Randomness in Biology [ICTP Spring College 2019]

HW problems numbered according to Lectures 1 – 7.

1A. Random number generators. In class we discussed how to make random number generators using fair coin flips (i.e. using a Bernoulli random variable X whose outcome is X = 1 with probability p and X = 0 with probability (1-p)).

a. On a computer, using Bernoulli bits, simulate a 10-bit random number generator that produces numbers uniformly distributed between 0 and 1. Plot the histogram of the output of this process and verify that it approaches a uniform distribution.

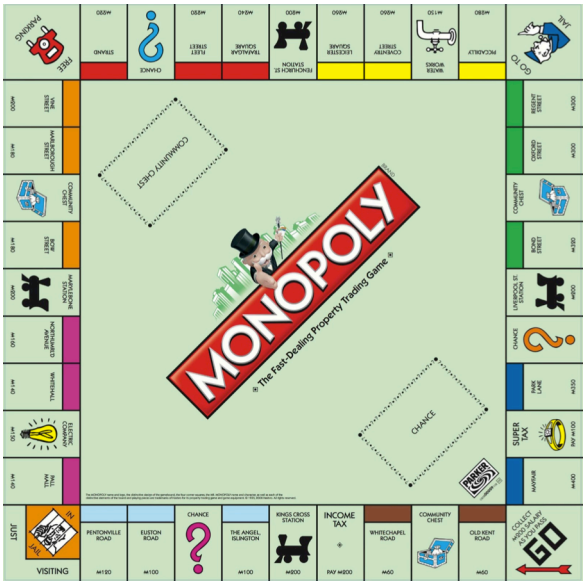
b. We discussed how the random walk approaches a Gaussian distribution. Based on this idea, how would you use Bernoulli bits to simulate a random number generator whose output is a Gaussian distribution with mean 0 and variance 1? Optional: Simulate this approach on a computer to see how well it works.

1B. Central limit theorem. I have a fair die, which generates each of the numbers 1 to 6 with equal probability at each roll. Let the outcome of the die roll X be represented by the random variable $x \in \{1,2,3,4,5,6\}$.

a. Compute the following: $< X >, < X^2 >, \sigma_X^2$

b. If I define $Y = X_1 + X_2 + \cdots + X_n$ for large n, approximately what distribution will Y obey? Give a precise mathematical formula for the approximate distribution.

2. Markov processes. The game of monopoly is played by moving tokens on a board with 40 locations.



At each step, a pair of dice is rolled and the player moves ahead by the corresponding number of steps. The player starts at "GO" (bottom right) and move clockwise. Two other important locations are "Jail" (bottom left) and "Go to jail" (top right). If you land on "Go to jail" then the token is immediately moved to "Jail". I.e. the probability of being found in the "Go to jail" location is always zero, so in reality there are only 39 accessible locations.

a. Write down the 39x39 Markov transition matrix *A* for this game. Number as follows: "GO"=1, "Jail"=11, "Free Parking"=21, "Mayfair"=39. Remember, the game wraps around, and throws that would otherwise have landed in "Go to jail" must be made to land in "Jail" itself.

b. Find the distribution after 50 steps if the token starts at "GO".

c. Find the eigenvector of *A* with the largest eigenvalue. This should be similar to what you found in (b).

d. Bonus. Apart from "Jail", some locations are more likely than others after a large number of steps. What do you think are the most overvalued properties and the best value-for-money properties? (Prices are listed as numbers on each location.)

3. Stochastic differential equations. The velocity of a Brownian particle is an example of an Ornstein-Uhlenbeck process described by the following Fokker-Planck equation:

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial v}\left(\frac{\Gamma v}{m}p\right) + \frac{\partial^2}{\partial v^2}\left(\frac{\Gamma kT}{m^2}p\right)$$

This is equivalent to the Langevin equation

$$\frac{dx}{dt} = v$$
$$m\frac{dv}{dt} = -\Gamma v + \xi(t) \qquad < \xi(t)\xi(t+t') \geq 2\Gamma kT\delta(t')$$

Or, more explicitly,

$$\Delta x = \frac{dx}{dt}\Delta t, \qquad \Delta v = (-\Gamma v/m)\Delta t + \alpha\sqrt{2\Gamma kT/m^2}\sqrt{\Delta t}$$

where α is normally distributed with mean 0 and variance 1. Let measure time in units where the relaxation time is unity ($m/\Gamma = 1$), and measure distance in units such that the RMS velocity is unity ($kT/m = 1$).

a. Focus just on the velocity equation. Find a numerical solution to the Langevin equation as follows.

1.Start with $v(t = 0) = 0$ and choose a timestep Δt=0.1.
2. Compute Δv as above.
3. Update: $v = v + \Delta v$, and $t = t + \Delta t$.
4. Continue up to a maximum time t = 100.
5. Repeat for e.g. 100 realizations, storing just the final velocity in each case.

6. Run again for all combinations of v $t = 0 = 0$, 10; and $\Delta t = 0.1$, 0.01, 0.001. For your solution, submit five sample trajectories each for all 6 cases above.

b. Do velocities converge to a reproducible distribution independent of Δt and $v(t = 0)$?
c. What is the RMS value of the final velocity for the six cases?
d. [Optional] For $v(t = 0) = 0$ and $\Delta t = 0.001$, numerically solve the full equation for x, v. What is the apparent value of the diffusion coefficient? (Plot x vs t for many realizations, and check how the value of $< \delta x^2 >$ changes with t. The slope of this curve will be 2D.)

4. Stochastic chemical kinetics: flipping a genetic switch. We have seen in class that the following differential equation describes a protein which activates its own transcription. This is equivalent to a double-well potential, where the two wells correspond to states of low (x_{low}^{SS}) and high (x_{high}^{SS}) gene expression.

$$\frac{dx}{dt} = \frac{v_0 + v_1 K_1 K_2 x^2}{1 + K_1 K_2 x^2} - \gamma x$$

Use the following parameters: $v_0 = 12.5, v_1 = 200, \gamma = 1, K_1 K_2 = 10^{-4}$ or $K_1 K_2 = 10^{-6}$.

For each value of $K_1 K_2$, estimate the steady-state distribution of gene expression levels using the following approaches:

a. The deterministic stable steady states x_{low}^{SS} and x_{high}^{SS}.

b. The approximate form $\mathcal{P}(x) = \frac{A}{f(x)+g(x)} \exp\left(2 \int \frac{f(x)-g(x)}{f(x)+g(x)} dx\right)$. Plot this as a graph, and mark the deterministic steady states.

c. A Langevin stochastic differential equation by adding an appropriate noise term, for 1000 replicates starting at $x = x_{low}^{SS}$ and running until you think steady state is reached. Overlay this with the graph from (b).

d. A Gillespie simulation, for 1000 replicates starting at $x = x_{low}^{SS}$ and running until you think steady state is reached. Overlay this with the graph from (b).

e. [Optional] Use the Langevin method to find the mean first passage time to transit between the two deterministic steady states for the case $K_1 K_2 = 10^{-4}$. I.e. start with states x_{low}^{SS} or x_{high}^{SS} and keep running until you hit the other state. Run this for 100 replicates and see how much time each of these trajectories take.

5A. Instantaneous code. Consider the following set of codewords:

(A,B,C,D,E,F,G,H) = (01, 11, 001, 0000, 0001, 1001, 1010, 1011).

a. Is this an instantaneous (prefix) code?
b. Verify that it satisfies the Kraft inequality
c. Construct a string which has no meaning under this system

5B. Entropy and mutual information. Given a joint distribution, calculate various quantities:

$y\backslash x$	0	1
0	1/4	1/4
1	0	1/2

a. H(X,Y). b. H(X). c. H(Y). d. H(X|Y). e. H(Y|X).
f. H(X) + H(Y) – H(X,Y) g. H(X) – H(X|Y) h. H(Y) – H(Y|X) i. I(X;Y)

6. Typical sequences. We had defined a stringently typical sequence as one containing exactly as many occurrences of each symbol as expected. Let's find out (a) what the probability of each such sequence is and (b) how many such sequences there are *exactly* (i.e. not using Stirling's approximation). Then let's see how much of the total probability space is occupied by these typical sequences.

Consider a DNA sequence of length 8 generated iid from the distribution

$$\wp(A,T,G,C) = (\tfrac{1}{2},\tfrac{1}{4},\tfrac{1}{8},\tfrac{1}{8}).$$

a. What is the single most probable sequence? What is its probability of occurrence?
b. What is the probability of a given 'stringently typical' sequence, defined as one in which letters occur precisely as often as expected?
c. How many stringently typical sequences are there (exact answer required)?
d. What is the total probability of getting some stringently typical sequence?
e. Redo the whole calculation if the length is 16. What is the total probability of getting a stringently typical sequence? Are we converging to 1?

You should find that, as the sequences get longer, fewer and fewer of them are 'typical' by this definition. This motivates the new definition of typical sequence we will make this week.

7. Channel capacity. The Z channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \qquad x, y \in \{0, 1\}.$$

Find the capacity of the Z channel and the maximizing input probability distribution.