Probability of Winning at Tennis I. Theory and Data

By Paul K. Newton and Joseph B. Keller

The probability of winning a game, a set, and a match in tennis are computed, based on each player's probability of winning a point on serve, which we assume are independent identically distributed (iid) random variables. Both two out of three and three out of five set matches are considered, allowing a 13-point tiebreaker in each set, if necessary. As a by-product of these formulas, we give an explicit proof that the probability of winning a set, and hence a match, is independent of which player serves first. Then, the probability of each player winning a 128-player tournament is calculated. Data from the 2002 U.S. Open and Wimbledon tournaments are used both to validate the theory as well as to show how predictions can be made regarding the ultimate tournament champion. We finish with a brief discussion of evidence for non-iid effects in tennis, and indicate how one could extend the current theory to incorporate such features.

1. Introduction

We wish to calculate the probability that one player, A, wins a tennis match against another player B. It is not enough to know the rankings of A and B, because there is no unambiguous way to translate rankings into probabilities of winning [1, 2]. However, it does suffice to know the probability p_A^R that A

Address for correspondence: Paul K. Newton, Department of Aerospace and Mechanical Engineering and Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1191; e-mail: newton@spock.usc.edu

Published by Blackwell Publishing, 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ, UK.

wins a rally when A serves, and the probability p_B^R that B wins a rally when B serves. Such probabilities have been used to calculate the probability of winning a game in other racquet sports, such as racquetball [3], squash [4], and badminton [5]. Models of this type for tennis were first considered by Hsi and Burych [6], followed by Carter and Crews [7], and Pollard [8]. All of these analyses, including ours, treat points in tennis as independent identically distributed (iid) random variables, hence p_A^R and p_B^R are taken as constant throughout a match. A recent statistical analysis of 4 years of Wimbledon data [9] shows that although points in tennis are not iid, for most purposes this is not a bad assumption as the divergence from iid is small. Other aspects of tennis that have been analyzed using probabilistic models include optimal serving strategies [10], the efficiency of various scoring systems [11], and the question of which is the most important point [12]. Statistical methods have also been used to study the effects of new balls [13], service dominance [14], and the probabilities of winning the final set of a match [15].

Our formulation unifies and extends some of the previous treatments by the use of hierarchical recurrence relations whose solutions yield the probability that A wins a game, a set, or a match against B in terms of p_A^R and p_B^R. We then calculate the probability that a player in a 128 player single elimination tournament reaches the second, third, $\ldots$, or final round, and the probability that a player who has reached the nth round will win the tournament. We also provide an explicit proof, based on the solutions of our recurrence relations, that the probability of winning a set or match does not depend on which player serves first.

Of course the probability p_A^R that A wins a rally on serve depends upon the opponent B as well as upon A. If data are not available for A serving to B, then data for A playing against players similar to B can be used. We illustrate this point with data from the 2002 U.S. Open Men's and Women's Singles Tournaments, and from the 2002 Wimbledon Men's and Women's Singles Tournaments. The data, shown in Tables 1 and 2, and in Figure 1, agree well with our theoretical calculation of p_A^G, the probability that A wins a game when A serves. In a companion paper (part II), we will compare the theory with Monte Carlo simulations.

A game in tennis is played with one player serving. The game is won by the first player to score four or more points and to be at least two points ahead of the other player. In a set, the players serve alternate games until a player wins at least six games and is ahead by at least two games. If the game score reaches 6–6, a 13-point tiebreaker is used to determine who wins the set, with the player who started serving the set serving the first point of the tiebreaker.[1] Then, the

[1] In the U.S. Open, a tiebreaker is used in every set, whereas in Wimbledon, in the French Open, and in the Australian Open, tiebreakers are not used in the third set of a two out of three set match (women's format), or the fifth set in a three out of five set match (men's format).

Table 1
Data for the Semifinalists in the 2002 U.S. Open Tournament

Player	A	B	C	D	E	F	G
Women							
S. Williams	240	349	52	57	0.69	0.91	0.89
V. Williams	270	428	56	70	0.63	0.8	0.79
L. Davenport	206	301	45	53	0.68	0.85	0.88
A. Mauresmo	287	457	58	75	0.63	0.77	0.79
Men							
P. Sampras	573	781	124	130	0.73	0.95	0.93
A. Agassi	443	676	96	110	0.66	0.87	0.85
L. Hewitt	436	654	91	107	0.67	0.85	0.86
S. Schalken	519	768	107	119	0.68	0.9	0.88

Column A: points won on serve; Column B: total points served; Column C: games won on serve; Column D: total games served; Column E: empirical probability p_A^R of winning a rally on serve = A/B; Column F: empirical probability p_A^G of winning a game on serve = C/D; Column G: theoretical probability p_A^G of winning a game on serve, given by (5), with p_A^R from Column E.

Table 2
Data for the Semifinalists in the 2002 Wimbledon Tournament

Player	A	B	C	D	E	F	G
Women							
S. Williams	276	390	57	64	0.71	0.89	0.91
V. Williams	273	352	51	62	0.67	0.82	0.86
J. Henin	252	427	48	66	0.59	0.73	0.71
A. Mauresmo	241	378	50	57	0.64	0.88	0.81
Men							
L. Hewitt	450	646	96	107	0.70	0.90	0.90
D. Nalbandian	516	847	94	128	0.61	0.73	0.76
T. Henman	457	683	92	110	0.67	0.84	0.86
X. Malisse	483	721	101	114	0.67	0.89	0.86

Column A: points won on serve; Column B: total points served; Column C: games won on serve; Column D: total games served; Column E: empirical probability p_A^R of winning a rally on serve = A/B; Column F: empirical probability p_A^G of winning a game on serve = C/D; Column G: theoretical probability p_A^G of winning a game on serve, given by (5), with p_A^R from Column E.

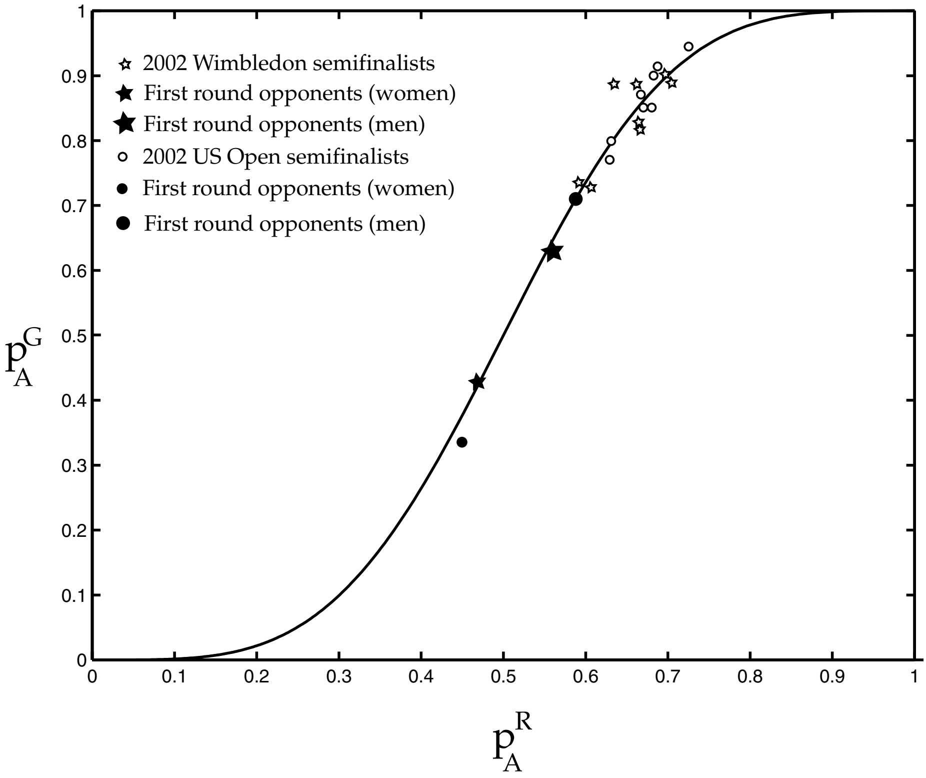

Figure 1. The probability p_A^G of A winning a game when A serves, i.e., of holding serve, as a function of p_A^R based on (6). The open circles correspond to data from eight semifinalists in the 2002 U.S. Open Men's and Women's Singles Tournaments and the open stars correspond to data from eight semifinalists in the 2002 Wimbledon Men's and Women's Singles Tournaments. The four left most data points represent the combined data from the semifinalists' first round opponents in each tournament.

players alternate serves, each serving two consecutive points, until someone wins at least seven points, and is ahead by at least two points. The winner of the tiebreaker wins the set with seven games to the opponents six games. To win a match, a player must win two out of three sets (women's format), or win three out of five sets (men's format), with the two players serving alternate games throughout the match. The initial server in the match is determined by a coin toss, with the winner given the choice of serving first or receiving first.

2. Probability of winning a game

Player A can win a game against player B by a score of (4, 0), (4, 1) or (4, 2), or else the score can become $(3, 3)$, called "deuce." Then, A can win by getting two points ahead of B, with a score of $(n + 5, n + 3)$ with $n \geq 0$. To calculate the probability p_A^G that A wins a game when A serves, we assume that p_A^R is the probability that A wins a rally when A serves, and set $q_A^R = 1 - p_A^R$, $q_A^G = 1 - p_A^G$. We also denote by $p_A^G(i, j)$ the probability that the score reaches i points for A and j points for B when A serves. Upon summing the probabilities of the different ways in which A can win, we get

$$p_A^G = \sum_{j=0}^{2} p_A^G(4, j) + p_A^G(3, 3) \sum_{n=0}^{\infty} p_A^{DG}(n + 2, n). \tag{1}$$

Here, $p_A^{DG}(n + 2, n)$ is the probability that A wins by scoring $n + 2$ while B scores n after deuce has been reached, with A serving. It is given by

$$\begin{aligned} p_A^{DG}(n + 2, n) &= \sum_{j=0}^{n} \left(p_A^R q_A^R\right)^j \left(q_A^R p_A^R\right)^{n-j} \frac{n!}{j!(n-j)!} \left(p_A^R\right)^2 \\ &= \left(p_A^R\right)^2 \left[p_A^R q_A^R\right]^n 2^n. \end{aligned} \tag{2}$$

Upon using (2) in (1), and summing the geometric series, we get

$$p_A^G = \sum_{j=0}^{2} p_A^G(4, j) + p_A^G(3, 3) \left(p_A^R\right)^2 \left[1 - 2p_A^R q_A^R\right]^{-1}. \tag{3}$$

Elementary combinatorial analysis yields

$$\begin{aligned} &p_A^G(4, 0) = \left(p_A^R\right)^4, \quad p_A^G(4, 1) = 4\left(p_A^R\right)^4 q_A^R, \quad p_A^G(4, 2) = \frac{5 \cdot 4}{2} \left(p_A^R\right)^4 \left(q_A^R\right)^2, \\ &p_A^G(3, 3) = \frac{6!}{(3!)^2} \left(p_A^R q_A^R\right)^3. \end{aligned} \tag{4}$$

Now using (4) in (3) gives the probability that A wins a game when A serves, i.e., that A holds serve:

$$p_A^G = \left(p_A^R\right)^4 \left[1 + 4q_A^R + 10\left(q_A^R\right)^2\right] + 20\left(p_A^R q_A^R\right)^3 \left(p_A^R\right)^2 \left[1 - 2p_A^R q_A^R\right]^{-1}. \tag{5}$$

This equation agrees with that given in [7]. Figure 1 shows p_A^G as a function of p_A^R, based upon (5). The open circles in the figure are data for the semifinalists in the 2002 U.S. Open Men's and Women's Singles Tournaments, shown in Table 1, while the stars are data for the semifinalists in the 2002 Wimbledon Men's and Women's Singles Tournaments shown in Table 2. The left most four points are totals for their first round opponents in both tournaments. They all lie close to the theoretical curve.

3. Probability of winning a set

3.1. Recursion equations

Let p_A^S denote the probability that player A wins a set against player B, with A serving first, and $q_A^S = 1 - p_A^S$. To calculate p_A^S in terms of p_A^G and p_B^G, we define $p_A^S(i, j)$ as the probability that in a set, the score becomes i games for A, j games for B, with A serving initially. Then,

$$p_A^S = \sum_{j=0}^{4} p_A^S(6, j) + p_A^S(7, 5) + p_A^S(6, 6) p_A^T. \tag{6}$$

Here, p_A^T is the probability that A wins a 13-point tiebreaker with A serving initially, and $q_A^T = 1 - p_A^T$.

To calculate $p_A^S(i, j)$, needed in (6), we use the following recursion formulas and initial conditions:

For $0 \le i, j \le 6$:

$$\begin{aligned} &\textbf{if } i - 1 + j \textbf{ is even:} \quad p_A^S(i, j) = p_A^S(i-1, j) p_A^G + p_A^S(i, j-1) q_A^G \\ &\qquad\qquad \textbf{omit } i - 1 \textbf{ term if } j = 6, i \le 5; \\ &\qquad\qquad \textbf{omit } j - 1 \textbf{ term if } i = 6, j \le 5 \end{aligned} \tag{7}$$

$$\begin{aligned} &\textbf{if } i - 1 + j \textbf{ is odd:} \quad p_A^S(i, j) = p_A^S(i-1, j) q_B^G + p_A^S(i, j-1) p_B^G \\ &\qquad\qquad \textbf{omit } i - 1 \textbf{ term if } j = 6, i \le 5; \\ &\qquad\qquad \textbf{omit } j - 1 \textbf{ term if } i = 6, j \le 5 \end{aligned} \tag{8}$$

Initial conditions:

$$p_A^S(0, 0) = 1; \quad p_A^S(i, j) = 0 \qquad \text{if } i < 0, \text{or } j < 0. \tag{9}$$

In terms of p_A^S (6, 5) and $p_A^S(5, 6)$, we have

$$p_A^S(7, 5) = p_A^S(6, 5) q_B^G; \quad p_A^S(5, 7) = p_A^S(5, 6) p_B^G. \tag{10}$$

The explicit solution of (7)–(10) is given in the Appendix.

3.2. Probability of winning a tiebreaker

To calculate p_A^T in terms of p_A^R and p_B^R, we define $p_A^T(i, j)$ to be the probability that the score becomes i for A, j for B in a tiebreaker with A serving initially. Then,

$$p_A^T = \sum_{j=0}^{5} p_A^T(7, j) + p_A^T(6, 6) \sum_{n=0}^{\infty} p_A^T(n + 2, n). \tag{11}$$

Because the sequence of serves in a tiebreaker is *A*, *BB*, *AA*, *BB*, etc., we have

$$p_A^T(n+2,n) = \sum_{j=0}^{n} \left(p_A^R p_B^R\right)^j \left(q_A^R q_B^R\right)^{n-j} \frac{n!}{j!(n-j)!} p_A^R q_B^R$$
$$= \left(p_A^R p_B^R + q_A^R q_B^R\right)^n p_A^R q_B^R. \tag{12}$$

Using (12) in (11) and summing yields

$$p_A^T = \sum_{j=0}^{5} p_A^T(7,j) + p_A^T(6,6) p_A^R q_B^R \left[1 - p_A^R p_B^R - q_A^R q_B^R\right]^{-1} \tag{13}$$

To calculate $p_A^T(i,\ j)$, we use the recursion formulas:
For $0 \le i,\ j \le 7$:

$$\textbf{if } i-1+j = 0,3,4,\ldots,4n-1,4n,\ldots$$
$$p_A^T(i,j) = p_A^T(i-1,j)p_A^R + p_A^T(i,j-1)q_A^R$$
$$\textbf{omit } j-1 \textbf{ term if } i=7, j \le 6$$
$$\textbf{omit } i-1 \textbf{ term if } j=7, i \le 6 \tag{14}$$

$$\textbf{if } i-1+j = 1,2,5,6,\ldots,4n+1,4n+2,\ldots$$
$$p_A^T(i,j) = p_A^T(i-1,j)q_B^R + p_A^T(i,j-1)p_B^R$$
$$\textbf{omit } j-1 \textbf{ term if } i=7, j \le 6$$
$$\textbf{omit } i-1 \textbf{ term if } j=7, i \le 6 \tag{15}$$

Initial conditions:

$$p_A^T(0,0) = 1; \quad p_A^T(i,j) = 0 \qquad \text{if } i < 0, \text{or } j < 0. \tag{16}$$

The solution of (14)–(16) is given in the Appendix.

Next we calculate p_A^T by using the solution of (14)–(16) in (13). Now we can calculate p_A^S by using the solution of (7)–(9), and (10), with the result for p_A^T, in (6).

Figure 2 shows the probability of player *A* winning a set against player *B* plotted as a function of $p_A^R \in [0, 1]$ for the full range values of p_B^R in increments of 0.1. The data shown are compiled from the 2002 U.S. Open Men's Singles event. Of the 117 completed matches played, there were 9 matches in which a player (designated player *B*) had a value of $p_B^R = 0.50 \pm 0.01$, 33 matches in which $p_B^R = 0.60 \pm 0.01$, and 20 matches with $p_B^R = 0.70 \pm 0.01$. Because each match involves three, four, or five sets, it is necessary to combine data from several matches to get meaningful statistics. Hence, each data point shown in the figure represents a compilation of several matches grouped according to corresponding values of p_A^R. Each of the three data points associated with the

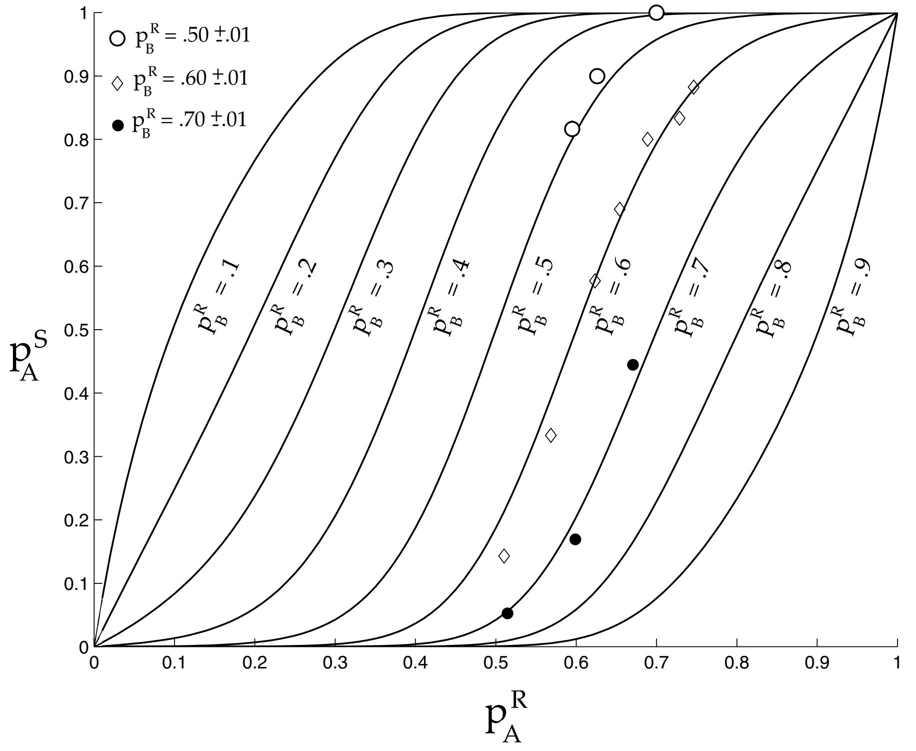

Figure 2. The probability p_A^S of player A winning a set plotted as a function of p_A^R for various values of p_B^R. Compiled data from the 2002 U.S. Open Men's Singles event are shown for the values $p_B^R = 0.50 \pm 0.01$, $p_B^R = 0.60 \pm 0.01$, and $p_B^R = 0.70 \pm 0.01$.

curve marked $p_B^R = 0.50$ represents three matches, each of the seven data points associated with the curve marked $p_B^R = 0.60$ represents approximately five matches, while each of the three data points associated with the curve marked $p_B^R = 0.70$ represents a compilation of approximately seven matches. Given the relatively small number of sets underlying each of the data points, the data fits the theoretical curves reasonably well. Figure 3 shows the probability of player A winning a tiebreaker against player B plotted as a function of $p_A^R \in [0, 1]$ for the full range values of p_B^R in increments of 0.1.

3.3. Serving or receiving first

In this section, we prove that there is no theoretical advantage to serving first by showing that the probability of player A winning the set when serving first, p_A^S, is equal to his probability of winning the set when receiving first, q_B^S. For this, we need formula (6) for p_A^S, along with the corresponding formula for q_B^S given by

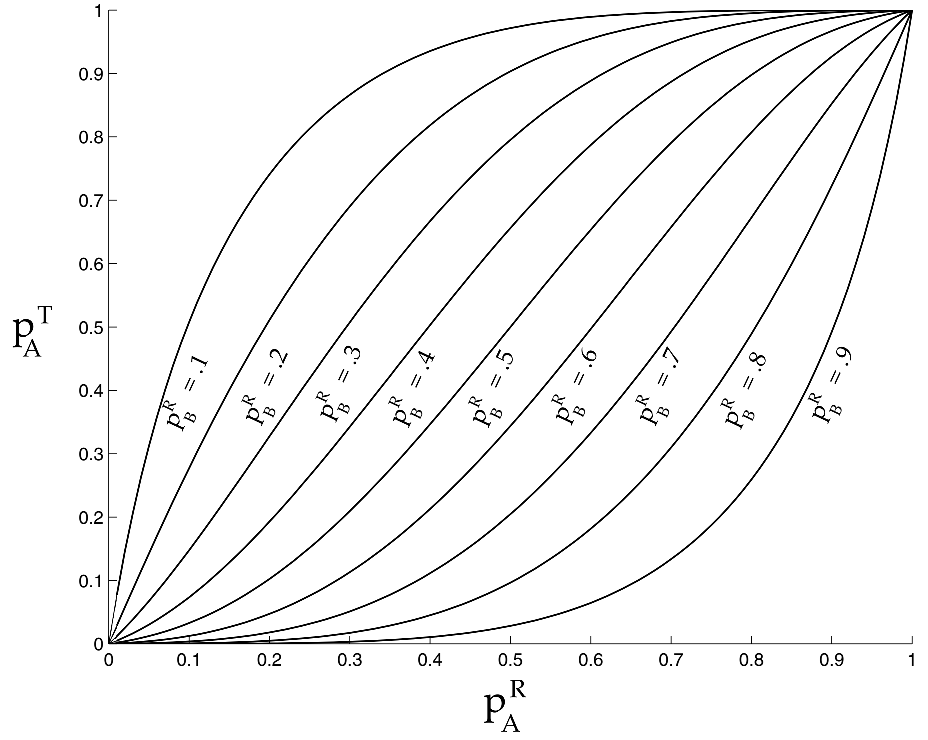

Figure 3. The probability p_A^T of player A winning a tiebreaker plotted as a function of p_A^R for various values of p_B^R.

$$q_B^S = \sum_{j=0}^{4} p_B^S(j, 6) + p_B^S(5, 7) + p_B^S(6, 6)q_B^T. \tag{17}$$

We obtain the terms $p_B^S(j, i)$ in (17) from $p_A^S(i, j)$ given in the Appendix, by interchanging $p_A^G \leftrightarrow q_B^G$, $p_B^G \leftrightarrow q_A^G$. From (A.1) and (A.6) it is immediate that

$$p_A^S(6, 0) = p_B^S(0, 6) \tag{18}$$

$$p_A^S(7, 5) = p_B^S(5, 7). \tag{19}$$

It is also clear from (A.7) that

$$p_A^S(6, 6) = p_B^S(6, 6). \tag{20}$$

Thus, it remains to show that

$$\sum_{j=1}^{4} p_A^S(6, j) = \sum_{j=1}^{4} p_B^S(j, 6) \tag{21}$$

and that

$$p_A^T = q_B^T. \tag{22}$$

To prove (21), we show that

$$p_A^S(6, 1) + p_A^S(6, 2) = p_B^S(1, 6) + p_B^S(2, 6) \tag{23}$$

and

$$p_A^S(6, 3) + p_A^S(6, 4) = p_B^S(3, 6) + p_B^S(4, 6). \tag{24}$$

By using formulas (A.2)–(A.5), and replacing $q_A^G = 1 - p_A^G$, $q_B^G = 1 - p_B^G$, we can write

$$p_A^S(6, 2n-1) + p_A^S(6, 2n) = \sum_{i=0}^{6+2n}\sum_{j=0}^{6+2n} a_{ij}^S(n)\left(p_A^G\right)^i\left(p_B^G\right)^j \tag{25}$$

$$p_B^S(2n-1, 6) + p_B^S(2n, 6) = \sum_{i=0}^{6+2n}\sum_{j=0}^{6+2n} b_{ij}^S(n)\left(p_A^G\right)^i\left(p_B^G\right)^j \tag{26}$$

for $n = 1, 2$. Then, it can be shown that the coefficients of each are equal, i.e., $a_{ij}^S(n) = b_{ij}^S(n)$. The values are listed in the Appendix. Figure 4 shows the probability of obtaining each of the scores that are independent of which player serves first for the case of evenly matched players.

To prove that $p_A^T = q_B^T$, we use the formula (11) for p_A^T and the corresponding one for q_B^T

$$q_B^T = \sum_{j=0}^{5} p_B^T(j, 7) + p_B^T(6, 6)\sum_{n=0}^{\infty} p_B^T(n, n+2). \tag{27}$$

We obtain the terms $p_B^T(j, i)$ in (27) from $p_A^T(i, j)$ given in the Appendix, by interchanging $p_A^R \leftrightarrow q_B^R$, $p_B^R \leftrightarrow q_A^R$. From (A.14) it is clear that $p_A^T(6, 6) = p_B^T(6, 6)$. Furthermore, from the symmetry under exchanging $p_A^R \leftrightarrow q_B^R$, $p_B^R \leftrightarrow q_A^R$ in (12), we have that

$$p_A^T(n+2, n) = p_B^T(n, n+2). \tag{28}$$

Thus, it remains to show that

$$\sum_{j=0}^{5} p_A^T(7, j) = \sum_{j=0}^{5} p_B^T(j, 7). \tag{29}$$

To prove this, we show that

$$p_A^T(7, 0) + p_A^T(7, 1) = p_B^T(0, 7) + p_B^T(1, 7), \tag{30}$$

$$p_A^T(7, 2) + p_A^T(7, 3) = p_B^T(2, 7) + p_B^T(3, 7), \tag{31}$$

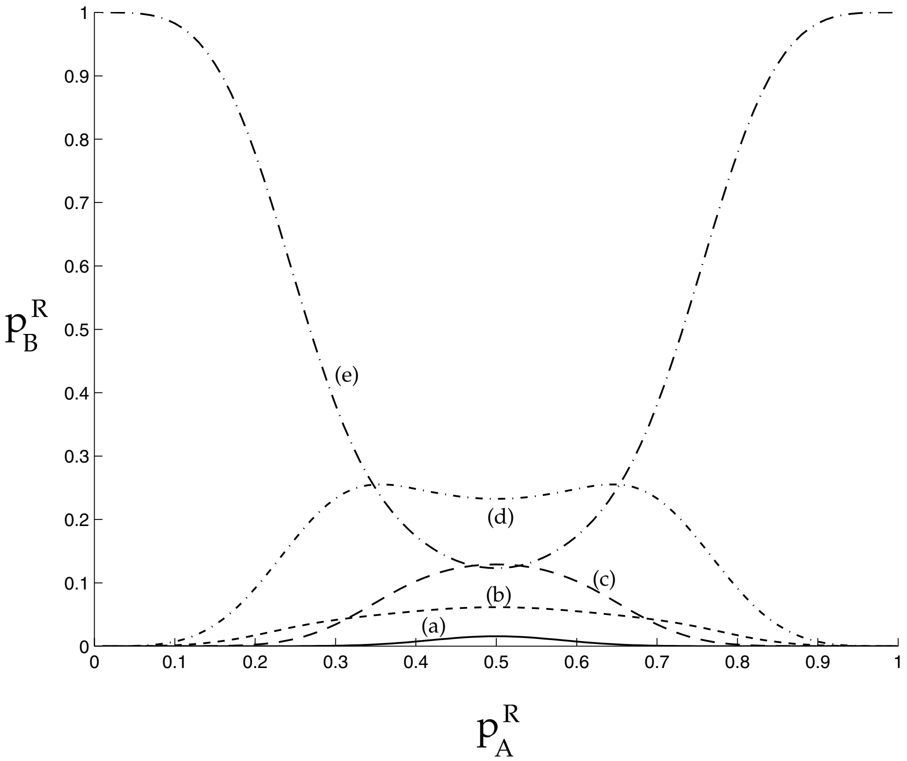

Figure 4. Set scores that are independent of which player serves first, plotted for two equal players $p_A^R = p_B^R$. (a) $p_A^S(6, 0)$, (b) $p_A^S(6, 1) + p_A^S(6, 2)$, (c) $p_A^S(6, 3) + p_A^S(6, 4)$, (d) $p_A^S(7, 5)$, and (e) $p_A^S(6, 6)$.

$$p_A^T(7,4) + p_A^T(7,5) = p_B^T(4,7) + p_B^T(5,7). \tag{32}$$

By using formulas (A.8)–(A.13) and replacing $q_A^R = 1 - p_A^R$, $q_B^R = 1 - p_B^R$, we can write

$$p_A^T(7,2n) + p_A^T(7,2n+1) = \sum_{i=0}^{4}\sum_{j=0}^{4} a_{ij}^T(n)\left(p_A^R\right)^i \left(p_B^R\right)^j \tag{33}$$

$$p_B^T(2n,7) + p_B^T(2n+1,7) = \sum_{i=0}^{4}\sum_{j=0}^{4} b_{ij}^T(n)\left(p_A^R\right)^i \left(p_B^R\right)^j \tag{34}$$

for $n = 0, 1, 2$. Then, it can be shown that the coefficients are equal, i.e., $a_{ij}^T(n) = b_{ij}^T(n)$. The values are listed in the Appendix. Figure 5 shows the probability of obtaining each of the tiebreaker scores that are independent of which player serves first, for equally matched players.

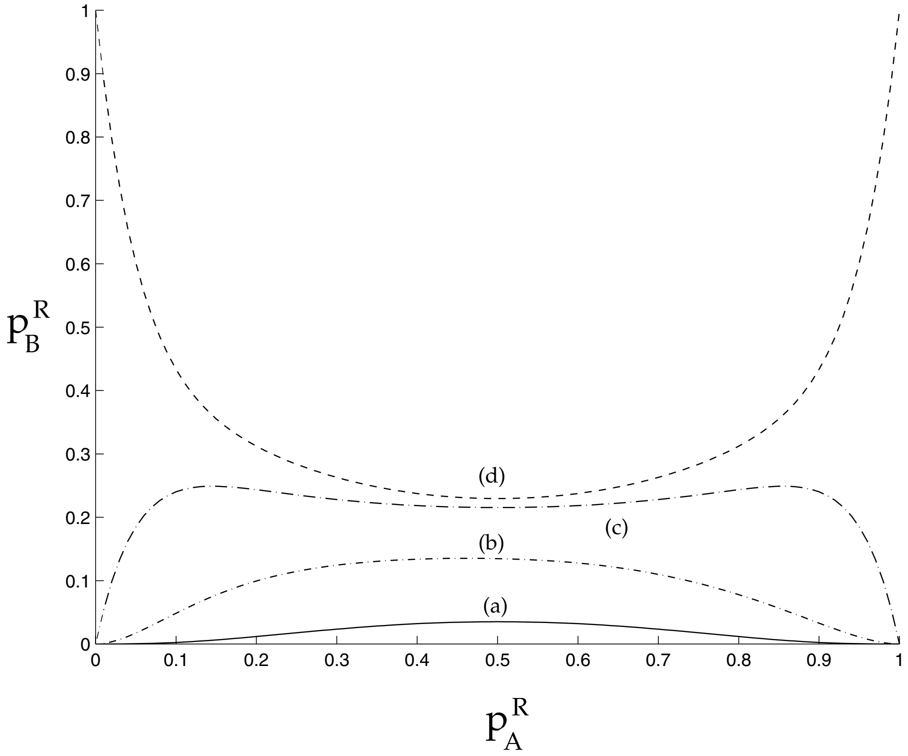

Figure 5. Tiebreaker scores that are independent of which player serves first, plotted for two equal players $p_A^R = p_B^R$. (a) $p_A^T(7, 0) + p_A^T(7, 1)$, (b) $p_A^T(7, 2) + p_A^T(7, 3)$, (c) $p_A^T(7, 4) + p_A^T(7, 5)$, and (d) $p_A^T(6, 6)$.

The question of whether to serve or receive first has received some attention in the literature. In an interesting combinatorial analysis of Kingston [16] (followed by a note [17]), a simplified scoring system (which he calls a "short set") is considered in which player A serves the first game of a match consisting of the best N of $2N - 1$ games. His striking result is that it does not matter whether the rules are such that the players alternate serves after each game, or whether the winner of the previous game continues to serve the next game. In either case, player A has the same probability of winning. At the end of the article, he asks how many games need to be played to give two equal players a reasonably equal chance of winning, *whoever* starts serving. As a consequence of the central limit theorem, player A's (approximate) probability of winning a short set is $\frac{1}{2} + \frac{1}{2}(p_A^R - \frac{1}{2})[\pi p_A^R(1 - p_A^R)(N - 1)]^{-1/2}$. Figure 2 in his paper shows the slow convergence to $\frac{1}{2}$ as $N \to \infty$, giving player A a distinct advantage, for finite N, if he serves first and $p_A^R > 0.5$. Thus, for *best* N of $2N - 1$ scoring, there is a theoretical advantage to serving first. For tennis scoring, the paper of Pollard [8] considers both classical scoring (no tiebreakers) and tiebreaker scoring, and implicit in his calculations (see, for example, his Tables 2 and 3) is the fact that $p_A^S = q_B^S$, although the result is not proven. There are other ways of proving and generalizing the result that do not rely on the explicit solutions for p_A^S and q_B^S as our proof does. In fact, one can prove that as long as the scoring system is such that the number of games served by player A minus the number of games served by player B is 1, 0, or -1, there is no advantage or disadvantage to serving first. Such scoring systems are termed "service neutral" and are discussed in [18].

4. Probability of winning a match

We now calculate p_A^M, the probability that player A wins a match against player B, with player A serving initially, and $q_A^M = 1 - p_A^M$. To do so we define $p_{AB}^M(i, j)$ to be the probability that in a match, the score becomes i sets for A and j sets for B, with A serving initially and B serving finally. We define $p_{AA}^M(i, j)$ similarly, but with A serving initially and finally.

To formulate recursion equations for $p_{AB}^M(i, j)$ and $p_{AA}^M(i, j)$, we introduce p_{AB}^S, p_{AA}^S, p_{BA}^S, and p_{BB}^S. Here, p_{XY}^S is the probability that X wins a set when X serves the first game and Y serves the last game, where X and Y are A or B.

To get an expression for p_{AA}^S we note that when A serves the first and last games, the total number of games must be odd. Then, by restricting the right side of (6) to odd numbers of games, we get

$$p_{AA}^S = \sum_{j=1,3} p_A^S(6, j) + p_A^S(6, 6) p_A^T. \tag{35}$$

Similarly when A serves the first game and B serves the last game, the total number of games is even. For even numbers of games, the right side of (6) yields

$$p_{AB}^S = \sum_{j=0,2,4} p_A^S(6, j) + p_A^S(7, 5). \tag{36}$$

Then, (6) is written

$$p_A^S = p_{AA}^S + p_{AB}^S. \tag{37}$$

We also define q_{AA}^S and q_{AB}^S as

$$q_{AA}^S = \sum_{j=1,3} p_A^S(j, 6) + p_A^S(6, 6) q_A^T, \tag{38}$$

$$q_{AB}^S = \sum_{j=0,2,4} p_A^S(j, 6) + p_A^S(5, 7). \tag{39}$$

Table 3
Probability p_A^M of Player A Winning a Match of Three Sets out of Five

		p_A^R										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0.0	*	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.1	0.0000	0.5000	0.9060	0.9956	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.2	0.0000	0.0940	0.5000	0.9136	0.9981	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.3	0.0000	0.0045	0.0864	0.5000	0.9380	0.9993	1.0000	1.0000	1.0000	1.0000	1.0000
	0.4	0.0000	0.0000	0.0019	0.0621	0.5000	0.9513	0.9995	1.0000	1.0000	1.0000	1.0000
	0.5	0.0000	0.0000	0.0000	0.0007	0.0487	0.5000	0.9513	0.9993	1.0000	1.0000	1.0000
P_B^R	0.6	0.0000	0.0000	0.0000	0.0000	0.0005	0.0487	0.5000	0.9380	0.9981	1.0000	1.0000
	0.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0621	0.5000	0.9136	0.9956	1.0000
	0.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0019	0.0864	0.5000	0.9060	1.0000
	0.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0045	0.0940	0.5000	1.0000
	1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	*

Values of p_A^R are along the top row and values of p_B^R are down the left column.
*Indicates that the match cannot end for these values.

Then,

$$q_A^S = q_{AA}^S + q_{AB}^S. \tag{40}$$

To get p_{BA}^S, p_{BB}^S, q_{BA}^S, and q_{BB}^S, we interchange A and B in (35)–(40). Note that because $p_A^S + q_A^S = 1$ and $p_B^S + q_B^S = 1$, we have

$$p_{AA}^S + q_{AA}^S + p_{AB}^S + q_{AB}^S = 1, \tag{41}$$

$$p_{BB}^S + q_{BB}^S + p_{BA}^S + q_{BA}^S = 1. \tag{42}$$

Now we can write the recursion equations satisfied by $p_{AB}^M(i, j)$ and $p_{AA}^M(i, j)$ as follows, for $i + j > 1$:

$$\begin{aligned} p_{AB}^M(i,j) = {} & p_{AB}^M(i-1,j)p_{AB}^S + p_{AA}^M(i-1,j)q_{BB}^S \\ & + p_{AB}^M(i,j-1)q_{AB}^S + p_{AA}^M(i,j-1)p_{BB}^S, \end{aligned} \tag{43}$$

$$\begin{aligned} p_{AA}^M(i,j) = {} & p_{AB}^M(i-1,j)p_{AA}^S + p_{AA}^M(i-1,j)q_{BA}^S \\ & + p_{AB}^M(i,j-1)q_{AA}^S + p_{AA}^M(i,j-1)p_{BA}^S. \end{aligned} \tag{44}$$

The initial conditions are

$$p_{AA}^M(0,0) = 1; \quad p_{AA}^M(i,j) = 0 \quad \text{if } i < 0 \text{ or } j < 0 \tag{45}$$

$$p_{AB}^M(0,0) = 1; \quad p_{AB}^M(i,j) = 0 \quad \text{if } i < 0 \text{ or } j < 0 \tag{46}$$

$$p_{AB}^M(1,0) = p_{AB}^S; \quad p_{AB}^M(0,1) = q_{AB}^S; \quad p_{AA}^M(1,0) = p_{AA}^S; \quad p_{AA}^M(0,1) = q_{AA}^S. \tag{47}$$

For the men's format of three sets out of five, (43)–(47) must be solved for $i, j = 0, 1, 2, 3$. When $j = 3$, the $i - 1$ terms must be omitted; when $i = 3$, the $j - 1$ terms must be omitted. The probability that player A wins a three out of five set match when serving first is given by

$$p_A^M = \sum_{j=0}^{2} \left[p_{AA}^M(3,j) + p_{AB}^M(3,j)\right]. \tag{48}$$

For a match of two sets out of three, (35) and (36) must be solved for i, $j = 0, 1, 2$. When $j = 2$, the $i - 1$ terms must be omitted; when $i = 2$, the $j - 1$ terms must be omitted. Then, the probability that player A wins a two out of three set match when serving first is

$$p_A^M = \sum_{j=0}^{1} \left[p_{AA}^M(2,j) + p_{AB}^M(2,j)\right]. \tag{49}$$

By using the solutions of (43) and (44) for $p_{AA}^M(2, j)$ and $p_{AB}^M(2, j)$ and taking advantage of (37) and (40), we can write (49) as

$$\begin{aligned} p_A^M = {} & p_{AA}^S q_B^S + p_{AB}^S p_A^S + p_{AA}^S p_{BA}^S q_B^S + p_{AA}^S p_{BB}^S p_A^S + p_{AB}^S q_{AA}^S q_B^S \\ & + p_{AB}^S q_{AB}^S p_A^S + q_{AA}^S q_{BA}^S q_B^S + q_{AA}^S q_{BB}^S p_A^S + q_{AB}^S p_{AA}^S q_B^S + q_{AB}^S p_{AB}^S p_A^S. \end{aligned} \tag{50}$$

Note that because the probability of winning a set is independent of which player serves first, the above formula (50) reduces to

$$p_A^M = (p_A^S)^2 + 2(p_A^S)^2 p_B^S \tag{51}$$

for the two out of three set format, and

$$p_A^M = (p_A^S)^3 + 3(p_A^S)^3 p_B^S + 6(p_A^S)^3 (p_B^S)^2 \tag{52}$$

for the three out of five set format.

Table 3 shows p_A^M for a match of three sets out of five based upon (48), and Table 4 shows p_A^M for a match of two sets out of three based upon (40). In both cases the results are shown as functions of p_A^R and p_B^R, ranging from 0 to 1 at intervals of 0.1. Figure 6 shows the data from the 2002 U.S. Open Men's Singles event as well as the theoretical curves for p_A^G, p_A^S, and p_A^M (three out of five set format) corresponding to the value $p_B^R = 0.60$. To obtain meaningful statistics for the three data points associated with the p_A^M curve, the 33 matches were grouped in clusters of approximately 11 matches per cluster.

5. Probability of winning a tournament

5.1. The 128-player tournament

We now consider a single elimination tournament of $128 = 2^7$ players numbered $i = 1, \ldots, 128$. We assume that we know the probability p_{ij}^M for player i to defeat player j in a match. We introduce the column vector of probabilities $\mathbf{p}^{(n)} \in R^{1\times 128}$;

$$\mathbf{p}^{(n)} = \begin{pmatrix} p_1^{(n)} \\ p_2^{(n)} \\ p_3^{(n)} \\ \vdots \\ p_{128}^{(n)} \end{pmatrix}. \tag{53}$$

Here, $p_i^{(n)}$ is the conditional probability that player i wins a match in the nth round, provided that he or she survives to that round of the tournament. From (48) or (49), we know p_{ij}^M, the probability that player i beats player j, which we write more simply as P_{ij}.

Table 4
Probability p_A^M of Player A Winning a Match of Two Sets out of Three

		p_A^R										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0.0	*	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.1	0.0000	0.5000	0.8539	0.9820	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.2	0.0000	0.1461	0.5000	0.8624	0.9898	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	0.3	0.0000	0.0180	0.1376	0.5000	0.8909	0.9947	1.0000	1.0000	1.0000	1.0000	1.0000
	0.4	0.0000	0.0005	0.0103	0.1091	0.5000	0.9079	0.9961	1.0000	1.0000	1.0000	1.0000
	0.5	0.0000	0.0000	0.0001	0.0053	0.0922	0.5000	0.9079	0.9947	0.9999	1.0000	1.0000
P_B^R	0.6	0.0000	0.0000	0.0000	0.0000	0.0039	0.0922	0.5000	0.8909	0.9898	0.9995	1.0000
	0.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0053	0.1091	0.5000	0.8624	0.9820	1.0000
	0.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0103	0.1376	0.5000	0.8539	1.0000
	0.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0180	0.1461	0.5000	1.0000
	1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	*

Values of p_A^R are along the top row and values of p_B^R are down the left column.
*Indicates that the match cannot end for these values.

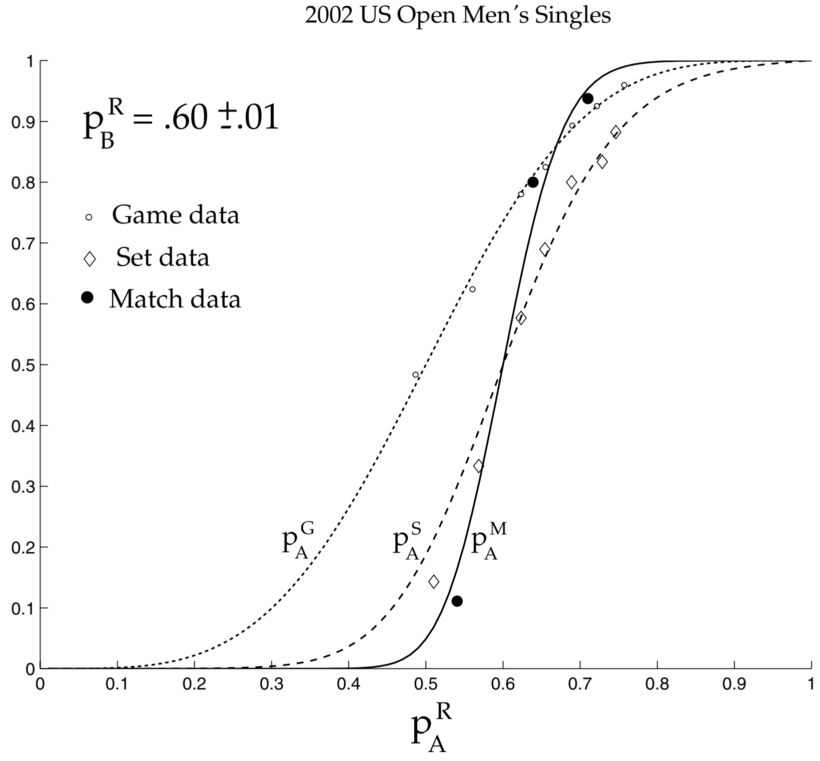

Figure 6. Theoretical curves for p_A^G (dotted), p_A^S (dashed), and p_A^M (solid) corresponding to values $p_B^R = 0.60$. Compiled data from the 2002 U.S. Open Men's Singles event are shown for all matches in which $p_B^R = 0.60 \pm 0.01$.

$\mathbf{p}^{(n)}$ satisfies the recursion formula

$$\mathbf{p}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \mathbf{p}^{(n)} = \mathbf{P}_n \mathbf{p}^{(n-1)} \quad (n = 1, \ldots, 6). \tag{54}$$

Here, $\mathbf{P}_n$ is a 128×128 matrix with block diagonal structure made up of 2^{7-n} blocks. We label them $\mathbf{P}_n^{(k)}$, $1 \le k \le 2^{7-n}$, and then P_n is given by

$$\mathbf{P}_n = \begin{bmatrix} \mathbf{P}_n^{(1)} & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{P}_n^{(2)} & 0 & \ldots & 0 \\ 0 & 0 & \mathbf{P}_n^{(3)} & \ldots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \mathbf{P}_n^{(2^{7-n})} \end{bmatrix}. \tag{55}$$

$\mathbf{P}_n^{(k)}$ is a $2^n \times 2^n$ off-diagonal block matrix:

$$\mathbf{P}_n^{(k)} = \begin{bmatrix} 0 & \mathbf{P}_{\alpha,\beta}^{(n,k)} \\ \mathbf{P}_{\alpha,\beta}^{(n,k)} & 0 \end{bmatrix}, \tag{56}$$

where $\alpha = (k-1)2^n + 1$, $\beta = k2^n$. The $\mathbf{P}_{\alpha,\beta}^{(n,k)}$ are $2^{n-1} \times 2^{n-1}$ matrices of the form

$$\mathbf{P}_{\alpha,\beta}^{(n,k)} = \begin{bmatrix} P_{\alpha,\beta+1-2^{n-1}} & \cdots & P_{\alpha,\beta-1} & P_{\alpha,\beta} \\ P_{\alpha+1,\beta+1-2^{n-1}} & \cdots & P_{\alpha+1,\beta-1} & P_{\alpha+1,\beta} \\ \vdots & \vdots & \vdots & \vdots \\ P_{\alpha+2^{n-1}-1,\beta+1-2^{n-1}} & \cdots & P_{\alpha+2^{n-1}-1,\beta-1} & P_{\alpha+2^{n-1}-1,\beta} \end{bmatrix}. \tag{57}$$

The entries of this matrix, P_{ij}, are obtained from (48) or (49).

As an example, for $n = 1$, (55) becomes

$$\mathbf{P_1} = \begin{bmatrix} \mathbf{P}_1^{(1)} & 0 & 0 & \dots & 0 \\ 0 & \mathbf{P}_1^{(2)} & 0 & \dots & 0 \\ 0 & 0 & \mathbf{P}_1^{(3)} & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{P}_1^{(64)} \end{bmatrix}. \tag{58}$$

$\mathbf{P}_1^{(k)}$ is a 2×2 matrix:

$$\mathbf{P}_1^{(k)} = \begin{bmatrix} 0 & P_{2k-1,2k} \\ P_{2k,2k-1} & 0 \end{bmatrix}. \tag{59}$$

Explicitly (59) yields

$$\mathbf{P_1^{(1)}} = \begin{bmatrix} 0 & P_{12} \\ P_{21} & 0 \end{bmatrix}, \mathbf{P_1^{(2)}} = \begin{bmatrix} 0 & P_{34} \\ P_{43} & 0 \end{bmatrix}, \dots, \mathbf{P_1^{(64)}} = \begin{bmatrix} 0 & P_{127,128} \\ P_{128,127} & 0 \end{bmatrix}. \tag{60}$$

The probability that player i ultimately becomes the tournament champion, which we denote p_i^{TC}, is the product of the conditional probabilities of winning each of the rounds. In vector form, this is given by

$$\mathbf{p}^{TC} \equiv \begin{pmatrix} p_1^{TC} \\ p_2^{TC} \\ p_3^{TC} \\ \vdots \\ p_{128}^{TC} \end{pmatrix} = \begin{pmatrix} \prod_{n=1}^{7} p_1^{(n)} \\ \prod_{n=1}^{7} p_2^{(n)} \\ \prod_{n=1}^{7} p_3^{(n)} \\ \vdots \\ \prod_{n=1}^{7} p_{128}^{(n)} \end{pmatrix}. \tag{61}$$

The factors in the last column are obtained by solving (54). Note that the components of the vector $\mathbf{p}^{TC}$ must sum to unity.

5.2. Predicting the fate of the semifinalists

Suppose that after the quarterfinal round, we wish to predict the probability of each of the four semifinalists becoming the tournament champion. We use the preceding recursion method, introducing the vectors $\mathbf{p}^{(0)}$, $\mathbf{p}^{(1)}$, and $\mathbf{p}^{(2)}$ of probabilities of winning the quarterfinal, semifinal, and final round

$$\mathbf{p}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{p}^{(n)} = \begin{pmatrix} p_1^{(n)} \\ p_2^{(n)} \\ p_3^{(n)} \\ p_4^{(n)} \end{pmatrix}, \quad (n = 1, 2). \tag{62}$$

The matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are given by

$$\mathbf{P}_1 = \begin{bmatrix} 0 & P_{12} & 0 & 0 \\ P_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{34} \\ 0 & 0 & P_{43} & 0 \end{bmatrix}, \tag{63}$$

$$\mathbf{P}_2 = \begin{bmatrix} 0 & 0 & P_{13} & P_{14} \\ 0 & 0 & P_{23} & P_{24} \\ P_{31} & P_{32} & 0 & 0 \\ P_{41} & P_{42} & 0 & 0 \end{bmatrix}. \tag{64}$$

The probability that player i wins a semifinal match is the ith component of

$$\mathbf{p}^{(1)} = \mathbf{P}_1\mathbf{p}^{(0)} = \begin{pmatrix} P_{12} \\ P_{21} \\ P_{34} \\ P_{43} \end{pmatrix}. \tag{65}$$

The probability that player i wins the final match if he or she plays in it is the ith component of

$$\mathbf{p}^{(2)} = \mathbf{P}_2\mathbf{p}^{(1)} = \begin{pmatrix} P_{13}P_{34} + P_{14}P_{43} \\ P_{23}P_{34} + P_{24}P_{43} \\ P_{31}P_{12} + P_{32}P_{21} \\ P_{41}P_{12} + P_{42}P_{21} \end{pmatrix}. \tag{66}$$

The vector of probabilities that each semifinalist wins the tournament is obtained by using (65) and (66) in (61):

$$\mathbf{p}^{TC} = \begin{pmatrix} P_{12}(P_{13}P_{34} + P_{14}P_{43}) \\ P_{21}(P_{23}P_{34} + P_{24}P_{43}) \\ P_{34}(P_{31}P_{12} + P_{32}P_{21}) \\ P_{43}(P_{41}P_{12} + P_{42}P_{21}) \end{pmatrix}. \tag{67}$$

6. 2002 U.S. Open and Wimbledon data

We now use the results of the 2002 U.S. Open and 2002 Wimbledon Singles events to show how the previous method can be applied to predict the tournament champion after the quarterfinal round ($n = 5$), based on the accumulated data through this round. Let $\alpha_i(n)$ be the total number of points won on serve by player i in round n, and let $\beta_i(n)$ be the total number of points served by player i in round n. Then, the empirical probability of player i winning a point on serve in round n is $\alpha_i(n)/\beta_i(n)$. The corresponding probability of winning a rally on serve in rounds 1–n is

$$p_i^R(n) = \sum_{j=1}^{n} \alpha_i(j) \Bigg/ \sum_{j=1}^{n} \beta_i(j). \tag{68}$$

We use this with $n = 5$ in (5) for each player in the semifinals and then compute their empirical probabilities of winning a match against any of the other remaining players. This allows us to compute the entries of the matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ in (63), (64), and arrive at values for $\mathbf{p}^{TC}$ in round $n = 6$ for each of the four semifinalists. To calculate $\mathbf{p}^{TC}$ for the two finalists after the semifinal round match, we repeat the same steps for the two finalists, using (68) with $n = 6$. The same method of calculating $\mathbf{p}^{TC}$ could be applied after round $n = 1$, and after each subsequent round as the tournament progresses to make running projections regarding tournament outcomes. Other forecasting methods which allow point by point updates as the match unfolds are described in [19].

6.1. Women's Tennis Association (WTA) data

Figure 7 shows the 2002 U.S. Open Women's Singles Draw from the semifinal round. Under each player, we show the value of $p_i^R(5)$, $p_i^R(6)$, and $p_i^R(7)$. Next to each player's name is their empirical probability of winning the upcoming match, P_{ij}, as well as their empirical probability of becoming the tournament champion, p_i^{TC}. After the quarterfinal round matches, L. Davenport would have been the slight favorite to win the tournament ($p_2^{TC} = 0.3599$), followed by V. Williams ($p_4^{TC} = 0.3047$), S. Williams ($p_1^{TC} = 0.2872$) and A. Mauresmo ($p_3^{TC} = 0.0482$), while after the semifinal round

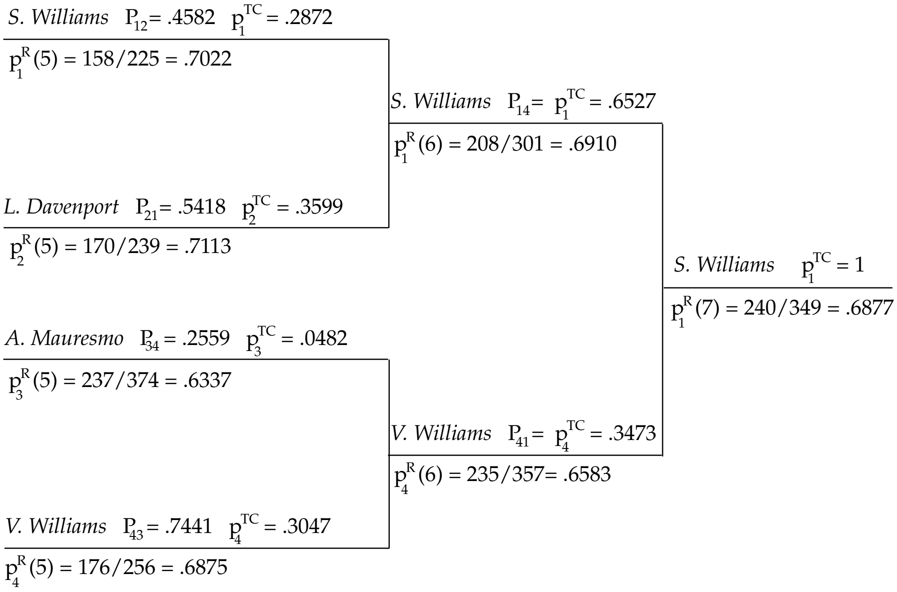
S. Williams P_{12}= .4582 p_1^{TC} = .2872
$p_1^R(5)$ = 158/225 = .7022
L. Davenport P_{21}= .5418 p_2^{TC} = .3599
$p_2^R(5)$ = 170/239 = .7113
A. Mauresmo P_{34} = .2559 p_3^{TC} = .0482
$p_3^R(5)$ = 237/374 = .6337
V. Williams P_{43} = .7441 p_4^{TC} = .3047
$p_4^R(5)$ = 176/256 = .6875
S. Williams P_{14}= p_1^{TC} = .6527
$p_1^R(6)$ = 208/301 = .6910
V. Williams P_{41}= p_4^{TC} = .3473
$p_4^R(6)$ = 235/357= .6583
S. Williams p_1^{TC} = 1
$p_1^R(7)$ = 240/349 = .6877

Figure 7. The probability P_{ij} of each of the four semifinalists in the 2002 U.S. Open Women's Singles tournament winning her match, and her probability p_i^{TC} of becoming the tournament champion.

matches, S. Williams ($p_1^{TC} = 0.6527$) was the favorite and ultimately won the tournament. Figure 8 shows the 2002 Wimbledon Women's Singles Draw from the semifinal round. Here, V. Williams ($p_1^{TC} = 0.4784$) was the favorite to win the tournament after the quarterfinal round match, followed by S. Williams ($p_4^{TC} = 0.3834$), A. Mauresmo ($p_3^{TC} = 0.1233$), and J. Henin ($p_2^{TC} = 0.0150$), while S. Williams ($p_4^{TC} = 0.5866$) was the favorite after the semifinal round match and ultimately won the tournament.

6.2. *Association of Tennis Professionals (ATP) data*

Figure 9 shows the 2002 U.S. Open Men's Singles Draw. After the quarterfinal round matches, P. Sampras was the heavy favorite to win the tournament ($p_1^{TC} = 0.6747$), followed by L. Hewitt ($p_4^{TC} = 0.1457$), A. Agassi ($p_3^{TC} = 0.0945$), and S. Schalken ($p_2^{TC} = 0.0851$). Sampras' chances of winning the tournament increased after his semifinal round match ($p_1^{TC} = 0.8856$) and he ultimately won the tournament. Figure 10 shows the results from the 2002 Wimbledon Men's Singles event. After their quarterfinal round matches, X. Malisse ($p_3^{TC} = 0.4573$) was favored to win the tournament, followed by L. Hewitt ($p_1^{TC} = 0.3364$), T. Henman ($p_2^{TC} = 0.1815$), and D. Nalbandian ($p_4^{TC} = 0.0247$). After the semifinal round matches, it was L. Hewitt, the ultimate tournament champion, who was the heavy favorite ($p_1^{TC} = 0.8698$).

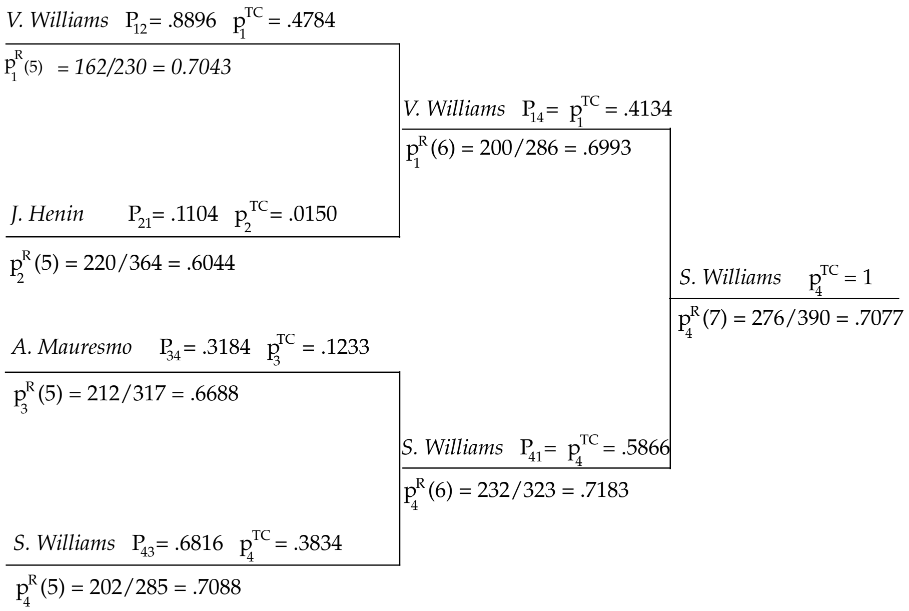
V. Williams P_{12}= .8896 p_1^{TC} = .4784
$p_1^R(5)$ = *162/230 = 0.7043*
J. Henin P_{21}= .1104 p_2^{TC}= .0150
$p_2^R(5)$ = 220/364 = .6044
A. Mauresmo P_{34}= .3184 p_3^{TC} = .1233
$p_3^R(5)$ = 212/317 = .6688
S. Williams P_{43}= .6816 p_4^{TC} = .3834
$p_4^R(5)$ = 202/285 = .7088
V. Williams P_{14}= p_1^{TC} = .4134
$p_1^R(6)$ = 200/286 = .6993
S. Williams P_{41}= p_4^{TC} = .5866
$p_4^R(6)$ = 232/323 = .7183
S. Williams p_4^{TC} = 1
$p_4^R(7)$ = 276/390 = .7077

Figure 8. The probability P_{ij} of each of the four semifinalists in the 2002 Wimbledon Women's Singles tournament winning her match, and her probability p_i^{TC} of becoming the tournament champion.

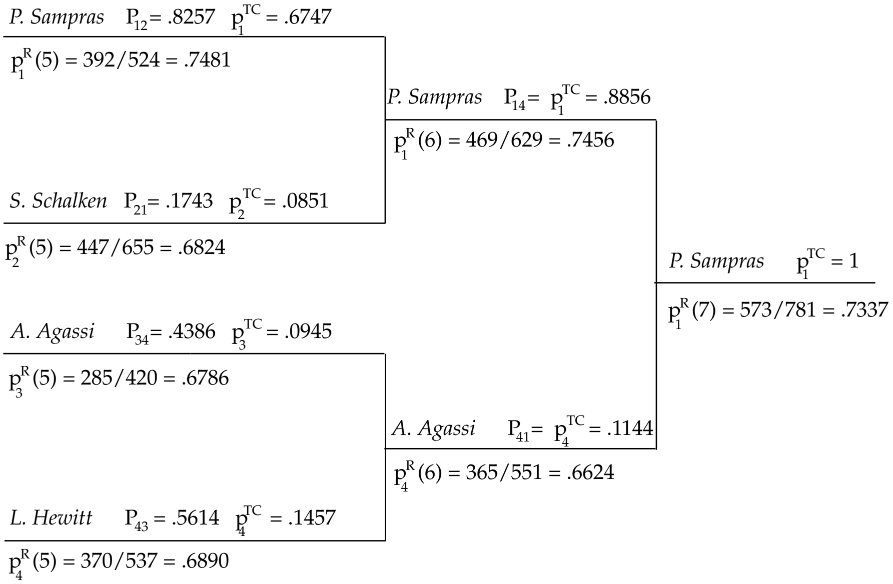
P. Sampras P_{12}= .8257 p_1^{TC} = .6747
$p_1^R(5)$ = 392/524 = .7481
S. Schalken P_{21}= .1743 p_2^{TC}= .0851
$p_2^R(5)$ = 447/655 = .6824
A. Agassi P_{34}= .4386 p_3^{TC} = .0945
$p_3^R(5)$ = 285/420 = .6786
L. Hewitt P_{43} = .5614 p_4^{TC} = .1457
$p_4^R(5)$ = 370/537 = .6890
P. Sampras P_{14}= p_1^{TC} = .8856
$p_1^R(6)$ = 469/629 = .7456
A. Agassi P_{41}= p_4^{TC} = .1144
$p_4^R(6)$ = 365/551 = .6624
P. Sampras p_1^{TC} = 1
$p_1^R(7)$ = 573/781 = .7337

Figure 9. The probability P_{ij} of each of the four semifinalists in the 2002 U.S. Open Men's Singles tournament winning his match, and his probability p_i^{TC} of becoming the tournament champion.

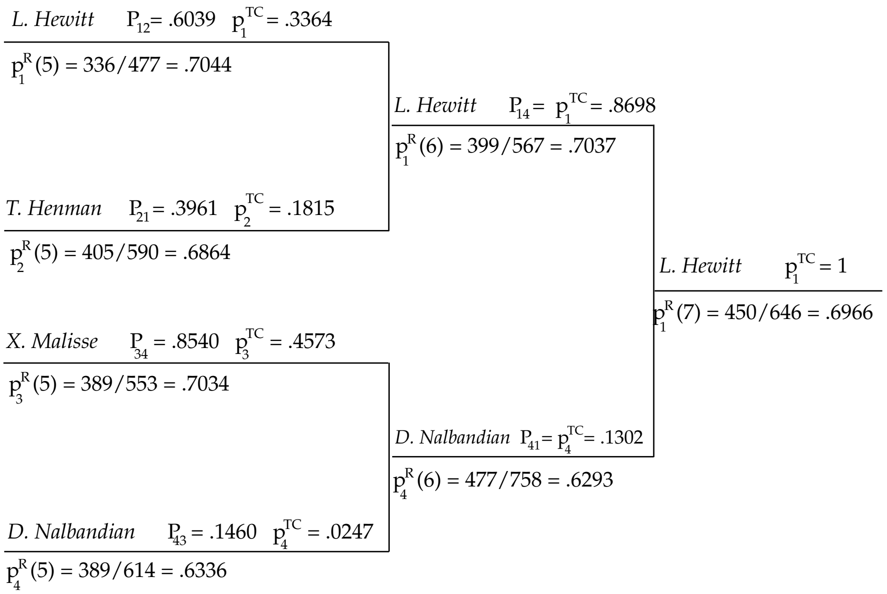
L. Hewitt P_{12}= .6039 p_1^{TC} = .3364
$p_1^R(5)$ = 336/477 = .7044
T. Henman P_{21} = .3961 p_2^{TC} = .1815
$p_2^R(5)$ = 405/590 = .6864
X. Malisse P_{34} = .8540 p_3^{TC} = .4573
$p_3^R(5)$ = 389/553 = .7034
D. Nalbandian P_{43} = .1460 p_4^{TC} = .0247
$p_4^R(5)$ = 389/614 = .6336
L. Hewitt P_{14} = p_1^{TC} = .8698
$p_1^R(6)$ = 399/567 = .7037
D. Nalbandian P_{41} = p_4^{TC} = .1302
$p_4^R(6)$ = 477/758 = .6293
L. Hewitt p_1^{TC} = 1
$p_1^R(7)$ = 450/646 = .6966

Figure 10. The probability P_{ij} of each of the four semifinalists in the 2002 Wimbledon Men's Singles tournament winning his match, and his probability p_i^{TC} of becoming the tournament champion.

7. Capturing non-iid effects

There are several papers documenting effects that cannot be captured with the assumption that points are independent and identically distributed. For example, Magnus and Klassen [20] analyze 90,000 points played at Wimbledon, and find evidence of a "first game effect," i.e., that the first game of a match is the hardest one to break. This indicates that it may be desirable to allow p_A^G and p_B^G to vary from game to game and perhaps depend on the specific pair of players who are competing. Jackson and Mosurski [21] give compelling evidence which indicates that points may not be independent. This includes what is commonly called the "hot-hand" phenomenon in which winning a previous point, game, or set, increases ones chances of winning the next, and the opposite of this, called the "back-to-the-wall" effect in which playing from behind can sometimes be a psychological advantage. From the analysis of Klassen and Magnus [9], one can assume that although these effects may be small when analyzing large heterogeneous data sets, they may be more important when analyzing specific head-to-head match-ups between two players, as, for example, the famous McEnroe–Borg series of matches [21] in which a "back-to-the-wall" phenomenon seems to be present.

A more refined analysis than the one described in this paper could incorporate these and other higher-order effects by allowing p_A^R and p_B^R to vary from point to point as the match unfolds, depending on the points "importance" [12] or by taking into consideration more detailed player characteristics such as rallying ability or strength of return of serve. For example, we could define the probability that player A wins a point on serve as

$$\hat{p}_A^R = p_A^R + \delta p_{AB}^R(i, j), \quad \left(0 \le \hat{p}_A^R \le 1\right) \tag{69}$$

where p_A^R is constant throughout the match, $p_{AB}^R(i,\ j)$ represents player A's probability of winning a point on serve against player B, when the score is i points for A and j points for B, and $\delta \ll 1$ is a small parameter reflecting the fact that, in most cases, the deviation from iid is small. The goal then would be to calculate the corresponding formulas for game, set, and match for each player, i.e., $\hat{p}_A^G, \hat{p}_A^S, \hat{p}_A^M$, and $\hat{p}_B^G, \hat{p}_B^S, \hat{p}_B^M$. The "leading-order" theory ($\delta = 0$) is the one described in this paper based on the iid assumption, while "higher-order" corrections could be treated perturbatively.

Acknowledgments

This work is supported by the National Science Foundation grants NSF-DMS 9800797 and NSF-DMS 0203581. Useful comments and observations by J. D'Angelo and G.H. Pollard on an early draft of the manuscript are gratefully acknowledged. The first author also thanks Andres Figueroa for skillfully performing Matlab calculations on the models developed in this manuscript as part of a summer undergraduate research project.

Appendix

The solution of (7)–(10) is

$$p_A^S(6,0) = \left(p_A^G q_B^G\right)^3 \tag{A.1}$$

$$p_A^S(6,1) = 3\left(p_A^G\right)^3 q_A^G \left(q_B^G\right)^3 + 3\left(p_A^G\right)^4 p_B^G \left(q_B^G\right)^2 \tag{A.2}$$

$$\begin{aligned} p_A^S(6,2) = {} & 12\left(p_A^G\right)^3 q_A^G p_B^G \left(q_B^G\right)^3 + 6\left(p_A^G\right)^2 \left(q_A^G\right)^2 \left(q_B^G\right)^4 \\ & + 3\left(p_A^G\right)^4 \left(p_B^G\right)^2 \left(q_B^G\right)^2 \end{aligned} \tag{A.3}$$

$$\begin{aligned} p_A^S(6,3) = {} & 24\left(p_A^G\right)^3 \left(q_A^G\right)^2 p_B^G \left(q_B^G\right)^3 + 24\left(p_A^G\right)^4 q_A^G \left(p_B^G\right)^2 \left(q_B^G\right)^2 \\ & + 4\left(p_A^G\right)^2 \left(q_A^G\right)^3 \left(q_B^G\right)^4 + 4\left(p_A^G\right)^5 \left(p_B^G\right)^3 q_B^G \end{aligned} \tag{A.4}$$

$$\begin{aligned} p_A^S(6,4) &= 60\big(p_A^G\big)^3\big(q_A^G\big)^2\big(p_B^G\big)^2\big(q_B^G\big)^3 + 40\big(p_A^G\big)^2\big(q_A^G\big)^3 p_B^G\big(q_B^G\big)^4 \\ &\quad + 20\big(p_A^G\big)^4 q_A^G\big(p_B^G\big)^3\big(q_B^G\big)^2 + 5p_A^G\big(q_A^G\big)^4\big(q_B^G\big)^5 \\ &\quad + \big(p_A^G\big)^5\big(p_B^G\big)^4 q_B^G \end{aligned} \tag{A.5}$$

$$\begin{aligned} p_A^S(7,5) &= 100\big(p_A^G\big)^3\big(q_A^G\big)^3\big(p_B^G\big)^2\big(q_B^G\big)^4 + 100\big(p_A^G\big)^4\big(q_A^G\big)^2\big(p_B^G\big)^3\big(q_B^G\big)^3 \\ &\quad + 25\big(p_A^G\big)^2\big(q_A^G\big)^4 p_B^G\big(q_B^G\big)^5 + 25\big(p_A^G\big)^5 q_A^G\big(p_B^G\big)^4\big(q_B^G\big)^2 \\ &\quad + p_A^G\big(q_A^G\big)^5\big(q_B^G\big)^6 + \big(p_A^G\big)^6\big(p_B^G\big)^5 q_B^G. \end{aligned} \tag{A.6}$$

To obtain $p_A^S(i, j)$ from $p_A^S(j, i)$, we interchange $p_A^G \leftrightarrow q_A^G$ and $p_B^G \leftrightarrow q_B^G$ in (A.1)–(A.6). Finally, $p_A^S(6, 6)$ in (6) is given by

$$p_A^S(6,6) = 1 - \left[\sum_{i=0}^{4}\big(p_A^S(i,6) + p_A^S(6,i)\big) + p_A^S(7,5) + p_A^S(5,7)\right]. \tag{A.7}$$

The solution of (14)–(16) yields:

$$p_A^T(7,0) = \big(p_A^R\big)^3\big(q_B^R\big)^4 \tag{A.8}$$

$$p_A^T(7,1) = 3\big(p_A^R\big)^3 q_A^R\big(q_B^R\big)^4 + 4\big(p_A^R\big)^4 p_B^R\big(q_B^R\big)^3 \tag{A.9}$$

$$\begin{aligned} p_A^T(7,2) &= 16\big(p_A^R\big)^4 q_A^R p_B^R\big(q_B^R\big)^3 + 6\big(p_A^R\big)^5\big(p_B^R\big)^2\big(q_B^R\big)^2 \\ &\quad + 6\big(p_A^R\big)^3\big(q_A^R\big)^2\big(q_B^R\big)^4 \end{aligned} \tag{A.10}$$

$$\begin{aligned} p_A^T(7,3) &= 40\big(p_A^R\big)^3\big(q_A^R\big)^2 p_B^R\big(q_B^R\big)^4 + 10\big(p_A^R\big)^2\big(q_A^R\big)^3\big(q_B^R\big)^5 \\ &\quad + 4\big(p_A^R\big)^5\big(p_B^R\big)^3\big(q_B^R\big)^2 + 30\big(p_A^R\big)^4 q_A^R\big(p_B^R\big)^2\big(q_B^R\big)^3 \end{aligned} \tag{A.11}$$

$$\begin{aligned} p_A^T(7,4) &= 50\big(p_A^R\big)^4 q_A^R\big(p_B^R\big)^3\big(q_B^R\big)^3 + 5\big(p_A^R\big)^5\big(p_B^R\big)^4\big(q_B^R\big)^2 \\ &\quad + 50\big(p_A^R\big)^2\big(q_A^R\big)^3 p_B^R\big(q_B^R\big)^5 + 5p_A^R\big(q_A^R\big)^4\big(q_B^R\big)^6 \\ &\quad + 100\big(p_A^R\big)^3\big(q_A^R\big)^2\big(p_B^R\big)^2\big(q_B^R\big)^4 \end{aligned} \tag{A.12}$$

$$\begin{aligned} p_A^T(7,5) &= 30\big(p_A^R\big)^2\big(q_A^R\big)^4 p_B^R\big(q_B^R\big)^5 + p_A^R\big(q_A^R\big)^5\big(q_B^R\big)^6 \\ &\quad + 200\big(p_A^R\big)^4\big(q_A^R\big)^2\big(p_B^R\big)^3\big(q_B^R\big)^3 + 75\big(p_A^R\big)^5 q_A^R\big(p_B^R\big)^4\big(q_B^R\big)^2 \\ &\quad + 150\big(p_A^R\big)^3\big(q_A^R\big)^3\big(p_B^R\big)^2\big(q_B^R\big)^4 + 6\big(p_A^R\big)^6\big(p_B^R\big)^5 q_B^R. \end{aligned} \tag{A.13}$$

To obtain $p_A^T(j, i)$ from $p_A^T(i, j)$, we interchange $p_A^R \leftrightarrow q_A^R$ and $p_B^R \leftrightarrow q_B^R$ in (A.9)–(A.13). Finally, $p_A^T(6, 6)$ in (13) is given by

$$p_A^T(6,6) = 1 - \left[\sum_{i=0}^{5}\big(p_A^T(i,7) + p_A^T(7,i)\big)\right]. \tag{A.14}$$

The nonzero coefficients $a_{ij}^{S}(n) = b_{ij}^{S}(n)$ are given by

$$\begin{aligned}
&a_{20}^{S}(1) = 6, && a_{21}^{S}(1) = -24, && a_{22}^{S}(1) = 36, && a_{23}^{S}(1) = -24, && a_{24}^{S}(1) = 6,\\
&a_{30}^{S}(1) = -9, && a_{31}^{S}(1) = 51, && a_{32}^{S}(1) = -99, && a_{33}^{S}(1) = 81, && a_{34}^{S}(1) = -24,\\
&a_{40}^{S}(1) = 3, && a_{41}^{S}(1) = -24, && a_{42}^{S}(1) = 60, && a_{43}^{S}(1) = -60, && a_{44}^{S}(1) = 21.
\end{aligned} \tag{A.15}$$

$$\begin{aligned}
&a_{10}^{S}(2) = 5, && a_{11}^{S}(2) = -25, && a_{12}^{S}(2) = 50, && a_{13}^{S}(2) = -50,\\
&a_{14}^{S}(2) = 25, && a_{15}^{S}(2) = -5,\\
&a_{20}^{S}(2) = -16, && a_{21}^{S}(2) = 124, && a_{22}^{S}(2) = -336, && a_{23}^{S}(2) = 424,\\
&a_{24}^{S}(2) = -256, && a_{25}^{S}(2) = 60\\
&a_{30}^{S}(2) = 18, && a_{31}^{S}(2) = -198, && a_{32}^{S}(2) = 696, && a_{33}^{S}(2) = -1080,\\
&a_{34}^{S}(2) = 774, && a_{35}^{S}(2) = -210\\
&a_{40}^{S}(2) = -8, && a_{41}^{S}(2) = 124, && a_{42}^{S}(2) = -560, && a_{43}^{S}(2) = 1060,\\
&a_{44}^{S}(2) = -896, && a_{45}^{S}(2) = 280\\
&a_{50}^{S}(2) = 1, && a_{51}^{S}(2) = -25, && a_{52}^{S}(2) = 150, && a_{53}^{S}(2) = -350,\\
&a_{54}^{S}(2) = 350, && a_{55}^{S}(2) = -126.
\end{aligned} \tag{A.16}$$

The nonzero coefficients $a_{ij}^{T}(n) = b_{ij}^{T}(n)$ are given by

$$\begin{aligned}
&a_{30}^{T}(1) = 4, && a_{31}^{T}(1) = -16, && a_{32}^{T}(1) = 24, && a_{33}^{T}(1) = -16, && a_{34}^{T}(1) = 4,\\
&a_{40}^{T}(1) = -3, && a_{41}^{T}(1) = 16, && a_{42}^{T}(1) = -30, && a_{43}^{T}(1) = 24, && a_{44}^{T}(1) = -7.
\end{aligned} \tag{A.17}$$

$$\begin{aligned}
&a_{20}^{T}(2) = 10, && a_{21}^{T}(2) = -50, && a_{22}^{T}(2) = 100, && a_{23}^{T}(2) = -100,\\
&a_{24}^{T}(2) = 50, && a_{25}^{T}(2) = -10\\
&a_{30}^{T}(2) = -24, && a_{31}^{T}(2) = 166, && a_{32}^{T}(2) = -424, && a_{33}^{T}(2) = 516,\\
&a_{34}^{T}(2) = -304, && a_{35}^{T}(2) = 70\\
&a_{40}^{T}(2) = 18, && a_{41}^{T}(2) = -166, && a_{42}^{T}(2) = 530, && a_{43}^{T}(2) = -774,\\
&a_{44}^{T}(2) = 532, && a_{45}^{T}(2) = -140\\
&a_{50}^{T}(2) = -4, && a_{51}^{T}(2) = 50, && a_{52}^{T}(2) = -200, && a_{53}^{T}(2) = 350,\\
&a_{54}^{T}(2) = -280, && a_{55}^{T}(2) = 84.
\end{aligned} \tag{A.18}$$

$$
\begin{array}{llll}
a_{10}^{T}(3) = 6, & a_{11}^{T}(3) = -36, & a_{12}^{T}(3) = 90, & a_{13}^{T}(3) = -120, \\
a_{14}^{T}(3) = 90, & a_{15}^{T}(3) = -36, & a_{16}^{T}(3) = 6 & \\
a_{20}^{T}(3) = -25, & a_{21}^{T}(3) = 230, & a_{22}^{T}(3) = -775, & a_{23}^{T}(3) = 1300, \\
a_{24}^{T}(3) = -1175, & a_{25}^{T}(3) = 550, & a_{26}^{T}(3) = -105 & \\
a_{30}^{T}(3) = 40, & a_{31}^{T}(3) = -510, & a_{32}^{T}(3) = 2200, & a_{33}^{T}(3) = -4500, \\
a_{34}^{T}(3) = 4800, & a_{35}^{T}(3) = -2590, & a_{36}^{T}(3) = 560 & \\
a_{40}^{T}(3) = -30, & a_{41}^{T}(3) = 510, & a_{42}^{T}(3) = -2750, & a_{43}^{T}(3) = 6750, \\
a_{44}^{T}(3) = -8400, & a_{45}^{T}(3) = 5180, & a_{46}^{T}(3) = -1260 & \\
a_{50}^{T}(3) = 10, & a_{51}^{T}(3) = -230, & a_{52}^{T}(3) = 1550, & a_{53}^{T}(3) = -4550, \\
a_{54}^{T}(3) = 6580, & a_{55}^{T}(3) = -5620, & a_{56}^{T}(3) = 1260, & \\
a_{60}^{T}(3) = -1, & a_{61}^{T}(3) = 36, & a_{62}^{T}(3) = -315, & a_{63}^{T}(3) = 1120, \\
a_{64}^{T}(3) = -1890, & a_{65}^{T}(3) = 1512, & a_{66}^{T}(3) = -462. &
\end{array}
\tag{A.19}
$$

References

1. S. R. Clarke and D. S. Dyte, Using official tennis ratings to estimate tournament chances, preprint, 2002.
2. R. T. Stefani, Survey of the major world sports rating systems, *J. Appl. Stat.* 24(6):635–646 (1997).
3. J. B. Keller, Probability of a shutout in racquetball, *SIAM Rev.* 26:267–268 (1984).
4. J. Renick, Optimal strategies at decision points in singles squash, *Res. Quart. Exercise Sport* 47:562–568 (1976).
5. J. B. Keller, Tie point strategies in badminton, preprint, 2003.
6. B. P. Hsi and D. M. Burych, Games of two players, *Appl. Stat.: J. R. Stat. Soc. C* 22(1):86–92 (1971).
7. W. H. Carter and S. L. Crews, An analysis of the game of tennis, *Am. Stat.* 28(4):130–134 (1974).
8. G. H. Pollard, An analysis of classical and tie-breaker tennis, *Austr. J. Stat.* 25:496–505 (1983).
9. F. J. G. M. Klaassen and J. R. Magnus, Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model, *J. Am. Stat. Assoc.* 96(454):500–509 (2001).
10. S. L. George, Optimal strategy in tennis: A simple probabilistic model, *Appl. Stat.: J. R. Stat. Soc. C* 22(1):97–104 (1973).
11. R. E. Miles, Symmetric sequential analysis: The efficiencies of sports scoring systems (with particular reference to those of tennis), *J. R. Stat. Soc. B* 46(1):93–108 (1984).

12. C. Morris, The most important points in tennis, in *Optimal Strategies in Sport* (S. P. Ladany and R. E. Machol, Eds.), pp. 131–140, Amsterdam; North-Holland, 1977.

13. J. R. Magnus and F. J. G. M. Klaassen, The effect of new balls in tennis: Four years at Wimbledon, *The Statistician* 48:239–246 (1999).

14. F. J. G. M. Klaassen and J. R. Magnus, How to reduce the service dominance in tennis? Empirical results from four years at Wimbledon, preprint, 2003.

15. J. R. Magnus and F. J. G. M. Klaassen, The final set in a tennis match: Four years at Wimbledon, *J. Appl. Stat.* 26(4):461–468 (1999).

16. J. G. Kingston, Comparison of scoring systems in two-sided competitions, *J. Comb. Theory A* 20:357–362 (1976).

17. C. L. Anderson, Note on the advantage of first serve, *J. Comb. Theory A* 23:363 (1977).

18. P. K. Newton and G. H. Pollard, Service neutral scoring strategies for tennis, in *Proceedings of the Seventh Autralasian Conference on Mathematics and Computers in Sport*, 2004.

19. F. J. G. M. Klaassen and J. R. Magnus, Forecasting in tennis, preprint, 2003.

20. J. R. Magnus and F. J. G. M. Klaassen, On the advantage of serving first in a tennis set: Four years at Wimbledon, *The Statistician* 48:247–256 (1999).

21. D. Jackson and K. Mosurski, Heavy defeats in tennis: Psychological momentum or random effects, *Chance* 10:27–34 (1997).

University of Southern California
Stanford University

(Received July 21, 2004)