- arbitrary initialize $x_1, \ldots, x_n$
- If $\varphi(x_1, \ldots, x_n)$ is true, return $(x_1, \ldots, x_n)$. Otherwise, let C be an unsatisfied clause and change random variable in C.
- repeat previous step at most $n/2$-times, then reinitialize randomly
- repeat previous step at most m-times
- say that no solution exists

Theorem 15. *The above algorithm gives wrong answer with probability at most e^{-t}, where $m = t \cdot 2 \cdot 3^{n/2}$. The running time is $O(n^4 \cdot 3^{n/2})$.*

Proof. We found already that one block of $n/2$ steps succeeds with probability at least $q = \frac{1}{23^k}$. Then we do a new attempt, thus blocks are independent at probability that all of them fail is at most

$$(1-q)^m \leq e^{-qm} = e^{-t}.$$

Here we used the inequality $1 - q \leq e^{-q}$ that is valid for any real q. □

TODO: what if there is a clause with just one variable?

Note that the above algorithm is not optimal. If we let each block run for $3n$ steps, and if we analyze it smarter, we get the running time down to $O(n^3(4/3)^n)$. But even $3^{n/2}$ is much better than the trivial solution $O(2^n)$.

2 Bayesian statistics

2.1 Two approaches to statistics

In the first semester we looked at the *classical (frequentists')* approach to statistics. In this approach:

- Probability is a long-term frequency (out of 6000 rolls of the dice, a six was rolled 1026 times, the ratio converges to the true probability). It is an objective property of the real world.
- Parameters are fixed, unknown constants. We can't make meaningful probabilistic statements about them.
- We design statistical procedures to have desirable long-run properties. E.g. 95 % of our interval estimates will cover the unknown parameter.

Now we are going to look at an alternative, so called *Bayesian approach*:

- Probability describes how much we believe in a phenomenon, how much we are willing to bet:
  (Prob. that T. Bayes had a cup of tea on December 18, 1760 is 90 %.)
  (Prob. that COVID-19 virus did leak from a lab is ?50? %.)
- We can make probabilistic statements about parameters (even though they are fixed constants): the "choice of universe" is the underlying elementary event.
- We compute the distribution of ϑ and form point and interval estimates from it, etc.

2.2 Preliminaries – conditionaly pmf, cdf, etc.

Before we get to the meat of the matter, let us first define/recall the needed definitions. TODO: improve

- pdf f_X is a function such that $P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$.
- Intuition: $f_X(x) = \lim_{t\to 0} \frac{P(x \leq X \leq x+t)}{t}$ – so indeed, it is a "density of probability"
- joint pmf
- conditional pmf $p_{X|Y}(x|y) = p_{X,Y}(x,y)/p_Y(y)$
- joint pdf $f_{X,Y}$ is a function such that $P(X \leq x \,\&\, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dsdt$.
- Intuition: $f_{X,Y}(x) = \lim_{t\to 0} \frac{P(x \leq X \leq x+t \,\&\, y \leq Y \leq y+t)}{t^2}$ – so indeed, it is a "density of probability"
- conditional pdf $f_{X|Y} = f_{X,Y}(x,y)/f_Y(y)$
- marginal pdf $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
- marginal pdf $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$

2.3 Bayesian method – basic description

- The unknown parameter is treated as a random variable Θ
- We choose *prior distribution*, the pmf $p_\Theta(\vartheta)$ or the pdf $f_\Theta(\vartheta)$ independent of the data.
- We choose a statistical model $p_{X|\Theta}(x|\vartheta)$ or $f_{X|\Theta}(x|\vartheta)$ that describes what we measure (and with what probability), depending on the value of the parameter.
- After we observe $X = x$, we compute the *posterior distribution* $f_{\Theta|X}(\vartheta|x)$
- and then derive what we need e.g. find a, b so that $P(a \le \Theta \le b \mid X = x) = \int_a^b f_{\Theta|X}(\vartheta|x) d\vartheta \ge 1 - \alpha$

2.4 Bayes theorem

Theorem 16 (Bayes theorem for discrete r.v.'s)**.** *X, Θ are discrete r.v.'s*

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')p_\Theta(\vartheta')}.$$

(terms with $p_\Theta(\vartheta') = 0$ are considered to be 0).

Theorem 17 (Bayes theorem for continuous r.v.'s)**.** *X, Θ are continuous r.v.'s with pdf's f_X, f_Θ and joint pdf $f_{X,\Theta}$*

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_\Theta(\vartheta)}{\int_{\vartheta' \in \mathbb{R}} f_{X|\Theta}(x|\vartheta')f_\Theta(\vartheta')d\vartheta'}.$$

(terms with $f_\Theta(\vartheta') = 0$ with $f_\Theta(\vartheta') = 0$ are considered 0).

Theorem 18 (Bayes theorem for discrete r.v.'s)**.** *X be discrete and Θ continuous r.v. Then*

$$f_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)f_\Theta(\vartheta)}{\int_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')f_\Theta(\vartheta')}.$$

(terms with $p_\Theta(\vartheta') = 0$ are considered to be 0).

2.5 Bayesian point estimates – MAP and LMS

Even when we know a distribution of a random variable it is unclear what is the best numerical value that represents it. is it the mean (expected value)? Or the mode (moste probable value)? Or the median? It turns out all choices have their justification. In the context of Bayesian statistics, we are interested in a random variable Θ conditioned on the event $X = x$. (You may concentrate on the discrete case, where the conditioning is easy to understand.)

MAP – Maximum A-Posteriori We choose $\hat{\vartheta}$ to maximize

- $p_{\Theta|X}(\vartheta|x)$ in the discrete case
- $f_{\Theta|X}(\vartheta|x)$ in the continuous case
- Essentially, we are replacing the random variable by its mode.
- Similar to the ML method in the classical approach if we choose a "flat prior" – Θ is supposed to be uniform/discrete uniform.

LMS – Least Mean Square Also the conditional mean method.

- We choose $\hat{\vartheta} = \mathbb{E}(\Theta \mid X = x)$, so we replace the random variable by its mean.
- What we get is an Unbiased point estimate that has the smallest possible LMS (least mean square) error:

$$\mathbb{E}((\Theta - \hat{\vartheta})^2 | X = x)$$

- (we will show this later.)

Similarly, if we take median (number m such that $P(\Theta \le m \mid X = x) = 1/2$) then we minimize absolut value of an error $\mathbb{E}((\Theta - \hat{\vartheta})^2 | X = x)$. But we will not pursue this approach further.

2.6 Bayesian inference – examples

2.6.1 Naive Bayes classifier – both Θ and X are discrete

This techniques can be used for any classification of objects into finite number of categories, using finite number of discrete features. For concreteness, we will explain it as a way to test whether some email is a spam or ham (that is, not spam). We let Ω be the set of all emails (together with the probability of receiving each of them). We can't possibly list of elements of Ω, but we consider the emails delivered to our inbox as sampling from this probability space.

Our interest lies in random variable Θ that is equal to 1 for spams and to 2 for hams. (Recall Θ is a function from Ω to $\mathbb{R}$, so for each email $\omega \in \Omega$ we need to define value of $\Theta(\omega)$.) In order to estimate value of Θ, we measure data: a list of Bernoulli variables $X_1, \ldots, X_n$, where $X_i(\omega) = 1$ if ω contains word w_i (and $X_i(\omega) = 0$ otherwise). So we imagine $w_1, \ldots, w_n$ is a list of all words that are useful to detect spams.

By the Bayes theorem we have

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta) p_\Theta(\vartheta)}{\sum_{t=1}^{2} p_{X|\Theta}(x|t) p_\Theta(t)}.$$

TODO finish it

2.6.2 Estimating bias of a coin – Θ is continuous, X is discrete

Consider a loaded coin with probability of heads being ϑ (which we assume to be an evaluation of a random variable Θ). Btw, everything applies to any procedure generating a Bernoulli random variable, but we stick with a coin for concreteness. Our goal is to find out the value of ϑ. In tune with the Bayesian methodology, we start with a prior distribution, that is a pdf f_Θ. (As we want to allow any real number in $[0,1]$ as the value of ϑ, we must take Θ to be a continuous random variable.) Then we take measurements: we choose a number n of coin tosses and check how many heads we get. If we know the value of θ, the distribution of this number (call it X), is clearly $Bin(n,\vartheta)$. So we get

$$p_{X|\Theta}(k|\vartheta) = \binom{n}{k}\vartheta^k(1-\vartheta)^{n-k}.$$

It remains to apply Theorem 18. We still haven't decided what prior to choose though. If we don't known anything (say it is not a real coin but a digital generator), we may take *flat prior* $\Theta \sim U(0,1)$. However, we need something more versatile to allow us to encode some prior knowledge.

Beta distribution It is convenient to use the following type of distribution for Θ:

$$f_\Theta(\vartheta) = \begin{cases} c\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1} & \text{for } 0 < \vartheta < 1 \\ 0 & \text{otherwise} \end{cases}$$

Here c is a normalizing constant that makes the following function a pdf. It is typically written as $1/B(\alpha,\beta)$, the reciprocal of a *beta function*. The r.v. Θ is said to have *beta distribution*. We will collect some useful properties of this distribution. All are easy to verify using basic knowledge of calculus, details are omitted though.

- $f_\Theta(\vartheta)$ is maximal for $\vartheta = \frac{\alpha-1}{\alpha+\beta-2}$ (mode of the distribution). This can be verified by a simple differentiation.
- $\mathbb{E}(\Theta) = \frac{\alpha}{\alpha+\beta}$ (mean of the distribution). This follows from the next part and easy calculation.
- $B(\alpha,\beta) = 1/\binom{\alpha+\beta-2}{\alpha-1}$. This can be shown by per-partes and induction over $\alpha+\beta$.

Now we have all set up to apply Theorem 18. Fortunately, we don't need to compute the integral in the denominator.

$$\begin{aligned} f_{\Theta|X}(\vartheta|k) &= c_1 p_{X|\Theta}(k|\vartheta) f_\Theta(\vartheta) \\ &= c_2\vartheta^k(1-\vartheta)^{n-k}\vartheta^{\alpha-1}(1-\vartheta)^{\beta-1} \\ &= c_2\vartheta^{\alpha+k-1}(1-\vartheta)^{\beta+n-k-1} \end{aligned}$$

The calculation is only valid for $\vartheta \in [0,1]$, otherwise $f_\Theta(\vartheta) = 0$, so the updated (posterior) pdf is also 0. How to find out c_2, if we need to? We use the fact that after conditioning on the event $\{X = k\}$ the random variable Θ still only attains values in $[0,1]$. Thus, c_2 takes such value that makes $f_{\Theta|X}(\vartheta|k)$ a pdf, a function with integral 1. Based on what we learned about Beta distribution, $c_2 = 1/B(\alpha', \beta')$ and $\Theta|X = k$ follows the Beta distribution with parameters $\alpha' = \alpha + k$ and $\beta' = \beta + n - k$.

TODO: wrap up

2.6.3 Estimating normal random variables – both Θ and X are continuous

3 Conditional expectation

We have already learned about expectation $\mathbb{E}(Y)$ of a random variable Y — average value over the whole probability space — and about conditional expectation $\mathbb{E}(Y \mid A)$ — average over a set $A \subseteq \Omega$. In this section we will learn about a related topic, where we will take averages of Y over sets defined by another random variable, X. We will restrict the discussion to the case of a discrete random variable X, the case of continuous X is more subtle. The variable Y can be discrete or continuous.

For any $x \in \mathbb{R}$ we let

$$g(x) := \mathbb{E}(Y \mid X = x).$$

This is obviously some real function of real variable. Next, we plug the random variable X to the function g and we define

$$\mathbb{E}(Y \mid X) := g(X).$$

Thus, $\mathbb{E}(Y \mid X)$ is a random variable. In case of discrete X, it is easy to understand what is going on: on each set of form $A_x = \{X = x\}$ we define $\mathbb{E}(Y \mid X)$ as $\mathbb{E}(Y \mid A_x)$. This leads to the following important observation:

Theorem 19 (Law of Iterated Expectation)**.**

$$\mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(Y)$$

Proof. By law of total probability we have

$$\mathbb{E}(Y) = \sum_x P(A_x)\mathbb{E}(Y \mid A_x)$$

while LOTUS says, that

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(g(X)) = \sum_x g(x)P(X = x),$$

which is the same as the expression above. □

Example 1: coin TODO

Example 2: stick TODO

Example 3; group of students TODO

$\hat{Y} = \mathbb{E}(Y \mid X)$
$\tilde{Y} = Y - \hat{Y}$
expression of $\mathrm{var}(Y)$ – Eve's rule

$$\mathrm{var}(Y) = \mathbb{E}(\mathrm{var}(Y \mid X)) + \mathrm{var}(\mathbb{E}(Y \mid X))$$

Theorem 20 (Conditional expectation gives LMS estimate)**.** *We switch to the statistical practice of using Θ for the parameter we care about, and X for the measured data. Let $\hat{\Theta}$ be any estimator (function of X that we use to estimate Θ). The mean square error*

$$\mathbb{E}((\hat{\Theta}(X) - \Theta)^2 | X)$$

is minimal, if we put $\hat{\Theta}(X) = \mathbb{E}(\Theta \mid X)$.

4 Stochastic processes

A stochastic process is a name for a sequence (or more generally, a collection) of random variables. We have already seen an important case of that when we looked at Markov chains (where we added an important condition, "independence on the past"). Another important example (that we just mention in passing) is the *Wiener process* W_t (here $t \in \mathbb{R}$, so it is not a sequence, but a "continuous time parametrized parameter", in other words a random function of a real variable). These processes are used to model Brownian motion and stock prices, to name a few.

Next, we will look at two models of *arrival times* – time till some random event occurs, you can imagine next email arrival, or next person walking in a store. The first model will consider discrete time, the second one continuous time.

4.1 Bernoulli process

A *Bernoulli process* is an infinite sequence of independent identically distributed Bernoulli trials, i.e., a sequence of RVs $X_1, X_2, \dots$ that are independent and each follows $Ber(p)$ distribution. As such, it is a very simple object. We will look at it from different angles though. As for terminology, we will call the fact $X_k = 1$ a success at time k, or an arrival (of a person/email/…) at time k.

Number of successes We let $N_t = T_1 + \cdots + T_t$ be the number of successes/arrivals up to time t. We know from the first semester that $N_t \sim Bin(t, p)$, so $\mathbb{E}(N_t) = tp$, $\mathrm{var}(N_t) = tp(1-p)$ and $P(N_t = k) = \binom{t}{k} p^k (1-p)^{t-k}$.

Arrival times We let $T = T_1$ be the time of first success/arrival, that is minimal t such that $X_t = 1$. It is easy to see that $T \sim Geom(p)$, thus $\mathbb{E}(T) = 1/p$, $\mathrm{var}(T) = (1-p)/p^2$ and $P(T = t) = (1-p)^{t-1}p$. More generally, we let T_k be the time of the k-th success/arrival. In other words, it is the minimal t such that $N_t = k$. We discuss its properties in a while.

Waiting times/Interarrival times Put $L_k = T_k - T_{k-1}$ (we put $T_0 = 0$ to simplify notation). In words, it is the time we are waiting for the k-th success. TODO memoryless property thus $L_k \sim Geom(p)$, that is $\mathbb{E}(L_k) = 1/p$, $\mathrm{var}(L_k) = (1-p)/p^2$ and $P(L_k = t) = (1-p)^{t-1}p$. Moreover, $L_1, L_2, \dots$ are independent.

Properties of T_k We can see that $T_k = L_1 + \cdots + L_k$. Thus by linearity, we have $\mathbb{E}(T_k) = k/p$. As the interarrival times are independent, we have $\mathrm{var}(T_k) = k(1-p)/p^2$. The PMF of T_k can be obtained by a careful application of the convolution formula. However, it is more convenient to derive from first principles: The fact $T_k = t$ means that $X_t = 1$ and there are exactly $k-1$ times $\tau < t$ where $X_\tau = 1$. This, together with independence of all the Bernoulli variables, implies

$$P(T_k = t) = \binom{t-1}{k-1} p^{k-1}(1-p)^{t-k}, \text{for } t \geq k,$$

clearly $P(T_k = t) = 0$ otherwise. This distribution is called *Pascal distribution of order k*. A related distribution is that of random variable $T_k - k$, the number of failures before k-th success. This is called *negative binomial distribution* due to its pmf that can be written as $(-1)^t\binom{-k}{t}(1-p)^t p^k$ TODO WRITE CORRECTLY

Alternative description Note that we can equivalently describe the situation by the interarrival times, that is by the sequence of i.i.d. random variables $L_1, L_2, \dots \sim Geom(p)$. Then we put $T_k = L_1 + \cdots + L_k$ and $X_k = 1$ whenever $T_t = k$ for some t. It is easy to see that this is an equivalent description, in other words, the sequence $X_1, X_2, \dots$ is a Bernoulli process.

Example: The number of days till the next rain follows the $Geom(p)$ distribution. (We assume each day is either rainy/not rainy, that is we have no finer distinction.) What is the probability that it will rain at days 10 and 20?

This seems very complicated and tedious. However, by the indicated description of Bernoulli process by interarrival times, the indicator variables of rain form a Bernoulli process. And the probability of rain at days 10 and 20 is

$$P(X_{10} = 1 \,\&\, X_{20} = 1) = P(X_{10} = 1) \cdot P(X_{20} = 1) = p \cdot p = p^2.$$

Merging of Bernoulli processes Consider two independent Bernoulli processes $(X_i) \sim Bp(p)$ and $(Y_i) \sim Bp(q)$. Put $Z_i = X_i \vee Y_i$. Then $(Z_i) \sim Bp(p+q-pq)$.

This is obvious as $P(Z_i = 1) = P(X_i = 1 \vee Y_i = 1)$ and we use the basic formula for probability of a union. However, from the point of view of arrival and/or waiting times this is nontrivial: Suppose time to a rainy day follows $Geom(p)$ and time to a snowy day follows $Geom(q)$. Then time to a day when it rains or snows follows $Geom(p+q-pq)$.

Splitting of Bernoulli processes Let $(Z_i) \sim Bp(p)$. If $Z_i = 0$, we put $X_i = Y_i = 0$. If $Z_i = 1$, we with probability q put $(X_i, Y_i) = (1, 0)$, and otherwise $(X_i, Y_i) = (0, 1)$. (Example to imagine: $Z_i = 1$ means we get a message, $X_i = 1$ it was from Ann, $Y_i = 1$ it was from Bob.) Then $(X_i) \sim Bp(pq)$ and $(Y_i) \sim Bp(p(1-q))$.

(Question: are the X_is and Y_is independent?)

4.2 Poisson process

Continuous-time version of Bernoulli processes. Assume, we want to deal with events that occur more often than once per day. We can stay with discrete time and measure it in hours, second, or nanoseconds. But instead of that, we will define a more elegant description that will allow any real values of the arrival times.

As for Bernoulli process, we will describe the process by several random variables:

- $T_1, T_2, T_3, \dots$ are the times of individual arrivals (events we want to describe), or *arrival times* for short
- $N((a, b])$ is the number of arrivals at time in $(a, b]$.
- $N_t = N((0, t])$
- $L_k = T_k - T_{k-1}$ – *waiting times* for next arrival

What is different now is that we do not have the underlying sequence of "coin tosses", that we denoted $X_1, X_2, \dots$ above. To derive the properties of this process we start with three axioms – that we pose as a natural "limit" version of the Bernoulli process for a very small time intervals.

(a) We are describing times of "arrival" in interval $[0, \infty)$. For any time interval $(a, b]$ we let $N((a, b])$ be the number of arrivals in this intervals. We postulate that the pmf of this random variable only depends on $\tau = b - a$. We denote $P(N((a, b]) = k)$ as $P(k, \tau)$.

(b) $N((a, b])$ and $N((0, a])$ are independent.

(c) For small values of τ we have the following approximation, for some $\lambda > 0$

- $P(0, \tau) = 1 - \lambda\tau + o(1)$
- $P(1, \tau) = \lambda\tau + o(1)$
- $P(k, \tau) = o(1)$ for $k > 1$

TODO: explain how this follows from approximating a Bernoulli process

From these axioms we derive (TODO) the following properties:

- $N_t \sim Pois(\lambda t)$
- $L_k \sim Exp(\lambda)$
- For any sequence $0 <= t_0 < t_1 < \cdots < t_k$ the random variables $N((t_{i-1}, t_i])$ for $i = 1, \dots, k$ are independent and the i-th of them follows $Pois(\lambda(t_i - t_{i-1}))$

From the distribution of L_ks we get information about the distribution of T_ks:

- First, $\mathbb{E}(T_k) = \mathbb{E}(L_1) + \cdots + \mathbb{E}(L_k) = k/\lambda$ by linearity of expectation.

- Next, $\mathrm{var}(T_k) = \mathrm{var}(L_1) + \cdots + \mathrm{var}(L_k) = k/\lambda^2$ by formula for variance of independent RVs.
- Finally, we can find the pdf of T_k, so-called *Erlang distribution of order* $k-1$

  $$f_{T_k}(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}.$$

  One possible proof: by induction on k. For $k = 1$ this is the pdf of $Exp(\lambda)$, as it should be. Then we use convolution formula to get from pdf of T_{k-1} to T_k:

  $$f_{T_k}(t) = \int_0^t f_{T_{k-1}}(s) f_{L_k}(t-s) ds$$

  Another one: use the formula $P(t \le T_k \le t + \delta) \approx \delta f_{T_k}(t)$ (TODO: more precisely) together with the expression

  $$\begin{aligned} P(t \le T_k \le t+\delta) &= P(k-1 \text{ arrivals in } [0,t]) P(\text{at least 1 arrival in } [t, t+\delta]) \\ &= e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} (1 - e^{-\lambda \delta}) \\ &\approx e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} \lambda \delta \end{aligned}$$

Splitting of Poisson processes Consider a sequence of arrival times $T_1, T_2, \ldots$ of a Poisson process of intensity λ. For each i we toss independently a coin and with probability p classify the arrival as type-1, with probability $1-p$ a type 2. We let $T'_1, T'_2, \ldots$ be the times of the arrivals of type 1 and $T''_1, T''_2, \ldots$ the times of the arrivals of type 2. Then $(T''_k)_k$ are arrival times of a Poisson process of intensity λp and Then $(T''_k)_k$ are arrival times of a Poisson process of intensity $\lambda(1-p)$. Moreover, these processes are independent. (TODO: explain precise meaning).

TODO: explain why it is so.

Example: customers buying a book? Emails arriving important or not.

Merging of Poisson processes Consider two Poisson processes: one with intensity λ, the other with intensity λ'. Then their merging is a Poisson process of intensity $\lambda + \lambda'$.

5 Balls and bins

Birthday paradox We start with a simple, but illustrative example: in a group of k people, what is the probability that two celebrate their birthdays in the same day? (Ignore leap years, twins, and irregularities of birthrate during the year.) Obviously, the probability that no such coincidence happens is exactly

$$(1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{k-1}{365}).$$

For small values of k we may use the well-known approximation $e^{-x} \doteq 1-x$ – but suprisingly we use it to approximate $1-x$. The expression above is close to

$$e^{-\frac{1}{365}} \ldots e^{-\frac{k-1}{365}} = e^{-\sum_{i=1}^{k-1} \frac{i}{365}} = e^{-\frac{k(k-1)}{2\cdot 365}}.$$

We can conclude that for $k^2 \doteq 2 \cdot 365$ the probability of no birthday "collision" is approximately $1/e$. Btw $\sqrt{730} \doteq 27$ and the exact formula for $k = 27$ gives XXX, so we see our approximations were pretty good. TODO When is the prob one half.

Balls and bins model Next, we describe an abstract model, that not only generalizes the above exercise, but mainly is used to analyze many random algorithms, some of which we will see below. We will be throwing m balls randomly into n bins. Each ball is thrown independently, and each bin has the same probability of being hit (also no ball ends up outside the bins). In this setting, if we put $n = 365$ and $m = k$, we have our birthday paradox problem again; now it is equivalent to asking, what is the probability that some bin will end up with at least two balls.

Let us look at some more easy questions to ask about this model:

- What is the number of balls in the first bin? (Or any fixed bin, really.)
  Obviously, it is a random variable. By recalling the definitions, we see that it follows the binomial distribution $Bin(m, 1/n)$. This is all we can say about this number – and all we need to answer further questions: e.g., the probability the first bin is empty is $\binom{m}{0}(1-1/n)^m \doteq e^{-m/n}$ (using again the approximation $1-x \doteq e^{-x}$).

- How many bins are empty (on average)?
  Using the previous item and linearity of expectation, this number is equal to $n(1-1/n)^m \doteq ne^{-m/n}$.

- What is maximal amount of balls in a bin?
  This is a harder problem, so let us first show why to care about it.

Application 1: hashing We seek a data structure to store strings and later answer membership question (has 'Cat' been stored?). We use a hash function h that assigns to every string an integer in $[n] = \{1, \ldots, n\}$. We assume h is "sufficiently random". This may be confusing, as h is a deterministic function (and we need it to give the same answer to each string when running next time). What we want is that for typical input strings s_1, …, s_n the hashes $h(s_1)$, …, $h(s_n)$ are independent and uniformly distributed in $[n]$.

To proceed with our data structure: we have n linked lists B_1, …, B_n, initially empty. We store string s to $B_{h(s)}$ in constant time. We look for s in $B_{h(s)}$, which takes time proportional to the length of this list. Our goal is to estimate the worst seek-time, which is (proportional to) the maximum size of a bin, so-called *max-load*. We must be careful what we ask for though: the worst time in the worst case is n, as we may get all balls in the same bin (i.e., all words can have the same hash). This is very unlikely though, with probability $1/n^M$. More precise result in this direction is the following upper bound.

Theorem 21. *For large enough n we have*

$$P(maxload \geq \frac{3 \log \log n}{\log n}) \leq \frac{1}{n}.$$

Proof. **Claim:** For any i, $P(|B_i| \geq M) \leq \binom{n}{M}\frac{1}{n^M}$.

To prove this, we use *union bound*: we first write event "$|B_i| \geq M$" as a union: for every set $S \subseteq [n]$ of size M we consider event $A_S =$ "all balls in S end in bin $B_i{}''$. Obviously, $P(A_S) = 1/n^M$. Also, "$|B_i| \geq M$" is simply $\bigcup_S A_S$ (we take union over all sets S of size M). So we get

$$P(\mid B_i) \geq M) = P(\bigcup_S A_s) \leq \sum_S P(A_S) = \sum_S \frac{1}{n^M} = \binom{n}{M}\frac{1}{n^M}.$$

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

Definition of binomial coefficient and Stirling-type estimate of factorial.

Claim: $P(maxload \geq M) \leq \sum_{i=1}^{n} P(\mid B_i) \leq n\left(\frac{e}{M}\right)^M$

We use again the union bound: event "

$$maxload \geq M$$

" is the same as "$\exists i : |B_i| \geq M$", which is a union of events: $\cup_{i=1}^{n}\{\omega : |B_i| \geq M\}$. (To test your understanding: what does ω stand for here?)

The rest is straightforward estimate: TODO □

Later we will see that the bound for maxload we just got is best possible, up to a multiplicative factor.

Application 2: Bucketsort We want to sort $n = 2^k$ input numbers as fast as possible. We will be assuming the inputs are l-bit integers, thus elements of $I = \{0, \ldots, 2^\ell - 1\}$. Crucially, we will also assume the inputs are uniformly random in this set and mutually independent.

For $x \in I$ we let $b(x)$ be the top k bits, thus $b(x) = \lfloor x/2^{\ell-k} \rfloor$ (or $b(x) = x >> (l-k)$ TODO: different typeset).

Bucketsort algorithm proceeds as follows:

1. Initialize n empty *buckets* – linked lists.
   For $i = 1, \ldots, n$: put input x_i to bucket $B_{b(x_i)}$.

2. For $j = 0, \ldots, n-1$: sort bucket B_j by bubblesort.

3. Join buckets $B_0, \ldots, B_{n-1}$.

Obviously, steps 1 and 3 take linear time (and we cannot do better). The interesting part is to analyze, how long step 2 takes. We let $X_j = |B_j|$. For each input to the algorithm, this will be a particular integer. However, we analyze the running time in avarage, on a random input. Thus we treat X_j as a random variable. As we saw before,

$X_j \sim Bin(n, 1/n)$. The running time of bubblesort is quadratic, so total running time of step 2 is

$$\sum_{j=0}^{n-1} \mathbb{E}(X_j^2).$$

The easiest way to compute the expectation is by using the formula

$$\mathrm{var}(X_j) = \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2$$

in reverse: we already know that $\mathbb{E}(X_j) = n \cdot \frac{1}{n} = 1$ and $\mathrm{var}(X_j) = n \cdot \frac{1}{n} \cdot (1 - \frac{1}{n}) \leq 1$. Thus $\mathbb{E}(X_j^2) = \mathrm{var}(X_j) + \mathbb{E}(X_j)^2 \leq 1 + 1 = 2$. So the expected running time of step 3 is at most $2n$.

Poisson approximation Next, we will prove the "likely lower bound" for maxload:

Theorem 22. *For large enough n we have*

$$P(maxload \leq \frac{\log \log n}{\log n}) \leq \frac{1}{n}.$$

In contrary with the upper bound, this will require quite a bit of prep work. We will invoke the magic of Poisson random variables to help us with this estimate. We will also need to set up some notation. We will use X_i (or $X_i^{(m)}$) for the number of balls in bin i when m balls are being thown. We already know that each X_i follows the $Bin(m, 1/n)$ distribution, which is well approximated by $Pois(m/n)$. Thus, with a leap of faith we let $Y_1, \ldots, Y_n$ be i.i.d. random variables, each with distribution $Pois(m/n)$. We will call the variables $X_1, \ldots, X_n$ the *exact case* and their approximation $Y_1, \ldots, Y_n$ the *Poisson case*. Note, that while $Y_1, \ldots, Y_n$ are independent, the $X_1, \ldots, X_n$ are definitely not! In fact, we have already met their distribution, it is the multinomial distribution and satisfies

$$P(\vec{X} = \vec{x}) = P(X_1 = x_1, \ldots, X_n = x_n) = \binom{m}{x_1, \ldots, x_n} \frac{1}{n^m},$$

where the multinomial coeeficient is defined by $\binom{m}{x_1, \ldots, x_n} = \frac{m!}{x_1! \ldots x_n!}$. The formula above is only true if $\sum_i x_i = m$, otherwise $P(\vec{X} = \vec{x}) = 0$. Thus, the distribution of $\vec{X}$ is definitely distinct from that of $\vec{Y}$: for instance we can have $Y_i = 0$ for each i with nonzero probability, while the probability of this is 0 in the exact case. However, this is in some sense the only thing distinguishing the exact and Poisson cases.

Observation 23. *The distribution of $\vec{X}$ is the same as that of $\vec{Y}$, given that $\sum_i Y = m$. Formally,*

$$P(\vec{X} = \vec{x}) = P(\vec{Y} = \vec{x} | \sum_{i=1}^{n} Y_n = m).$$

Proof. Both probabiities are clearly 0 if $\sum_i x \neq m$. Otherwise, we compute TODO □

Thus, we can simulate the balls&bins process just by using independent Poisson variables. However, the conditioning makes computations complicated. The real magic comes in the next theorem, where we study what happens when we truly embrace the Poisson case of independent Poisson variables.

Theorem 24. *Let $f : \mathbb{Z}^n \to [0, \infty)$ be any function. With the notation as above, we have*

$$\mathbb{E}(f(\vec{X})) \leq e\sqrt{m}\mathbb{E}(f(\vec{Y})).$$

Moreover, if the left-hand-side is monotone in m (the number of balls), we can replace $e\sqrt{m}$ by 2.

Corollary 25. *Let A be any event expressed in terms of sizes of the bins, so $A \subseteq \mathbb{Z}^n$. Then the probability that A happens in the exact case is less or equal that the probability it happens in the Poisson case times a factor $e\sqrt{m}$. Formally,*

$$P(\vec{X} \in A) \leq e\sqrt{m}P(\vec{Y} \in A).$$

Proof. (Theorem implies Corollary) It is enough to let $f(a) = 1$ if $a \in A$ and $f(a) = 0$ otherwise (f is a characteristic function of A. Then $\mathbb{E}(f(\vec{X})) = P(\vec{X} \in A)$ and similarly for $\vec{Y}$. Thus the corollary follows. □

Proof. (of Theorem)

Let $Y = \sum_{i=1}^{n} Y_i$ (do not confuse with $\vec{Y}$!). By law of total expectation (using decomposition $\Omega = \cup_{y=0}^{\infty}\{Y = y\}$) we get the following

$$\begin{aligned}
\mathbb{E}(f(\vec{Y})) &= \sum_{y=0}^{\infty} P(Y = y)P(\mathbb{E}(f(\vec{Y}))|Y = y) \\
&\geq P(Y = m)P(\mathbb{E}(f(\vec{Y}))|Y = m) \\
&= P(Y = m)P(\mathbb{E}(f(\vec{X})))
\end{aligned}$$

The inequality is clear (all terms in the sum are nonnegative) and the equality in the last row follows from the Observation above. It remains to recall that sum of Poisson random variables is again Poisson and thus $Y \sim Pois(m)$. So $P(Y = m) = e^{-m}\frac{m^m}{m!}$. Now we use an estimate for factorial: $m! \leq (m/e)^m e\sqrt{m}$ and we are done.

Notes: for the extended version with monotone left-hand side we replace the $\sum_{y=0}^{\infty}$ by $\sum_{y=0}^{m}$ or $\sum_{y=m}^{\infty}$ (base on wheter the LHS is decreasing or increasing). □

Now we test our technique by estimating probability that maxload is low – Theorem 22. We let $M = \log n / \log\log n$. The probability in the Poisson case can be estimated as follows:

$$\begin{aligned}
P(\max_i Y_i < M) &= P(\forall i : Y_i < M) && \text{property of } \max \\
&= \prod_i P(Y_i < M) && \text{independence of } Y_i\text{s} \\
&\leq \prod_i (1 - P(Y_i = M)) && \text{we give away a lot here} \\
&= \prod_i (1 - e^{-1}\frac{1^M}{M!}) && \text{def. of Poisson} \\
&= \big(1 - \frac{1}{eM!}\big)^n \\
&\leq e^{-\frac{1}{eM!}} && 1 - t \leq e^{-t} \text{ again}
\end{aligned}$$

By Corollary 25, the probability in the exact case can be estimated as

$$P(\max_i X_i < M) \leq e\sqrt{n}e^{-\frac{1}{eM!}}.$$

So it remains to show, that $e\sqrt{n}e^{-\frac{1}{eM!}} < \frac{1}{n}$ (for large enough n). To do this, we will show that $e^{-\frac{1}{eM!}} < \frac{1}{n^2}$. TODO

6 Permutation test

How to compare two random variables, if their distribution can be arbitrary? For a concrete example, suppose we want to compare two gadgets by looking at their ratings. If the gadgets have similar features and price, perhaps this is the best way to decide – so if one gadget has average rating 4.1, then it surely is better than another one of rating 3.9, right? But wait, what about randomness – what if the gadgets are exactly equal, they still won't receive exactly the same rating, with high probability. So what how to decide what deviation we can priradit to randomness, and what is a mark of true difference?

–> compare other possibilities

–> describe permutation test

Wilcoxon signed rank test https://stats.stackexchange.com/questions/348057/wilcoxon-signed-rank-symmetry-assumption

7 Moment Generating Functions and their applications

In this section we will meet an old friend from combinatorics – generating functions. We will see how they can be applied to study random variables and help us to prove two important results from the first semester – Central Limit Theorem and Chernoff bound.

Definition 26 (MGF). *A* moment generating function *for a random variable X is the function*

$$M_X(s) := \mathbb{E}(e^{sX}).$$

Observation 27.

- $M_X(0) = 1$
- $\lim_{s\to-\infty} M_X(s) = P(X = 0)$

Example 28. *Let $X \sim Ber(p)$. Then*

$$M_X(s) = \mathbb{E}(e^{sX}) = (1-p)e^{s\cdot 0} + pe^{s\cdot 1} = pe^s + 1 - p.$$

Theorem 29.

$$M_X(s) = \sum_{k\geq 0} \mathbb{E}(X^k)\frac{s^k}{k!}$$

Proof.

$$\begin{aligned} M_X(s) &= \mathbb{E}(e^{sX}) && \text{by definition} \\ &= \mathbb{E}(\sum_{k\geq 0}\frac{(sX)^k}{k!}) && \text{by Taylor expandion of exponential} \\ &= \mathbb{E}(\sum_{k\geq 0} X^k\frac{s^k}{k!}) \\ &= \sum_{k\geq 0}\mathbb{E}(X^k)\frac{s^k}{k!}) && \text{by linearity of expectation} \end{aligned}$$

TODO: care must be taken, as we are using the linearity for infinite sum. □

The theorem above explains the name of M_X: the coefficient of s^k is the *k-th moment*, that is the value $\mathbb{E}(X^k)$ (divided by $k!$), so M_X can be thought as a GF for this sequence of numbers of interest. We can sometimes use this to compute the moments easily:

Example 30. *Let $X \sim Exp(\lambda)$. Then*

$$M_X(s) = \mathbb{E}(e^{sX}) = \int_0^\infty e^{sx}\lambda e^{-\lambda x}dx = \lambda\int_0^\infty e^{(s-\lambda)x}dx = \frac{\lambda}{\lambda - s}$$

if $s < \lambda$, while $M_X(s) = \infty$ otherwise. We can now expand this function as a power series:

$$M_X(s) = \frac{1}{1 - s/\lambda} = \sum_{k\geq 0}\frac{s^k}{\lambda^k}.$$

This and Theorem 29 shows, that $\mathbb{E}(X^k) = k!/\lambda^k$.

Theorem 31. $M_{aX+b}(s) = e^{sb}M_X(as)$

Proof. TODO □

Theorem 32. *If X, Y are independent, then $M_{X+Y} = M_X M_Y$.*

Proof. TODO □

MGFs do uniquely determine the distribution of corresponding random variable:

Theorem 33. *Suppose for some $\varepsilon > 0$ two MGFs are equal on $(-\varepsilon, \varepsilon)$, that is for random variables X, Y we have*

$$M_X(s) = M_Y(s) \qquad \text{for all } s \in (-\varepsilon, \varepsilon).$$

Then $F_X = F_Y$.

(No proof.) (Note: we cannot hope for any stronger result than equality of CDFs, we certainly cannot expect $X = Y$, for instance!) We also have the following limit version that will be used later.

Theorem 34. *Suppose for some $\varepsilon > 0$ and for random variables Z, Y_1, Y_2, $\ldots$ we have*

$$M_Z(s) = \lim_{n\to\infty} M_{Y_n}(s) \qquad \text{for all } s \in (-\varepsilon, \varepsilon).$$

Then $F_Z(x) = \lim_{n\to\infty} F_{Y_n}(x)$ provided F_Z is continuous.

Example 35. *If $X \sim Pois(\lambda)$ then by definition*

$$M_X(s) = \sum_{k\geq 0} e^{sk} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda(e^s-1)}.$$

If $Y \sim Pois(\mu)$ we get $M_Y(s) = e^{\mu(e^s-1)}$. By Theorem 32 we get $M_{X+Y}(s) = e^{\lambda(e^s-1)}e^{\mu(e^s-1)} = e^{(\lambda+\mu)(e^s-1)}$ and Theorem 33 we see that $X + Y \sim Pois(\lambda + \mu)$.

Example 36. *Let $X = X_n \sim Bin(n, p)$. We know that X is a sum of n independent $Ber(p)$, thus $M_X(s) = (1 - p + pe^s)^n$. (This can also be verified independently from the definition.) Let $p = \lambda/n$, so $M_{X_n}(s) = (1 + \frac{\lambda(e^s-1)}{n})^n$ and we can see that $\lim_{n\to\infty} M_{X_n}(s) = e^{\lambda(e^s-1)}$. By Theorem 34 this shows that $Bin(n, \lambda/n)$ converges in distribution to $Pois(\lambda)$.*

Example 37. *Let $X \sim N(0, 1)$. Then*

$$\begin{aligned} M_X(s) = \mathbb{E}(e^{sX}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-s)^2/2} e^{s^2/2} dx \\ &= e^{s^2/2}. \end{aligned}$$

As $e^{s^2/2} = 1 + s^2/2 + \frac{(s^2/2)^2}{2!} + \ldots$, this gives a formula for all moments of standard normal distribution.

7.1 Proof of CLT

First, we recall the statement of CLT:

Theorem 38. *Let $X_1, X_2, \dots$ be i.i.d. RVs with mean μ and variance $\sigma^2 > 0$. Define*

$$Y_n = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma}.$$

Then $Y_n \xrightarrow{d} N(0,1)$.

Proof. We may assume that $\mu = 0$, otherwise put $X_n' = X_n - \mu$; this does not change σ. We compute first few terms of the MGF of X_n: $M_{X_i}(s) = 1 + \sigma^2 s^2/2 + O(s^3)$. This gives the formula for the MGF of Y_n:

$$M_{Y_n}(s) = M_X(\frac{s}{\sigma\sqrt{n}})^n = \big(1 + \frac{1}{2}(\frac{s}{\sqrt{n}})^2 + O(s^3)\big)^n \to e^{s^2/2}.$$

TODO: add more details □

7.2 Chernoff inequality

Theorem 39. *Suppose i.i.d. $X_1, \dots, X_n$ are ± 1, each with probability $1/2$. Put $X = \sum_i X_i$. We have $\sigma^2 = n$. Then for any $t > 0$ we have*

$$P(X \geq t) = P(X \leq -t) \leq e^{-t^2/2\sigma^2}.$$

Proof. For any $s > 0$ we have

$$P(X \geq t) = P(e^{sX} \geq e^{st}) \leq \frac{\mathbb{E}(e^{sX})}{e^{st}}$$

by Markov inequality. The numerator of the last term is $M_X(s)$, so let us estimate this. By Theorem 31 we have

$$M_X(s) = M_{X_1}(s) M_{X_2}(s) \dots M_{X_n}(s) = M_{X_1}(s)^n.$$

By definition,

$$M_{X_1}(s) = \frac{e^s}{2} + \frac{e^{-s}}{2} = \sum_{k \geq 0} \frac{s^{2k}}{(2k)!}.$$

This can be estimated from above by $e^{s^2/2}$ by looking at each term separately: for every k we have

$$\frac{s^{2k}}{(2k)!} \leq \frac{(s^2/2)^k}{k!}$$

and the terms on the right hand side sum up to $e^{s^2/2}$.

Thus we have $M_X(s) \leq e^{ns^2/2}$ and we have the bound

$$P(X \geq t) \leq e^{ns^2/2 - st}.$$

We optimize this bound by choosing $s = t/n$ (which is positive, so we can do that) and we get the desired bound. □

TODO: compare Chernoff and CLT.

7.3 Applications of Chernoff

Fair coin Let H be the number of Heads we get in n throws of a fair coin, let $X = 2H - n = 2(H - n/2)$. We expect H to be rather close to $n/2$, thus X to be rather small, but how small exactly? Exact answer is given by CDF of $Bin(n, 1/2)$, but Chernoff has a convenient estimate:

$$P(|X| > t) \le 2e^{-t^2/2n}.$$

So if $t = \sqrt{2n \ln n}$, we have the above probability at most $2/n$.

Set Balancing Consider sets $S_1, \ldots, S_n \subseteq [m]$. We want to find a set $T \subseteq [m]$ that divides each of S_i as fairly as possible. (Application: design of statistical experiments.) Specifically, we want to minimize the *discrepancy* $\max_i \big||S_i \cap T| - |S_i \setminus T|\big|$. We can design various algorithms to do that, but a simple argument gives us a solution with discrepancy at most $\sqrt{4m \ln n}$: we choose T as a random subset of $[m]$, with each element having probability $1/2$ for being selected. We will show that probability of discrepancy larger than $d = \sqrt{4m \ln n}$ is at most $2/n$.

We can ignore sets for which $|S_i| \le d$. (TODO: is this needed?) If $|S_i| = k \ge d$, we express $X = |S_i \cap T| - |S_i \setminus T|$ as a sum $X = \sum_j X_j$ where $X_j = +1$ if the j-th element of S_i is selected to T, and $X_j = -1$ otherwise. By Chernoff bound (Theorem 39) we have

$$P(X \ge d) \le e^{-d^2/2k} \le e^{-4m \ln n/(2m)} = 1/n^2.$$

Thus $P(|\,X) \le 2/n^2$.

For our next application we will need a modified version of Chernoff bound:

Theorem 40. *Suppose $X_1, \cdots, X_n$ are independent random variables taking values in $\{0, 1\}$ (not necessarily identically distributed). Let $X = \sum_i X_i$ and $\mu = \mathbb{E}(X)$.*

$$\begin{aligned}
\Pr(X \ge (1+\delta)\mu) &\le e^{-\delta^2\mu/(2+\delta)}, && 0 \le \delta, \\
\Pr(X \le (1-\delta)\mu) &\le e^{-\delta^2\mu/2}, && 0 < \delta < 1, \\
\Pr(|X - \mu| \ge \delta\mu) &\le 2e^{-\delta^2\mu/3}, && 0 < \delta < 1.
\end{aligned}$$

Balls-and-bins revisited Recall our model of putting m into n bins, now with $m \gg n$. We will be again interested in variable $maxload = \max X_i$, where X_i is the number of balls in the i-th bin. Obviously, $\mathbb{E}(X_i) = m/n$. We know that $X_i \sim Bin(m, 1/n)$ but we will not use this. Put $\delta = 1$. Then $P(X_i \ge 2m/n) \le e^{-\frac{m}{3n}}$. Thus, if $m \gg n$, this probability is $o(1/n)$, thus also

$$P(maxload \ge 2m/n) \le \sum_i P(X_i \ge 2m/n) = o(1).$$

Note that this is in strong contrast to our analysis in the case $m = n$.

TODO: more versions of Chernoff: `https://en.wikipedia.org/wiki/Chernoff_bound`