Lecture 12. Stationary distribution. Bins and Balls. Bucket sort

Martin Klazar

December 22, 2020

Stationary distribution

In this lecture every Markov chain $M = (X_0, X_1, \dots)$ is finite and has the set of states $S := [n] = \{1, 2, \dots, n\}$, $n \in \mathbb{N}$. The evolution of M is determined by (i) the initial *distribution*

$$\overline{p} = (p_1, p_2, \dots, p_n) \in \mathbb{R}^n_{\geq 0},\ p_i = \Pr(X_0 = i)\ ,$$

satisfying $\sum_i p_i = 1$, which is the starting distribution of probabilities on the states of M, and (ii) the stochastic *transition matrix*

$$P = (p_{i,j})^n_{i,j=1} \in \mathbb{R}^{n\times n}_{\geq 0}$$

satisfying $\sum_j p_{i,j} = 1$ for every $i \in [n]$, which records transition probabilities between the states: $p_{i,j}$ is the conditional probability that M evolves in one step from the state i to the state j, so $p_{i,j} = \Pr(X_{t+1} = j \mid X_t = i)$ for every $t \in \mathbb{N}_0$ if the right side is defined.

Exercise (on matrices). If in matrices $A \in \mathbb{R}^{k\times l}$ and $B \in \mathbb{R}^{l\times m}$ every row sum equals 1, then so does in their product AB. □

This part of the lecture will actually be more a lecture in linear algebra or linear programming than a lecture in the probabilistic method. Our goal is to show that for any P satisfying certain obvious necessary conditions there exists a unique *stationary distribution* $\overline{p}$ that is moreover *attractive*. A distribution $\overline{p}$ is stationary if it is a fixed point in the evolution of M,

$$\overline{p}P = \overline{p}$$

for $\overline{p}$ written as a $1 \times n$ row vector. Attractivity of $\overline{p}$ means that M evolves towards it from any start, for every distribution $\overline{q}$ we have

that, coordinatewise,

$$\lim_{n\to\infty} \overline{q}P^n = \overline{p}\ .$$

As before we associate with P the weighted *transition digraph*

$$D_P = ([n],\ E,\ h)$$

where $e = (i,j) \in E$ iff $p_{i,j} > 0$, and the edge (arrow) e has weight $h(e) = p_{i,j}$. In any digraph $D = (V, E)$ (so $E \subset V \times V$), a *walk* w is any sequence of vertices

$$w = (v_0,\ v_1,\ \ldots,\ v_n) \subset V,\ n \in \mathbb{N}_0\ ,$$

such that $(v_{i-1}, v_i) \in E$ for every $i = 1, 2, \ldots, n$. If $v_0 = v_n$, we call w a *cycle*. We write $|w| := n$ for the *length* of a walk or a cycle. If $v_0 = u$ and $v_n = v$, we call w a *u-v walk*. It follows that in any u-v walk w with minimum length no vertex is repeated and therefore $|w| \le |V| - 1$. We call such w a *shortest path from u to v*.

We may think of M in terms of the following game in a sandbox. The given distribution $\overline{p}$ distributes one kilogram of sand into heaps p_i at the vertices $i \in [n]$ of D_P, and the matrix P is a redistribution rule. In one step we simultaneously move, according to the arrows

$$\overset{i}{\bullet} \xrightarrow{p_{i,j}} \overset{j}{\bullet}$$

of D_P, for every vertex $i \in [n]$ and any $j \in [n]$ the fraction $p_{i,j} = h(i,j)$ of the heap p_i (that is, the amount $p_i p_{i,j}$) to the vertex j. By this the old heap p_i disappears (as $\sum_j p_{i,j} = 1$) but a new heap p'_j is formed at each vertex $j \in [n]$. In this way we redistribute the sand to the vertices and get its new distribution

$$(p'_1,\ p'_2\ \ldots,\ p'_n) = \left(\textstyle\sum_i p_i p_{i,1},\ \sum_i p_i p_{i,2},\ \ldots,\ \sum_i p_i p_{i,n}\right)\ .$$

The total amount of sand is indeed preserved because

$$\sum_{j=1}^{n}\sum_{i=1}^{n} p_i p_{i,j} = \sum_{i=1}^{n} p_i \sum_{j=1}^{n} p_{i,j} = \sum_{i=1}^{n} p_i = 1\ .$$

If this transformation has not changed the distribution of sand, $\overline{p'} = \overline{p}$, the distribution $\overline{p}$ is stationary.

The necessary conditions on P for existence of a unique stationary distribution that is attractive, mentioned above, are (i) *irreducibility of D_P* and (ii) *aperiodicity of D_P*. The former means that for every two vertices $i, j \in [n]$ there exists in D_P an i-j walk, and the latter requires that the lengths of cycles in D_P be altogether coprime: the only $d \in \mathbb{N}$ dividing all of them is $d = 1$. One also says that the matrix P, or the Markov chain M, is irreducible, respectively aperiodic.

It is easy to see that (i) and (ii) are in general necessary. Consider the non-irreducible digraph $D_P = ([3], E, h)$ with arrows $1 \to 2$ and $1 \to 3$, loops $2 \to 2$ and $3 \to 3$, and weight $\frac{1}{2}$ on the arrows and 1 on the loops (the corresponding matrix P is stochastic). A stationary distribution exists but is not unique, both distributions $(0, 1, 0)$ and $(0, 0, 1)$ are stationary. Consider the non-aperiodic digraph $D_P = ([2], E, h)$ with only two arrows $1 \to 2$ and $2 \to 1$, weighted by 1 (the corresponding matrix P is stochastic). There is the unique stationary distribution $(\frac{1}{2}, \frac{1}{2})$ but it is not attractive, any distribution $(p, 1-p)$, $p \in [0, 1]$, only evolves to $(1-p, p)$ and back.

For the proof of the theorem on stationary distribution we need four lemmas. In the proof I follow the lecture notes of J. Sgall.

• *Jiří Sgall* is, by the Czech mutation of Wikipedia, Czech computer scientist and mathematician working in the areas of approximation algorithms, online algorithms and theory of scheduling. Currently he is the vice-dean of the School of Computer Science at the MFF UK (and hence lecturer's superior).

Lemma 1. *If the numbers $a_1, \ldots, a_k \in \mathbb{N}$, $k \in \mathbb{N}$, are altogether coprime then there is an $n_0 \in \mathbb{N}_0$ such that for every integer $n \geq n_0$ there exist numbers $m_1, \ldots, m_k \in \mathbb{N}_0$ with*

$$n = \sum_{i=1}^{k} m_i a_i \ .$$

Proof. It is a well known result in elementary number theory that then $1 = \sum_{i=1}^{k} r_i a_i$ for some numbers $r_i \in \mathbb{Z}$. We set $r := \max_i |r_i| \in \mathbb{N}$ and $n_0 := \sum_{i=1}^{k} (ra_1)a_i$. Adding to each of the numbers $n_0, n_0 + a_1, n_0 + 2a_1, \ldots$ repeatedly the linear combination expressing 1, we see that

every natural number from n_0 on is an integral linear combination of the a_is with nonnegative (here even positive) coefficients. □

Exercise (on compactness). Show that every set $A \subset \mathbb{N}$ of altogether coprime numbers (the only $d \in \mathbb{N}$ dividing all of the numbers is 1) has a finite subset of altogether coprime numbers. □

Lemma 2. *For every irreducible and aperiodic stochastic matrix $P = (p_{i,j})$ in $\mathbb{R}^{n\times n}_{\geq 0}$, $n \in \mathbb{N}$, there exists a number $k \in \mathbb{N}$ such that every entry in the matrix P^k is positive, and hence every entry in P^k is at least some $\delta > 0$.*

Proof. Recall that by a proposition in the last lecture, for every $i, j \in [n]$ the i,j-entry in P^k is the sum of weights of all i-j walks in D_P with length k, where the weight of a walk in D_P is the product of weights of its edges. Thus it suffices to show that there is a $k \in \mathbb{N}$ such that for every $i, j \in [n]$ there is an i-j walk in D_P with length k. By aperiodicity of P and the above exercise on compactness there exist cycles $C_1, \ldots, C_r$, $r \in \mathbb{N}$, in D_P with altogether coprime lengths. For each cycle C_l we fix a vertex $v_l \in C_l$. It is clear by irreducibility of D_P that for any given $i, j \in [n]$ there is an i-j walk w in D_P such that $|w| \leq (r+1)(n-1)$ and w goes through all vertices $v_1, \ldots, v_r$: we take the shortest path from i to v_1, then from v_1 to v_2, ..., and finally from v_r to j. Let $n_0 \in \mathbb{N}$ be the number guaranteed by Lemma 1 for $a_l := |C_l|$, $l = 1, 2, \ldots, r$. We take the walk w and extend it by an appropriate detour over the cycles C_l to an i-j walk w' in D_P with length

$$|w'| = k := (r+1)(n-1) + n_0 \, .$$

□

For $\overline{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $x \in \mathbb{R}$ we introduce the notation $|\overline{x}|_1 := \sum_i |x_i|$ (L_1 norm), $x^+ := \max(0, x)$ and $x^- := \max(0, -x)$. Then $x = x^+ - x^-$. Further,

$$\overline{x}^+ := (x_1^+, x_2^+, \ldots, x_n^+)$$

and similarly for $\overline{x}^-$. The vectors $\overline{x}^+$ and $\overline{x}^-$ have disjoint supports (sets of indices of the nonzero entries), $\overline{x} = \overline{x}^+ - \overline{x}^-$ and $|\overline{x}|_1 = |\overline{x}^+|_1 + |\overline{x}^-|_1$.

Lemma 3. *If $\overline{v}, \overline{w} \in \mathbb{R}^n_{\geq 0}$, $n \in \mathbb{N}$, are vectors with each entry at least $\alpha \geq 0$, then*

$$|\overline{v} - \overline{w}|_1 \leq |\overline{v}|_1 + |\overline{w}|_1 - 2n\alpha \ .$$

Proof. It suffices to prove this inequality only for $n = 1$: if $0 \leq \alpha \leq a \leq b$ then $b - a \leq a + b - 2\alpha$, which clearly holds as it is equivalent with $2\alpha \leq 2a$. □

Lemma 4. *Let $\delta > 0$ be a real number, M be a matrix in $[\delta, +\infty)^{n \times n}$, $n \in \mathbb{N}$, and with every row sum equal to 1, and let $\overline{x} \in \mathbb{R}^n$. Then the following hold.*

1. *$|\overline{x}M|_1 \leq |\overline{x}|_1 - 2\delta n \min(|\overline{x}^+|_1, |\overline{x}^-|_1)$.*
2. *If $\overline{x}$ has both positive and negative entries then $\overline{x}M \neq \overline{x}$.*
3. *If $\sum_i x_i = 0$ then $|\overline{x}M|_1 \leq (1 - \delta n)|\overline{x}|_1$.*

Proof. 1. We define $\overline{v} := \overline{x}^+ M$ and $\overline{w} := \overline{x}^- M$. Since the vectors $\overline{x}^+$, $\overline{x}^-$, $\overline{v}$ and $\overline{w}$ have nonnegative entries and the row sums of M are 1, we have that $|\overline{v}|_1 = |\overline{x}^+|_1$ and $|\overline{w}|_1 = |\overline{x}^-|_1$. Hence $|\overline{v}|_1 + |\overline{w}|_1 = |\overline{x}|_1$. The entries in M are at least δ and therefore we have for every $i \in [n]$ that $v_i \geq \delta \sum_i x_i^+ = \delta |\overline{x}^+|_1$ and, similarly, $w_i \geq \delta |\overline{x}^-|_1$. Applying Lemma 3 with $\alpha := \delta \min(|\overline{x}^+|_1, |\overline{x}^-|_1)$, we get that

$$\begin{aligned} |\overline{x}M|_1 &= |\overline{v} - \overline{w}|_1 \\ &\leq |\overline{v}|_1 + |\overline{w}|_1 - 2\delta n \min(|\overline{x}^+|_1, |\overline{x}^-|_1) \\ &= |\overline{x}|_1 - 2\delta n \min(|\overline{x}^+|_1, |\overline{x}^-|_1) \ . \end{aligned}$$

2. By the assumption, $\min(|\overline{x}^+|_1, |\overline{x}^-|_1) > 0$. Thus by part 1, $|\overline{x}M|_1 < |\overline{x}|_1$ and $\overline{x}M \neq \overline{x}$.

3. By the assumption, $|\overline{x}^+|_1 = |\overline{x}^-|_1 = |\overline{x}|_1/2$. Thus by part 1,

$$|\overline{x}M|_1 \leq |\overline{x}|_1 - 2\delta n \min(|\overline{x}^+|_1, |\overline{x}^-|_1) = (1 - \delta n)|\overline{x}|_1 \ .$$

□

Now we can prove the main theorem.

Theorem (on stationary distribution). *Any irreducible and aperiodic Markov chain M with the states $[n]$ has a unique stationary distribution $\overline{p}$ which is moreover attractive. If P is the transition matrix then for every $i, j \in [n]$ one has that*

$$p_j = \lim_{n\to\infty} (P^n)_{i,j} \ .$$

Proof. Let $k \in \mathbb{N}$ and $\delta > 0$ be as in Lemma 2. We claim that the system

$$\overline{x}P = \overline{x}$$

of n homogeneous linear equations with n unknowns $x_1, \ldots, x_n$ has a nontrivial solution, different from $\overline{0} = (0, \ldots, 0)$. Indeed, if we move the right sides to the left to get the canonical form $\overline{x}(P-I) = \overline{0}$, where I is the identity $n \times n$ matrix, the obtained matrix $P - I$ is singular because the sum of its columns is the zero column (P is stochastic). We denote this nontrivial solution by $\overline{x}$. From $\overline{x}P = \overline{x}$ we get that also $\overline{x}P^k = \overline{x}$ and see by part 2 of Lemma 4 and the exercise on matrices that $\overline{x} \geq \overline{0}$ (the vector $\overline{x}$ has only nonnegative entries) or $\overline{x} \leq \overline{0}$. Thus we may define the distribution

$$\overline{p} := \left(\textstyle\sum_i x_i\right)^{-1} \cdot \overline{x} \ .$$

We show that $\overline{p}$ is the unique stationary distribution of M. Since $\overline{x}P = \overline{x}$, also $\overline{p}P = \overline{p}$. If $\overline{q} \neq \overline{p}$ were another distribution satisfying $\overline{q}P = \overline{q}$, we would have also

$$(\overline{p} - \overline{q})P^k = \overline{p} - \overline{q} \ .$$

But this contradicts part 2 of Lemma 4 because the vector $\overline{p} - \overline{q}$ has both positive and negative entries.

We show that $\overline{p}$ is attractive: for any distribution $\overline{q}$ we have coordrdinatewise that

$$\lim_{n\to\infty} \overline{q}P^n = \overline{p} \ .$$

Then the vector $\overline{q} = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 at the i-th place yields the formula in the statement of the theorem.

Let $\overline{a}$ be any distribution. For $s = 0, 1, 2, \ldots$ let $\overline{v}^{(s)} := \overline{a}P^{sk} - \overline{p}$. Stationarity of $\overline{p}$ implies that $\overline{v}^{(s)} = \overline{a}P^{sk} - \overline{p}P^{sk} = (\overline{a} - \overline{p})P^{sk}$ and

therefore $\overline{v}^{(s+1)} = \overline{v}^{(s)}P^k$. By the exercise on matrices the coordinates of each vector $\overline{v}^{(s)}$ sum to 0. Since each entry in P^k is at least δ, by part 3 of Lemma 4 one has that

$$|\overline{v}^{(s+1)}|_1 = |\overline{v}^{(s)}P^k|_1 \leq (1-\delta n)|\overline{v}^{(s)}|_1 \ .$$

So $|\overline{v}^{(s)}|_1 \leq (1-\delta n)^s|\overline{v}^{(0)}|_1$ and $|\overline{v}^{(s)}|_1 \to 0$ for $s \to \infty$. Here $1-\delta n$ is in $[0,1)$ because every entry in P^k is at least $\delta > 0$ and every row sum of the $n \times n$ matrix P^k is 1. Therefore

$$\lim_{s\to\infty} \overline{a}P^{sk} = \overline{p} \ .$$

We use this result with $\overline{a} := \overline{q}P^l$ for any $l = 0, 1, \ldots, k-1$ and get that

$$\lim_{s\to\infty} \overline{a}P^{sk} = \lim_{s\to\infty} \overline{q}P^{sk+l} = \overline{p} \ .$$

Thus $\lim_{n\to\infty} \overline{q}P^n = \overline{p}$. □

Example (more on the Ehrenfest model). I will write here on the Diaconis–Shakshahani result on the mixing time in the Laplace–Bernoulli diffusion. □

Balls and bins. Bucket sort

An intuitive probabilistic model widely used in computer science is that of m balls and n bins, $m, n \in \mathbb{N}$, where the balls are placed randomly and mutually independently in the bins. One describes by it various situations in design of randomized algorithms. One can investigate various parameters of the (random) distribution of balls in bins.

For example, if $m \leq n$, or better if $m = o(n)$, then one can estimate for $n \to \infty$ the probability $B_m(n)$ that no bin contains two or more balls:

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

The first equality here follows by placing the first ball in any of the n bins, the second ball in any of the remaining $n-1$ empty bins, the third ball in any of the remaining $n-2$ empty bins, and so on.

Exercise. Recall that for two functions $f, g\colon \mathbb{N} \to \mathbb{C}$ the notation $f \sim g$ means that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. Show that for $m = o(n)$ the last $\approx$ is indeed $\sim$. Show that for $m = o(n^{2/3})$ the first $\approx$ is $\sim$. $\square$

We can interpret $B_m(n)$ as the probability that if there are n possibilities for birthdays (usually $n = 365$) then in a group of m people on a party no two of them have birthdays on the same day. (Assuming that the people and their birth dates are uncorrelated; for example, the party is not the first-day-of-the-year-baby convention.) For example, for $m \sim \sqrt{2(\log 2)n}$ we get (by the exercise) that $B_m(n) \sim \frac{1}{2}$.

As a sample result (taken from Mitzenmacher and Upfal) in the balls and bins model we prove a probabilistic bound on the maximum load of a bin.

Proposition (maximum load). *Let $m = n$ and $X_{\max}$ be the random variable recording the maximum number of balls in a bin ("maximum load"). Then*

$$\Pr\big(X_{\max} > 3\tfrac{\log n}{\log\log n}\big) \le \tfrac{1}{n},\ n \ge n_0\ .$$

Proof. For any $M \in \mathbb{N}$ we get by the union bound that

$$\Pr(\ge M \text{ balls in bin 1}) \le \binom{n}{M}\left(\frac{1}{n}\right)^M \le \frac{1}{M!} \le \Big(\frac{e}{M}\Big)^M\ .$$

Thus, again by the union bound, for $M \ge \frac{3\log n}{\log\log n}$ we have that

$$\begin{aligned}
&\Pr(\ge M \text{ balls in any bin}) \le n\left(\frac{e\log\log n}{3\log n}\right)^{3\log n/\log\log n}\\
&\le n\left(\frac{\log_{(2)} n}{\log n}\right)^{3\log n/\log_{(2)} n}\\
&= \exp(\log n)\exp\big((\log_{(3)} n - \log_{(2)} n)\tfrac{3\log n}{\log_{(2)} n}\big)\\
&= \exp\big(-2\log n + \tfrac{3\log n\log_{(3)} n}{\log_{(2)} n}\big)\\
&\le \frac{1}{n},\ n \ge n_0\ .
\end{aligned}$$

$\square$

Bucket sort. We sort $n = 2^m$ elements which are random integers from $[0, 2^k)$ where $k \geq m$. Using bucket sort we sort them in expected time $O(n)$. The algorithm is deterministic but the input is random.

First stage. We place the elements in n buckets $b_1, \ldots, b_n$ where we put in b_j the elements whose first m binary digits give number j. Thus $b_1 < b_2 < \cdots < b_n$ meaning that for $k < l$ any element in b_k is smaller than any element in b_l. We assume that each element can be placed in its bucket in $O(1)$ time. This stage therefore takes $O(n)$ time. Note that the random variable X recording the number of elements in a fixed bucket has the binomial distribution $B(n, 1/n)$ (see Lecture 9 for its definition).

Second stage. The elements in every bucket are sorted by some algorithm, for example bubblesort, in quadratic time. We estimate the expected running time of the second stage. If X_j is the random variable recording the number of elements in b_j, they are sorted in time $\leq cX_j^2$ for some constant $c > 0$. Thus the expected running time of the second stage is at most

$$\mathbb{E}\big(\textstyle\sum_{j=1}^n cX_j^2\big) = c\sum_{j=1}^n \mathbb{E}X_j^2 = cn\mathbb{E}X_1^2 = cn(2 - 1/n) = O(n) \; .$$

We used linearity of expectation and the formula for the second moment of a binomial distribution, computed in Lecture 9. In total, both stages together, bucket sort runs in time $O(n)$.

More on bucket sort. Time permitting, I write here what DEK writes on it in *TAOCP, Volume 3 (Sorting and Searching).*

Thank you!

(semifinal version — more information on the Ehrenfest model and bucket sort will be added — of January 13, 2021)