BE/APh 161: Physical Biology of the Cell, Winter 2025
Homework #5
Due at the start of lecture, 2:30PM, February 12, 2025.

Problem 5.1 (Visualizing random walks, 20 pts)**.**
In this problem, we will develop intuition about random walks by exploring them computationally.

a) Write a computer program to generate two-dimensional random walks. Generate five random walks of 10^5 steps with unit step size. Plot these five random walks on the same plot (but space them apart so they do not overlap). Comment on anything you find striking about the plot.

b) Write a computer program to generate the start-to-end distance of a one-dimensional random walk. You will generate three sets of random walks, with each walk containing 10^4 steps. The first set has 10 walks, the second has 1000, and the third has 100,000. For each set of walks, plot either an ECDF (preferred) or a normalized histogram of the displacement (the distance from the origin of the last step of the walk). (*Hint*: To plot the ECDF in Python using Bokeh, you can use the iqplot package.) Overlay the corresponding solution to the continuum diffusion equation. (If you plot an ECDF, the solution to the diffusion equation is found by converting the probability density to a cumulative density function.) Comment on the plots. *Hint*: You do not actually need to "take" each walk. Think about how the end-to-end distance is distributed and you can draw random numbers out of that distribution.

Problem 5.2 (Transcriptional pausing and random walks, 40 pts)**.**
Transcription, the process by which DNA is transcribed into RNA, is key process in the central dogma of molecular biology. RNA polymerase (RNAP) is at the heart of this process. This amazing machine glides along the DNA template, unzipping it internally, incorporating ribonucleotides at the front, and spitting RNA out the back. Sometimes, though, the polymerase pauses and then backtracks, pushing the RNA transcript back out the front, as shown in the figure below, taken from Depken, et al., *Biophys. J.*, **96**, 2189-2193, 2009.

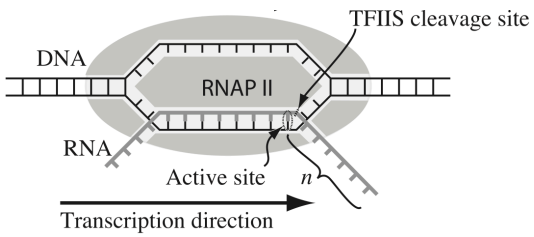

Figure 1: Schematic of an RNA polymerase in a backtrack. In this schematic, n is the distance of the backtrack. In our notation, $x = -n$.

To escape these backtracks, a cleavage enzyme called TFIIS cleaves the bit on RNA hanging out of the front, and the RNAP can then go about its merry way.

Researchers have long debated how these backtracks are governed. Single molecule experiments can provide some much needed insight. The groups of Carlos Bustamante, Steve Block, and Stephan Grill, among others, have investigated the dynamics of RNAP in the absence of TFIIS. They can measure many individual backtracks and get statistics about how long the backtracks last.

One hypothesis is that the backtracks simply consist of diffusive-like motion along the DNA stand. That is to say, the polymerase can move forward or backward along the strand with equal probability once it is paused. This constitutes a one-dimensional random walk. So, if we want to test this hypothesis, we would want to know how much time we should expect the RNAP to be in a backtrack so that we could compare to experiment.

So, we seek the probability distribution of backtrack times, $P(t_{\text{bt}})$, where t_{bt} is the time spent in the backtrack. This is an example of a **first-passage problem**, an important class of problems when working with random walks. The backtrack time is the amount of time it takes for the polymerase to first arrive at the point where it exits a backtrack. Specifically, if $+1$ is the position where it can begin elongating and zero is the position where it has backtracked one base pair (which, by definition is the start of the backtrack), then t_{bt} is the time for which the polymerase first arrives at position $x = 1$ starting from $x = 0$.

In this problem, we will first address the problem of the first passage time for a one-dimensional random walk with discrete steps analytically. Then, we will simulate the process numerically and compare the results.

a) Consider a one-dimensional random walk starting at zero where the walker may step right or left one unit. That is, the allowed positions x are integers. Let t be the number of steps in the walk, which we will refer to as "time," even though it is the integer number of steps the walker has taken. Note that Depken, et al., report that the time each step takes in typical single-molecules set-ups is about 0.5 seconds, but we will simply refer to t as the number of steps of the walk for convenience.

   i) Let $N(x,t)$ be the number of possible random walks that bring the walker to position x in t steps. Write an expression for $N(x,t)$ in terms of x and t.

   ii) Let $N_y(x,t)$ be the number of random walks that walk to or past point y and end at point x with $y > x$. The reflection principle states that $N_y(x,y) = N(2y-x,t)$. Provide an argument as to why this is true.

   iii) Let $N_F(x,t)$ be the number of first-passage paths from zero to x. A first passage path to x is one that visits x exactly once, at the very end of the

walk. Provide an argument as to why $N_F(x,t) = N_F(x-1,t-1)$.

iv) Provide an argument as to why $N_F(x,t) = N_F(x-1,t-1) = N(x-1,t-1) - N_x(x-1,t-1)$.

v) Show that $N_F(x,t) = xN(x,t)/t$.

vi) Provide an argument as to why, in the limit of large t,

$$N(x,t) \propto \frac{1}{\sqrt{t}}\, \mathrm{e}^{-x^2/2t}. \tag{5.1}$$

vii) Using all these results, deduce that, in the limit of large t_{bt},

$$P(t_{\mathrm{bt}} \propto t_{\mathrm{bt}}^{-3/2}\, \mathrm{e}^{-1/2t_{\mathrm{bt}}}. \tag{5.2}$$

b) Show that the mean backtrack time is infinite(!).

c) Let $F(t_{\mathrm{bt}})$ be the cumulative distribution function. Though t_{bt} as we have treated it here is discrete, we can in the limit of large t_{bt} treat it as continuous such that the complementary cumulative distribution function is

$$S(t_{\mathrm{bt}}) = 1 - F(t_{\mathrm{bt}}) \approx \int_{t_{\mathrm{bt}}}^{\infty} \mathrm{d}t'_{\mathrm{bt}}\, P(t'_{\mathrm{bt}}). \tag{5.3}$$

The complementary cumulative distribution function is sometimes called the **survival function**. In this case, it has the interpretation that it is the probability that the walker has *not* exited the backtrack prior to or at time t_{bt}. Sketch the survival function on a log-log plot making any necessary annotations.

d) Now write code to sample out of the distribution defined by $P(t_{\mathrm{bt}})$. This is equivalent to simulating a 1D random walk starting at $x = 0$ and then recording how many steps it takes to get to position $x = 1$. Take many samples and then make a plot of the empirical complementary cumulative distribution function. (*Hint*: To plot the ECDF in Python using Bokeh, you can use the iqplot package.) Does your simulation show the same scaling as the theoretical result?

e) Comment on what you have learned about the amount of time a polymerase remains in a backtrack from this analysis. More generally, are you surprised about how first-passage times for a random walker are distributed?

Problem 5.3 (Sedimentation, the Einstein-Smoluchowski equation, and the Stokes-Einstein-Sutherland relation, 40 pts)**.**
In this problem, we will derive some landmark results in statistical physics, and learn something about a technique for studying protein structure in the process.

In equilibrium sedimentation experiments, a tube of solution of a protein or protein complex of interest is placed in a centrifuge. The concentration of protein is

measured along the tube. The shape of this concentration profile is used to infer information about the size and shape of the protein. For our analysis, let ω be the angular velocity of the rotor of the centrifuge and r describe the distance from the center of the rotor to a given position in the tube of solution. Let $\rho_{\mathrm{H_2O}}$ be the density of the solvent and ρ_{p} be the density of the protein. Note that $\rho_{\mathrm{p}}/\rho_{\mathrm{H_2O}} \approx 1.4$ (BNID 104272). Let a be the radius of gyration of the protein.

a) Due to its density being greater than water, the protein will tend to fall toward the bottom of the tube with steady state velocity v. As it falls through the solvent, it experiences a friction f, such that it experiences a drag force of $F_{\mathrm{drag}} = -fv$. At steady state, the drag force balances the centrifugal force. Use this fact to compute the velocity with which it falls in terms of ρ_{p}, $\rho_{\mathrm{H_2O}}$, a, ω, r, and f. *Hint*: The centrifugal force is given by $F_{\mathrm{centrifugal}} = m_e \omega^2 r$, where m_e is the effective mass of the protein.

b) Show that at steady state, the sedimentation velocity is given by

$$v = D \frac{\mathrm{d} \ln c(r)}{\mathrm{d}r}, \tag{5.4}$$

where D is the diffusion coefficient of the protein.

c) Now use equilibrium statistical mechanics to derive an expression for the concentration profile of the protein. I.e., compute $c(r)$ as a function of $k_B T$ and the other variables describing the system. Note that if $P(r)$ is the probability density for a given particle being at position r in the centrifuge, $c(r) \propto P(r)$. Assume that in the absence of centrifugation, the solution has a uniform concentration of c_0. *Hint*: In part (a), you used an expression for the centrifugal force. Recall that a force $F(r)$ acting on a particle in a potential $U(r)$ is given by $F(r) = -\mathrm{d}U(r)/\mathrm{d}r$.

d) Use your expressions from parts (b) and (c) to derive an expression for D in terms of f. This is the Einstein-Smoluchowski equation, an example of a fluctuation-dissipation theorem. It has this name because it relates equilibrium fluctuations to response to applied perturbations. This is a profound and important concept in statistical physics.

e) The friction f is given by Stokes's law (applicable for spherical particles),

$$f = 6\pi\eta a. \tag{5.5}$$

This was derived by George Stokes by solving for fluid flow around a spherical object. Insert this result into your result in part (d) to get the Stokes-Einstein-Sutherland relation.

f) The sedimentation coefficient S is the ratio of the sedimentation velocity to the acceleration applied to it. It therefore has units of time. Derive an expression for S.

g) Ribosomes are often named by their sedimentation coefficient. A typical unit is a svedberg, which is equal to 10^{-13} s. The 70S ribosome has a sedimentation coefficient of approximately 70 svedbergs. Estimate the diameter of the 70S ribosome. Compare this estimate to what is reported on BioNumbers and explain any discrepancies.