Randomness in Biology [ICTP Spring College 2019] – Solutions to HW

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.

First choose some value of *N*, say *N* = 10, giving us $2^{10} = 1024$ possible ways the flips can land (0s and 1s). Each of these values corresponds to a binary number *x* between 0 and 1023. Spit out the output of the function as *x*/1024, which will give values evenly spaced between 0 and 0.999. To implement this on a computer, some of you might know how to get a random Bernoulli bit directly. Others might have used a uniform random number generator whose output is *y*, and used this to generate a Bernoulli variable by setting *z* = 1 if *y* > 0.5, and *z* = 0 otherwise. Ironically, this is a way to convert one uniform RNG into another.

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.

The trick here is only to keep track of the variance. Again, choose some value of *N*, say *N* = 1000. Let each coin flip be described by the variable *z*. The total number of 1s in each of these trials, call it *x*, has a nearly Gaussian distribution, according to the Central Limit Theorem.

$< x >= N < z >= 1000\times0.5 = 500.$
$\sigma_x^2 = N\sigma_z^2 = N(< x^2 > -< x >^2) = 1000\times(0.5 - 0.25) = 250.$

So the variable you want to spit out is actually: $y = (x - 500)/\sqrt{250}.$
This has mean 0 and variance 1.

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: $< X >, < X^2 >, \sigma_X^2$

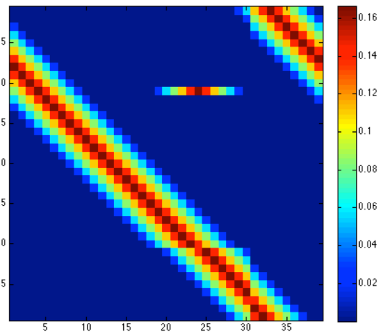
$< X >= \frac{1+2+3+4+5+6}{6} = \frac{21}{6}$ $< X^2 >= \frac{1^2+2^2+\cdots 6^2}{6} = \frac{91}{6}$ $\sigma_X^2 =< X^2 > -< X >^2$ ~2.9

b. If I define $Y = X_1 + X_2 + \cdots + X_n$ for large *n*, approximately what distribution will *Y* obey? Please be specific, give a mathematical formula for that distribution.

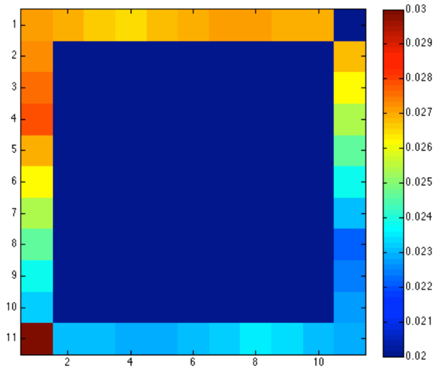
$p(Y = y) = \frac{1}{\surd(2\,\pi n\sigma_X^2)} \exp\left(-(y-< X >)^2/(2n\sigma_X^2)\right)$

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

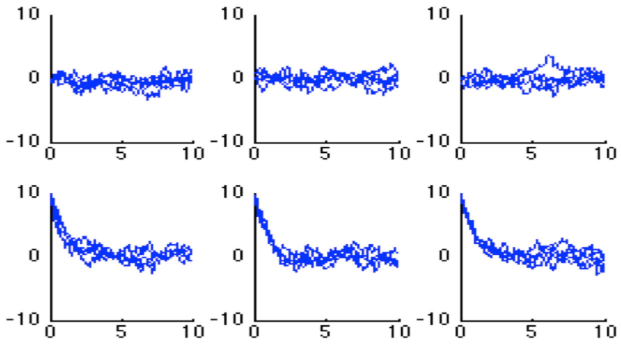


3. Stochastic differential equations. Velocity of a Brownian particle. Langevin description:

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

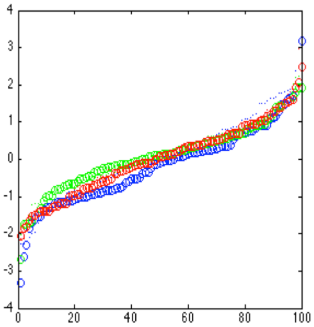
where α is normally distributed with mean 0 and variance 1. Let measure time in units where the relaxation time is unity ($m/\Gamma = 1$), and measure distance in units such that the RMS velocity is unity ($kT/m = 1$).

a. Sample trajectories:



b. Do velocities converge to a reproducible distribution independent of Δt and $v(t=0)$?

Yes. See cumulative distributions of final velocity below. Open circles: v0=10. Dots: v0=0. Blue: dt=0.1. Green: dt=0.01. Red: dt=0.001.



c. What is the RMS value of the final velocity for the six cases? Here is what I get from 100 samples each.

[v0,dt]	[0, {0.1,0.01,0.001}]:	1.13	0.93	1.03
	[10,{0.1,0.01,0.001}]:	1.06	0.88	0.98

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: $v_0 = 12.5, v_1 = 200, \gamma = 1, K_1K_2 = 10^{-4}$ or $K_1K_2 = 10^{-6}$.

For each value of K_1K_2, estimate the steady-state distribution of gene expression levels using the following approaches:

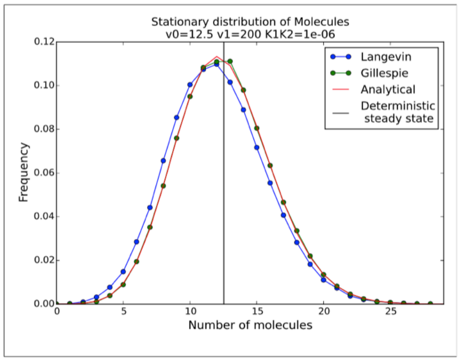
a. The deterministic stable steady states x_{low}^{SS} and x_{high}^{SS}.

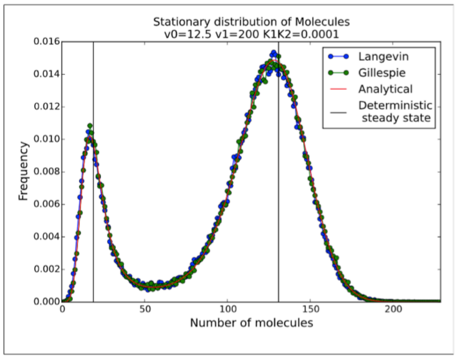
$K_1K_2 = 10^{-4}$: the stable states are: 19.10 and 130.90
$K_1K_2 = 10^{-6}$: the stable state is: 12.53

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 $K_1K_2 = 10^{-4}$. 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.

Mean First Passage Time (100000 trials): from 19.10 to 130.90: 182.72
from 130.90 to 19.10: 716.31

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?

Yes. All words are distinct, and no word is the prefix of any other.

b. Verify that it satisfies the Kraft inequality

$2\times 2^{-2} + 2^{-3} + 5\times 2^{-4} = \frac{15}{16} < 1.$

c. Construct a string which has no meaning under this system

By writing out the code on a tree, you can easily see that the symbol 1000 is a valid word which is not a prefix of any other, but not assigned to a meaning. So any string beginning with 1000… cannot be decoded. If we assign a dummy symbol Z to this string, then it is possible to show that *any* arbitrary string of 1s and 0s decodes to a unique set of letters.

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)

Soln (all answers in bits):

a. 1.5   b. 0.81   c. 1   d. 0.5   e. 0.69
f. 0.31   g. 0.31   h. 0.31   i. 0.31

6. 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?

The most likely sequence is one where each symbol is chosen from the set of most likely symbols: $AAAAAAAA$. Its probability of occurrence is $\left(\frac{1}{2}\right)^8 = 3.9\times 10^{-3}$.

b. What is the probability of a given 'stringently typical' sequence, defined as one in which letters occur precisely as often as expected?

A stringently typical sequence is any sequence containing 4 As, 2 Ts, 1 G, and 1 C. The entropy of the original distribution is

$$H = -\sum p\log(p) = \left(\frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{4}3\right) = 1.75 \text{ bits}$$

So the probability of a stringently typical sequence of length 8 is: $2^{-8H} = 6.1\times 10^{-5}$.

c. How many stringently typical sequences are there (exact answer required)?

The number of stringently typical sequences is $\frac{8!}{4!2!1!1!} = 840$.

d. What is the total probability of getting some stringently typical sequence?

The total probability of getting a stringently typical sequence is the product of the previous two answers: 0.05.

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?

For a sequence of length 16 we have $\frac{16!}{8!4!2!2!} = 5405400$ stringently typical sequences, each with probability $2^{-16H} = 3.72\times 10^{-9}$, so the total probability is 0.02. You can verify by using even longer sequences that by using this stringent definition of typicality, we do not cover most of the sequence space.

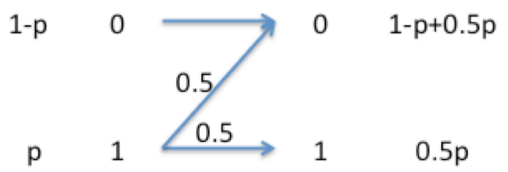
7. Channel capacity. The Z channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

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

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

Soln: This problem requires a bit of work. Of course it's called the Z channel because that's the shape of the cartoon. Let's just start with the definitions.

Define $p = \wp(X = 1)$. We need to maximize $I(X;Y)$ as a function of p. So:



$H(Y) = H(\frac{p}{2})$; $\qquad H(Y|X) = (1-p)\times 0 + p\times H(\frac{1}{2})$;
$I(X;Y) = f(p) = H(Y) - H(Y|X) = H(\frac{p}{2}) - p$.

Maximizing wrt p:

$$0 = \frac{df}{dp} = -\frac{1}{2}\log\left(\frac{p}{2}\right) - \frac{1}{2} + \frac{1}{2}\log\left(1 - \frac{p}{2}\right) + \frac{1}{2} - 1 = \frac{1}{2}\log\left(\frac{2-p}{p}\right) - 1$$
$$\Rightarrow \frac{(2-p)}{p} = 4 \Rightarrow p = \frac{2}{5}.$$

This gives $C = \max I(X;Y) = 0.322$.
Notice that you use the noisy symbol “1” less frequently than “0”.