Harvard John A. Paulson
School of Engineering
and Applied Sciences

Michael P. Brenner and Chris Gumb

Homework 2

Random Walks: From Basketball scores to Stock prices

Issued:	September 30, 2024
Due:	October 17, 2024

Problem 1: *Basketball Scoring as a Random Walk*

This problem explores modeling basketball scoring as a random walk process, building ideas we introduced in class.

1. Consider this game from the NBA season, and calculate the probability per unit time (30s in this case) of one of the two teams (your choice) scoring $n = 0, 1, 2, 3, ...$ points. For this, you might find this analysis of the box scores to be useful, giving the total number of 1,2,3 point shots by each team. Divide the game in windows of 30s, then count how many points were scored by this team in each 30s window and calculate the probabilities of each scoring you registered to happen. Note: Follow the code we gave in class for how to download and manipulate the code from the game you have chosen.

2. Using these probabilities, simulate a random walk representing the score progression of the team in a basketball game. Generate multiple simulations for a full game duration (48 minutes), the plot the score distribution after 12, 24, 36, and 48 minutes of play.

3. Derive the associated continuum equation for this process. Your answer should be in the form of a differential equation.

   (**Hint:** To do this, follow the derivation of the diffusion equation explained in class and outlined in the notes on the class website (Eqn. 19 on page 9, and subsequent calculations). Your starting point should have the form:

   $$P(s, t+1) = \sum_{n=0}^{N} p_n P(s-n, t)$$

   where $P(s,t)$ is the probability of having score s at time t, and p_n is the probability of scoring n points in one time step (with $n = 0, 1, 2, 3, 4, ...$ representing points scored by the team). N will be the maximum number of points scored in a single 30-second time step that you registered in (1).)

4. Using your derived equation, calculate and plot the expected score distribution after 12, 24, 36, and 48 minutes of play. Assume the team starts at 0 points, that is $s(t = 0) = \delta(s)$.

5. Compare the theoretical distributions to the ones you generated in (2). How does the variance of the score grow with time? Is this consistent with a standard random walk?

6. Discuss the limitations of this model. What aspects of real basketball games are not captured by this simple random walk approach?

7. Following the class discussion, explain the connection between this model and ELO scores.

Extra Credit Problem: *ELO Scores and Random Walks*

This problem explores the connection between ELO scoring systems and random walk models in sports. Consider two teams, A and B, whose performance in a match can be modeled as independent random walks. Let μ_A and μ_B be the mean performance of teams A and B respectively, and σ_A^2 and σ_B^2 their variances.

1. Show that the distribution of the score difference $\Delta s = s_B - s_A$ follows a random walk with mean $\tilde{\mu} = \mu_B - \mu_A$ and variance $\tilde{\sigma} = \sigma_A^2 + \sigma_B^2$.

2. A basketball match runs for a specified time T. Argue about why the distribution of Δs at the end of the match follows a Gaussian distribution. What are its mean and variance?

3. The probability of team A winning can be expressed as $P(A \text{ wins}) = P(\Delta s < 0)$. Using results from the previous questions, express this probability in terms of the cumulative distribution function of a standard normal distribution.

4. Plot this probability as a function of $\tilde{\mu} = \mu_B - \mu_A$, assuming $\sigma_A^2 = \sigma_B^2 = 1$. What shape does this curve take?

5. Recall that in the ELO rating system, the probability of team A (with rating R_A) winning against team B (with rating R_B) is given by:

$$P(A \text{ wins}) = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

   Compare this to your plot from part (4). What is the relationship between the ELO rating difference and the mean performance difference in the random walk model?

6. Now consider the case where the random walks of the two teams have standard deviations $\sigma_A = 1$ and $\sigma_B = 3$. Plot the probability of team A winning as a function of $\tilde{\mu} = \mu_B - \mu_A$. How does this curve differ from the one in part (4)? And from the ELO score formula in part (5)? (Hint: use the analytic formula based on the cumulative standard normal distribution you worked out before)

7. Under what conditions would the ELO system and the random walk model give approximately equivalent results? Discuss any assumptions required.

Problem 2: *Geometric Brownian Motion*

We discussed geometric brownian motion as a model for stock prices.

1. Pick your favorite stock (or if you don't have a favorite stock, choose a random stock) and import it using the Yahoo Finance Python API (introduced in the lecture notebook).

2. Use it to make a plot of the stock price variation over the time period 2020-2023.

3. Geometric brownian motion has two parameters, μ and σ. Find values of the parameters that are consistent with the stock price variation over the time period.

4. Use the code we gave you in class to simulate 100 instances of geometric brownian motion for this stock.

5. Are your plots consistent with the change you observe in the stock price?

Problem 3: *Kaggle Competition: Portfolio Optimization*

We are providing you with the stock prices of 10 stocks, from 2010-2022. Carry out a Markovitz style portfolio optimization to find the optimal portfolio going forward. Your answer is a vector denoting the proportion of each stock going forward, following the template of the competitio. Upload your weights as your entry to the Kaggle competition.

Please be careful in formulating your entry. We have generated these stock prices synthetically, and this is **not** as straightforward as the initial lecture notebook exercise. Our advice is to carry out the train/dev set splitting we discussed in class and validate your results yourself before submitting to the competition.