RANDOM WALKS IN TENNIS

Yaping Liu

Tennis is one of the most popular sports around the world. In the history of sports, people have always wanted to predict the winners. Knowing the past performance of two tennis players in a singles match and assuming that they will play at the same levels as in the past, how well can we predict who is going to win? To answer this question, we need to investigate how the outcome of a match depends on the point winning probabilities. In [3] a stochastic model is proposed to illustrate this relation, but no calculating formula is provided. To fully analyze the problem, we will first derive the necessary formulas and then use them to show, as one would expect, that when the top players meet in a match, the outcome is highly unpredictable.

This problem will be an excellent project for a mathematical modeling class. While the basic model is easy to understand for anyone with minimal knowledge in probability theory, the problem is open ended. Variations and extensions of the basic model will be ideal assignments for independent studies.

1. ABC of Tennis Scoring. We will briefly introduce the rules for tennis scoring. The reader is referred to [2] or other rule books for more details. Consider a singles match in which two players participate. A game is finished when one player receives at least 4 points with a two-point margin. Derived from the face values of French coins, the values of the first three points earned by a player are 15, 15, and 10, respectively. When a player wins the first point, the score is 15:0 (or 0:15). When the scores are a tie at 40-all or above, it is called a *deuce*. If the server scores (or loses) the point after deuce, it is *advantage in* (or *advantage out*). If the server wins the next point following 40:0, 40:15, 40:30, or advantage in, he takes the game.

A set is finished when one player wins at least 6 games with a two-game margin. If the score reaches 6-all, a tiebreaker game will be played to decide who takes the set. Thus, the possible winning scores in a set are: 6:0, 6:1, 6:2, 6:3, 6:4, 7:5, 7:6. Before the tiebreaking system was introduced in the early 70s, the score of a set frequently went over 20 games. For example, on June 21, 1969, in the longest singles match ever played at Wimbledon, Richard Gonzalez defeated Charles Pasarell 22-24, 1-6, 16-14, 6-3, 11-9.

A match consists of three or five sets, played as the best of three or five. If one player is much better than the other, one set will be enough to determine who is the better player and therefore the indisputable winner. If, however, the two players are very close, more sets should be played to demonstrate who is more superior. Ideally the match should be long enough so that the better player has almost a sure chance of winning and short enough so that the players will not be totally wrecked.

So what is the right number of sets to play? To shed some light on this question, it is necessary to look at the dynamics of a game first.

2. Modeling of a Game.

2.1. A Markov Chain. Denoting the two players by A and B and denoting the probability of an event X by $P(X)$, we let $p = P(\text{A wins the next point})$ and $q = P(\text{B wins the next point})$. Then $p \geq 0$, $q \geq 0$, $p + q = 1$. For simplicity, we assume that p and q are constants throughout a match, although it is possible to let p, q change values. To determine the game winning probabilities $p_1 = P(\text{A wins a game})$ and $q_1 = P(\text{B wins a game})$, we can observe the evolution of points in a game that is represented by the diagram shown below. It turns out that the diagram is actually a *finite Markov chain.*

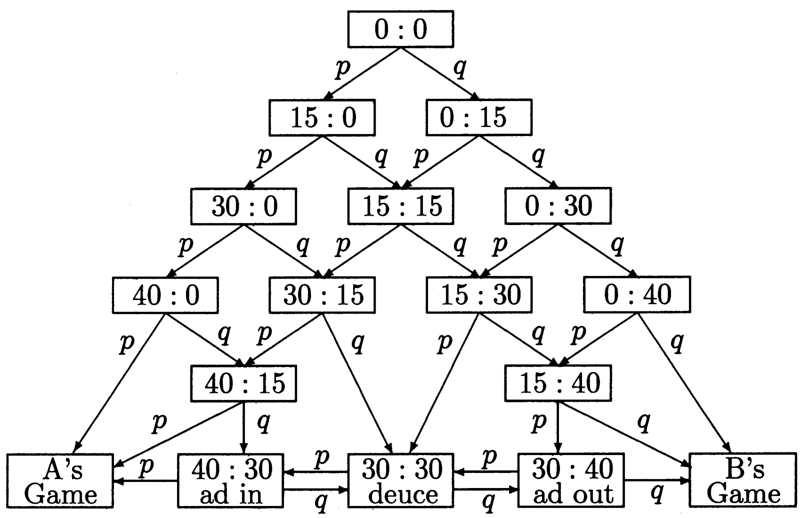

Figure 1. A Markov Chain for a Game.

The result of every exchange of points is called a *state.* To simplify the calculations we have identified the scores 40:30, 30:30, and 30:40 with the states advantage in, deuce, and advantage out, respectively. These three states may be visited any number of times before one of the players takes the game. In this way the five states in the bottom row of the above graph form a *random walk with absorbing barriers* at the two end states. See Figure 2.

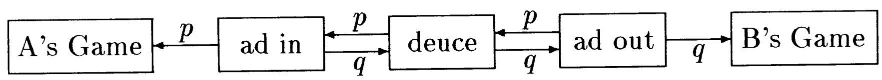

Figure 2. Random Walk for a Game.

Label the five states in the random walk by 1, 2, 3, 4, 5 from left to right. The probability that A wins a game can be calculated in two steps: first the probability of going from 0:0 directly to any state i in the bottom row of the Markov chain (call this probability u_i), then the ultimate probability of going from state i to A's game in the random walk (call this probability v_i). The vector $\mathbf{u} = (u_1, u_2, u_3, u_4, u_5)$ then represents the probability distribution of reaching the random walk from 0:0 after 4 or 5 exchanges of points. Let $\mathbf{v} = (v_1, v_2, v_3, v_4, v_5)$. It is easy to see that $p_1 = P(\text{A wins a game}) = \mathbf{u} \cdot \mathbf{v}$. The transition probabilities of going from 0:0 to any state in the Markov chain can be determined by using some basic principles in probability theory, as shown in Figure 3 below.

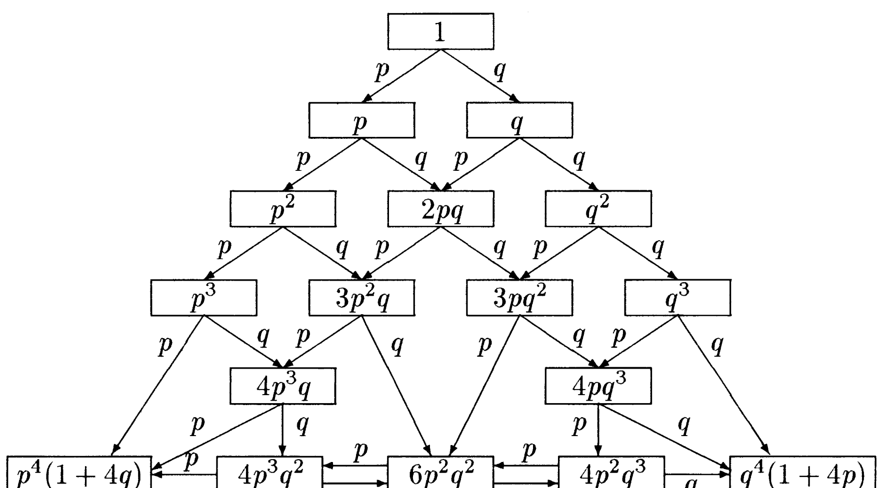

Figure 3. Transition Probabilities for a Game.

Note that the binomial theorem plays a role here so the diagram resembles Pascal's triangle. The probabilities shown in the bottom row gives

$$\mathbf{u} = (p^4(1+4q), 4p^3q^2, 6p^2q^2, 4p^2q^3, q^4(1+4p)).$$

Now a formula for the game winning probability p_1 will reveal itself if we can determine the vector $\mathbf{v}$. This can be done in a few different ways.

2.2. The Transition Matrix. One way to determine $\mathbf{v}$, suggested in [3], is to use the so-called *transition matrix* for the random walk, which is given by

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p & 0 & q & 0 & 0 \\ 0 & p & 0 & q & 0 \\ 0 & 0 & p & 0 & q \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Here the ij-entry in T is the probability of going from state i to state j in the random walk by an exchange of one point. The ultimate transition probabilities are given by the limit of the n-step transition matrix T^n: $T^\infty = \lim_{n\to\infty} T^n$ (provided it exists). Then $\mathbf{v}$ is exactly the first column of T^∞. A straightforward way to find T^∞ is to diagonalize T using eigenvalues and eigenvectors. Tedious calculations lead to

$$T^\infty = \frac{1}{p^2+q^2}\begin{pmatrix} p^2+q^2 & 0 & 0 & 0 & 0 \\ p(1-pq) & 0 & 0 & 0 & q^3 \\ p^2 & 0 & 0 & 0 & q^2 \\ p^3 & 0 & 0 & 0 & q(1-pq) \\ 0 & 0 & 0 & 0 & p^2+q^2 \end{pmatrix}.$$

It follows that

$$\mathbf{v} = \frac{1}{p^2+q^2}(p^2+q^2, p(1-pq), p^2, p^3, 0).$$

2.3. Gambler's Ruin. Another way to determine the vector $\mathbf{v}$ is to transform the problem into an equivalent but familiar problem, the so-called *gambler's ruin*, which is stated below:

Two gambler's, A and B, are betting on the outcome of coin tosses. A always bets a head which appears with probability p. B always bets a tail which appears with probability $q = 1-p$. The total fortune of the two gamblers is m dollars. The bet for each toss is one dollar. The game continues until one of the gamblers goes broke.

Let $w_k = P$(A starts with k dollars and wins it all). It is known that if $p \neq q$, then

$$w_k = \frac{p^{m-k}(p^k - q^k)}{p^m - q^m},$$

while if $p = q = 0.5$, then $w_k = k/m$. In the case $p > q$, $\lim_{k\to\infty} w_k = 1$. Thus, if the game is in favor of A and A is rich, there is little chance that B can win in the long run, even if B is also rich. As a famous principle puts it: the favored rich gets richer.

Gambler's ruin can be modeled as a random walk with absorbing barriers at 0 and m, in which state i means that B has i dollars in possession. State 0 will be reached when A wins it all and state m will be reached when B wins it all. See Figure 4.

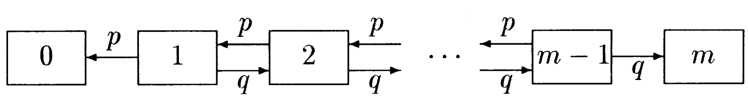

Figure 4. Random Walk for Gambler's Ruin.

If $m = 4$, then gambler's ruin is the same random walk as for a tennis game. Using the formulas above, we have

$$w_0 = 0,\ w_1 = \frac{p^3}{p^2+q^2},\ w_2 = \frac{p^2}{p^2+q^2},\ w_3 = \frac{p(1-pq)}{p^2+q^2},\ w_4 = 1.$$

Note that $v_i = w_{5-i}$. This confirms the fact that

$$\mathbf{v} = \frac{1}{p^2+q^2}(p^2+q^2, p(1-pq), p^2, p^3, 0).$$

2.4. There is a Better Way. Probably the easiest way to determine $\mathbf{v}$ is to solve a system of linear equations. A moment reflection on the random walk will convince us that

$$v_1 = 1,\ v_2 = p + qv_3,\ v_3 = pv_2 + qv_4,\ v_4 = pv_3,\ v_5 = 0.$$

Solving this system we obtain

$$\mathbf{v} = \frac{1}{p^2 + q^2}(p^2 + q^2, p(1 - pq), p^2, p^3, 0).$$

More generally, one can use this idea to derive a formula for the "ultimate ruin" probabilities in a random walk with absorbing barriers [1].

2.5. Game Winning Probabilities. From the relation $p_1 = \mathbf{u} \cdot \mathbf{v}$ we have

$$p_1 = P(\text{A wins a game}) = \frac{p^4(1 + 2q + 4q^2 + 8q^3)}{p^2 + q^2}.$$

Symmetrically

$$q_1 = P(\text{B wins a game}) = \frac{q^4(1 + 2p + 4p^2 + 8p^3)}{p^2 + q^2}.$$

Check that $p_1 + q_1 = 1$ for all valid choices of p and q.

We conclude that more games should be played in order to determine the indisputable winner. For $p = 0.6$, A is a much better player than B, yet $p_1 \approx 0.7357$, so B still can win about 1 game out of 4 on the average.

3. Modeling of a Set. We will only consider a set in which no tiebreaking system is administered. The modeling for a set with tiebreaker is not much more complicated (see the last section for some remarks on tiebreaker games). In much the same way one can obtain a Markov chain in which the five states in the bottom row form a random walk of the same type as in Figure 2. The probability distribution of reaching the random walk from 0:0 after 5–8 exchanges of games is

$$\mathbf{u} = (p_1^6(1 + 6q_1 + 21q_1^2 + 56q_1^3), 56p_1^5q_1^4, 70p_1^4q_1^4, 56p_1^4q_1^5, q_1^6(1 + 6p_1 + 21p_1^2 + 56p_1^3))$$

while the probabilities of going from any state in the random walk to A's set are given in

$$\mathbf{v} = \frac{1}{p_1^2+q_1^2}(p_1^2+q_1^2, p_1(1-p_1q_1), p_1^2, p_1^3, 0).$$

It follows that

$$p_2 = P(\text{A wins a set}) = \frac{p_1^6(1+4q_1+11q_1^2+26q_1^3+56q_1^4+112q_1^5)}{p_1^2+q_1^2}.$$

We conclude that if one player is clearly better than the other (say $p = 0.6$), they only have to play one set to determine who should win ($p_2 \approx 0.9661$). However, if the two players are closely matched (say $p = 0.55$), the outcome will not be convincing enough because the winning probability p_2 is not sufficiently close to 1 ($p_2 \approx 0.8216$). The following figure illustrates how the game winning probability function p_1 and the set winning probability function p_2 depend on the point winning probability p.

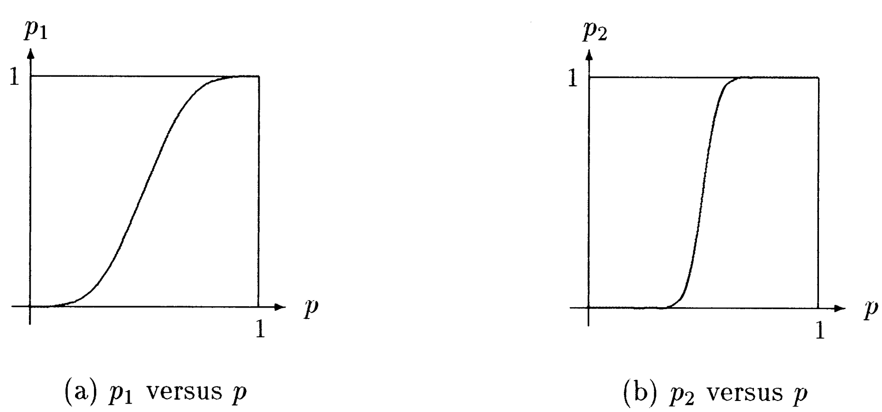

Figure 5. The Game and Set Winning Probability Functions.

4. Modeling of a 5-Set Match. A 5-set match is played as the best of 5. Whoever wins three sets first takes the match. As a result, there is no random walk in the following Markov chain.

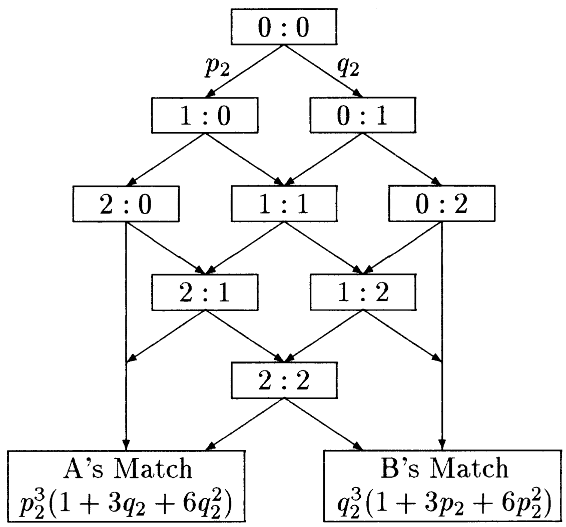

Figure 6. Markov Chain for a 5-Set Match.

The probability for winning a match is simply $p_3 = P(\text{A wins a match}) = p_2^3(1+3q_2+6q_2^2)$. Numerical calculation indicates that for players who are very closely matched (say $p = 0.51$), the outcome of a 5-set match is far from being convincing. One may conclude that the players must continue to play. Should seven sets be played? Most people will probably say no because a 7-set match will last for too long. In addition, a 7-set match will not improve the result very much. Fixing $p = 0.51$, we have

- for a 3-set match, $p_3 \approx 0.6093$,
- for a 5-set match, $p_3 \approx 0.6357$,
- for a 7-set match, $p_3 \approx 0.6571$.

5. Some Remarks. Let n be the number of sets played in a match. One can see that as $n \to \infty$, the graph of p_3 versus p approaches that of the heavyside function. For any fixed $p > 0.5$, it will be interesting to estimate the rate at which $p_3 \to 1$ as $n \to \infty$.

In the modeling of a set earlier, we have ignored the tiebreaking system. When the score reaches 6-all in a set, a tiebreaker game will be played in which one player takes the game when he receives at least 7 points with a 2-point margin. (This is the so-called Wimbledon method. There are other tiebreaking systems.) The modeling of a tiebreaker game again results in a Markov chain with a random walk. In the following table we compare the probabilities of winning a match with or without a tiebreaking system administered (we use $\bar{p}_3$ for a match with tiebreakers).

p	p_3	$\overline{p}_3$
0.60	0.9996	0.9995
0.55	0.9574	0.9530
0.51	0.6357	0.6317

It should be evident that the sole purpose of the tiebreaking system is to shorten the duration of a match. Playing tiebreaker games has almost no impact on the final outcome, which is exactly what we wanted.

One may also consider the difference being a server or a receiver in a game. Serving usually gives a player an edge. Since the two players serve alternately, the point winning probability p should have two different values, which results in two different values for p_1. So the formula for p_2 should be modified.

Finally, it is possible to incorporate more features into the model to reflect psychological, emotional, and other factors such as a player's strategy and physical endurance. The problem is open ended in this regard.

References

1. G. R. Grimmett and D. R. Stirzaker, *Probability and Random Process*, 2nd edition, Clarendon Press, Oxford, 1992.

2. J. D. Johnson and P. J. Xanthos, *Tennis*, 5th edition, Wm. C. Brown Publishers, Dubuque, Iowa, 1988.

3. L. E. Sadovskiĭ and A. L. Sadovskiĭ, *Mathematics and Sports*, American Mathematical Society, 1993.

Yaping Liu
Department of Mathematics
Pittsburg State University
Pittsburg, Kansas 66762
email: yliu@pittstate.edu