Stoch. Proc. HW assignment 3. Due Friday, September 29 at start of class

1. The USA and Canada both have 20 museums of modern art. Each of these 40 museums have one abstract expressionist painting, namely 20 Pollock paintings and 20 Rothko paintings. Each month, the international art committee chooses uniformly at random an abstract expressionist painting in Canada and swaps it with a uniformly chosen abstract expressionist painting in the US in order to entertain the art-loving audience in both countries with something new. Denote by X_n the number of Pollock paintings in the US in the n'th month.

   (a) Show that (X_n) is a birth/death chain and find its transition probabilities.

   (b) Find the stationary distribution of (X_n).

   Solution: This is a famous Markov chain, it is called the *Bernoulli-Laplace model* and it is a simple discrete model for the diffusion of two incompressible gases between two containers.

   (a) (X_n) is a birth-death chain with state space $S = \{0, 1, \dots, 20\}$ since $X_{n+1} - X_n \in \{-1, 0, 1\}$.
   If $x \in \{0, 1, \dots, 20\}$ is the number of Pollock paintings in the US in the current month then this number will increase by one next month if the committee picks a Rothko in the US (this happens with probability $\frac{20-x}{20}$ since the number of Rothkos in the US is $20 - x$) and a Pollock in Canada (this happens with probability $\frac{20-x}{20}$, since the number of Pollocks in Canada is $20 - x$). Now since the committee picked the two paintings to be swapped independently, we have

   $$p_x = \mathbb{P}[X_{n+1} = x + 1 \,|\, X_n = x] = \frac{20-x}{20}\frac{20-x}{20} = \frac{(20-x)^2}{400}.$$

   Similarly,

   $$q_x = \mathbb{P}[X_{n+1} = x - 1 \,|\, X_n = x] = \frac{x}{20}\frac{x}{20} = \frac{x^2}{400}.$$

   (b) We will keep track of the status of each museum: a museum can be either in a „Pollock" state or a „Rothko" state. Let us label the museums by the set $\{1, 2, \dots, 40\}$, where $\{1, 2, \dots, 20\}$ are the museums in the US and $\{21, 22, \dots, 40\}$ are the museums in Canada. Denote by $\widetilde{X}_n$ the set of museums with a Pollock in the n'th month. Thus $\widetilde{X}_n$ is a random subset of $\{1, 2, \dots, 40\}$, but the cardinality $|\widetilde{X}_n|$ of the set $\widetilde{X}_n$ is equal to 20. Thus the state space of the Markov process $(\widetilde{X}_n)$ is

   $$\widetilde{S} = \{\, A \subseteq \{1, \dots, 40\} \,:\, |A| = 20 \,\}.$$

   Note that $|\widetilde{S}| = \binom{40}{20}$. Note that $(\widetilde{X}_n)$ is an irreducible Markov chain, since the state $\{1, \dots, 20\} \in \widetilde{S}$ can be achieved by transporting all Pollocks to the US (and all of the Rothkos to Canada), and then any other state $A \in \widetilde{S}$ can be achieved from this state. The transition matrix of $(\widetilde{X}_n)$ is

   $$\mathbb{P}[\widetilde{X}_{n+1} = A' \,|\, \widetilde{X}_n = A] =$$
   $$\frac{1}{400}\mathbb{1}[\, |(A\Delta A') \cap \{1, \dots, 20\}| = 1,\ |(A\Delta A') \cap \{21, \dots, 40\}| = 1 \,], \qquad A, A' \in S, \qquad A \neq A',$$

   where $A\Delta A' = (A \setminus A') \cup (A' \setminus A)$ is the symmetric difference of the sets A and A'. In words, if $A \neq A'$ but A can be obtained from A' by swapping a painting from a US museum and a painting from a Canadian museum, then $\mathbb{P}[\widetilde{X}_{n+1} = A' \,|\, \widetilde{X}_n = A] = \frac{1}{20} \cdot \frac{1}{20}$, since the committee has to choose exactly these two paintings for this specific transition to occur.
   Now we can see that $\mathbb{P}[\widetilde{X}_{n+1} = A' \,|\, \widetilde{X}_n = A] = \mathbb{P}[\widetilde{X}_{n+1} = A \,|\, \widetilde{X}_n = A']$, i.e., the transition matrix of the „big" Markov chain $(\widetilde{X}_n)$ is a symmetric matrix. We have learnt (c.f. page 44 of scanned lecture notes) that this implies that the stationary distribution of $(\widetilde{X}_n)$ is uniform on the „big" state space $\widetilde{S}$. Therefore the stationary distribution of the original Markov chain (X_n) is

   $$\pi_x = \frac{|\{A \in \widetilde{S} \,:\, |A \cap \{1, \dots 20\}| = x\}|}{|\widetilde{S}|} = \frac{\binom{20}{x}\binom{20}{20-x}}{\binom{40}{20}} = \frac{\binom{20}{x}^2}{\binom{40}{20}}, \qquad x \in \{0, \dots, 20\}.$$

   Sanity check: $\pi_x p_x = \pi_{x+1} q_{x+1}$, since $\binom{20}{x}^2 \frac{(20-x)^2}{400} = \left(\frac{20!}{x!(19-x)!}\right)^2 \frac{1}{400} = \binom{20}{x+1}^2 \frac{(x+1)^2}{400}$. We have indeed constructed a measure π on S such that the *detailed balance* condition holds for π and the transition probabilities of the Markov chain (X_n).

2. Miles Raymond is a wine enthusiast, he drinks one bottle every day and puts the empty bottle on the shelf. His wife Maya, each evening, with probability 1/2 looks at the shelf and recycles all of the wine bottles. Every time five wine bottles pile up on the shelf, Miles himself immediately recycles all of them. Let B_n denote the number of wine bottles on the self on the n'th day (after the potential recycling).

   (a) Write down the transition matrix of the Markov chain (B_n).

   (b) Show that this Markov chain is irreducible and find its stationary distribution.

   (c) Suppose they have been doing this protocol for a long long time. What is the expected number of wine bottles on the shelf tonight (after the potential recycling)?

Solution:

(a)

$$\underline{\underline{P}} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(b) The Markov chain is irreducible because we can jump to state 0 from any state and then we can jump to 1 from 0, we can jump to 2 from 1, etc. All states can be reached from any other state, therefore the accessibility graph is strongly connected (see page 51 of the scanned lecture notes), hence the Markov chain (B_n) is irreducible. The unique stationary distribution π satisfies the equations

$$\begin{aligned} \pi_0 &= \frac{1}{2}\pi_0 + \frac{1}{2}\pi_1 + \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3 + \pi_4, \\ \pi_1 &= \frac{1}{2}\pi_0, \\ \pi_2 &= \frac{1}{2}\pi_1, \\ \pi_3 &= \frac{1}{2}\pi_2, \\ \pi_4 &= \frac{1}{2}\pi_3, \\ 1 &= \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4. \end{aligned}$$

The first five equations together state that $\pi\underline{\underline{P}} = \pi$, while the sixth equation guarantees that π is a probability distribution. Note that the first five equations are redundant, because if we add them up then we obtain the trivial identity

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 = \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4.$$

The reason for this redundancy is that the row sums of $\underline{\underline{P}}$ are all equal to 1. We can thus throw away one of the first five equations without changing the set of solutions. We choose to throw away the first equation. Thus

$$\pi_1 = \frac{1}{2}\pi_0, \quad \pi_2 = \frac{1}{4}\pi_0, \quad \pi_3 = \frac{1}{8}\pi_0, \quad \pi_4 = \frac{1}{16}\pi_0,$$

thus

$$\pi_0 = \frac{16}{31}, \quad \pi_1 = \frac{8}{31}, \quad \pi_2 = \frac{4}{31}, \quad \pi_3 = \frac{2}{31}, \quad \pi_4 = \frac{1}{31}.$$

(c) First note that all of the entries of the matrix $\underline{\underline{P}}^5$ are positive, since from every state we can get to every other state with five jumps: first jump to state 0, then possibly stay at state 0 for a few steps, then make increasing jumps to reach the desired state by step five. Thus by the theorem stated on page 53 of the lecture notes, the distribution of B_n converges to the stationary distribution as $n \to \infty$. This means that if Raymond and Maya have been doing this protocol for a long time, then we can assume that tonight the distribution of wine bottles on the shelf (after potential recycling) is π. Thus the desired expectation m is equal to

$$m = \sum_{x=0}^{4} \pi_x \cdot x = \frac{26}{31}. \tag{1}$$

3. There are N locations where a flower can grow. There are two types of flowers (dandelions and chamomiles, say) and each location is occupied by exactly one flower. Next year a new flower grows on each location and the seed of that flower came from a uniformly chosen flower from last year's population (this choice is independent for all locations). Denote by X_n the total number of dandelions in the n'th year.

   (a) Write down the transition matrix of the Markov chain (X_n).
   Hint: It is a good idea to first recall the notion of *binomial distribution*.

   (b) If $N = 5$, assuming we start with $X_0 = k$ dandelions, calculate the probability that we end up with all-dandelion population on the long run.
   Hint: Use symmetry to reduce the number of unknowns.

   (c) Now $N = 1000$. Assuming we start with $X_0 = k$ dandelions, calculate the probability that we end up with all-dandelion population on the long run.
   Hint: In order to solve (c), first solve (b) and make a guess about the form of the solution for the case of $N = 1000$, and then verify that your guess was indeed correct.

Solution: This is a famous Markov chain, it is called the *Wright-Fisher model*. It is a very simplistic yet popular stochastic model for reproduction in population genetics.

(a) If this year we have k dandelions, where $k \in \{0, 1, \dots, N\}$, then next year, on a fixed location we have a dandelion with probability k/N. Thus k/N is the dandelion-success probability of each of the N possible locations, the trials for the N locations are independent, and the number of dandelions next year is equal to the number of successful trials. Thus the number of dandelions next year has Binomial distribution $\text{BIN}\left(N, \frac{k}{N}\right)$.
This means that the transition matrix $\underline{\underline{P}} = (p(k,\ell))_{k,\ell=0}^{N}$ of the Markov chain (X_n) is

$$p(k,\ell) = \binom{N}{\ell}\left(\frac{k}{N}\right)^{\ell}\left(1-\frac{k}{N}\right)^{N-\ell}, \qquad k,\ell \in \{0,1,\dots,N\}. \tag{2}$$

Note that if we have zero dandelions this year then we have zero dandelions next year and if we have N dandelions this year then we have N dandelions next year. In other words, 0 and N are absorbing states of the Markov chain (X_n).

(b) Denote by $h(k)$ the probability that we end up with all-dandelion population on the long run if we start with k dandelions, i.e., that the Markov chain reaches state N before reaching state 0, starting from state k. The function h will satisfy the equations

$$h(k) = \sum_{\ell=0}^{N} p(k,\ell)h(\ell), \qquad 0 < k < N, \tag{3}$$

with boundary conditions $h(0) = 0,\ h(N) = 1$. There are $N-1$ unknowns: $h(1), h(2), \dots, h(N-1)$. Note that by symmetry we must have $h(N-k) = 1 - h(k)$ because either the dandelions or the chamomiles will win this battle and the chance of a dandelion victory starting with k dandelions is the same as the chance of a chamomile victory starting from k chamomiles (i.e., $N-k$ dandelions). Thus if $N = 5$ then there are only two unknowns ($h(1)$ and $h(2)$) and the equations read:

$$h(1) = p(1,0)\cdot 0 + p(1,1)\cdot h(1) + p(1,2)\cdot h(2) + p(1,3)\cdot(1-h(2)) + p(1,4)\cdot(1-h(1)) + p(1,5)\cdot 1,$$
$$h(2) = p(2,0)\cdot 0 + p(2,1)\cdot h(1) + p(2,2)\cdot h(2) + p(2,3)\cdot(1-h(2)) + p(2,4)\cdot(1-h(1)) + p(2,5)\cdot 1,$$

that is, we have

$$h(1) = 5\frac{1}{5}\left(\frac{4}{5}\right)^{4}h(1) + 10\left(\frac{1}{5}\right)^{2}\left(\frac{4}{5}\right)^{3}h(2) + 10\left(\frac{1}{5}\right)^{3}\left(\frac{4}{5}\right)^{2}(1-h(2)) + 5\left(\frac{1}{5}\right)^{4}\frac{4}{5}(1-h(1)) + \left(\frac{1}{5}\right)^{5}\cdot 1,$$
$$h(2) = 5\frac{2}{5}\left(\frac{3}{5}\right)^{4}h(1) + 10\left(\frac{2}{5}\right)^{2}\left(\frac{3}{5}\right)^{3}h(2) + 10\left(\frac{2}{5}\right)^{3}\left(\frac{3}{5}\right)^{2}(1-h(2)) + 5\left(\frac{2}{5}\right)^{4}\frac{3}{5}(1-h(1)) + \left(\frac{2}{5}\right)^{5}\cdot 1.$$

That is, we have

$$h(1) = \frac{256}{625}h(1) + \frac{128}{625}h(2) + \frac{32}{625}(1-h(2)) + \frac{4}{625}(1-h(1)) + \frac{1}{3125},$$
$$h(2) = \frac{162}{625}h(1) + \frac{216}{625}h(2) + \frac{144}{625}(1-h(2)) + \frac{48}{625}(1-h(1)) + \frac{32}{3125}.$$

Solving this system of equations with Wolfram Alpha (free online math tool) we obtain $h(1) = \frac{1}{5}$ and $h(2) = \frac{2}{5}$. So by $h(5-k) = 1 - h(k)$ we have

$$h(k) = \frac{k}{5}, \qquad 0 \le k \le 5.$$

(c) We conjecture that for general N the function $\widetilde{h}(k) = \frac{k}{N}$ will satisfy the equations

$$\widetilde{h}(k) = \sum_{\ell=0}^{N} p(k,\ell)\widetilde{h}(\ell), \qquad 0 < k < N, \tag{4}$$

with boundary conditions $\widetilde{h}(0) = 0$ and $\widetilde{h}(N) = 1$. Let's prove this:

$$\sum_{\ell=0}^{N} p(k,\ell)\widetilde{h}(\ell) = \sum_{\ell=0}^{N} p(k,\ell)\frac{\ell}{N} = \frac{1}{N} \cdot \sum_{\ell=0}^{N} p(k,\ell)\ell \overset{(*)}{=} \frac{1}{N} \cdot N\frac{k}{N} = \frac{k}{N} = \widetilde{h}(k),$$

where the equation marked by $(*)$ holds by (2) and the fact that the expected value of a random variable with $\mathrm{BIN}\,(N,p)$ distribution is equal to Np.

Thus if $h(k)$ is the probability of the event that the Markov chain reaches state N before reaching state 0 starting from state k and $\widetilde{h}(k) = k/N$, then by (3) and (4) we see that h and $\widetilde{h}$ solve the same system of equations with the same boundary condition.

Does that mean that $\widetilde{h} \equiv h$?

In other words, is the solution of the system of equations (3) with boundary condition $h(0) = 0$ and $h(N) = 1$ unique? Let's prove that this is indeed the case.

Let us consider two solutions h and $\widetilde{h}$ of (3) that satisfy the same boundary conditions. Let us define

$$v(k) = h(k) - \widetilde{h}(k) \qquad \text{for any} \qquad k \in \{0, 1, \dots, N\}.$$

Then we have

$$v(k) = \sum_{\ell=0}^{N} p(k,\ell)v(\ell), \qquad 0 \le k \le N, \tag{5}$$

moreover v satisfies the boundary conditions

$$v(0) = 0, \qquad v(N) = 0. \tag{6}$$

Let us define

$$M = \max_{1 \le k \le N-1} |v(k)|$$

and let us consider some particular $k^* \in \{1, \dots, N-1\}$ for which $M = |v(k^*)|$. We have

$$M = |v(k^*)| \overset{(5)}{=} \left| \sum_{\ell=0}^{N} p(k^*,\ell)v(\ell) \right| \overset{(6)}{=} \left| \sum_{\ell=1}^{N-1} p(k^*,\ell)v(\ell) \right| \le \sum_{\ell=1}^{N-1} p(k^*,\ell)M = (1 - p(k^*,0) - p(k^*,N))M.$$

Now $(1 - p(k^*,0) - p(k^*,N)) \in (0,1)$ thus we must have $M = 0$. Therefore $v \equiv 0$ and $h \equiv \widetilde{h}$.

Therefore if $N = 1000$, we have $h(k) = k/1000$ for any $k \in \{0, \dots, 1000\}$.