A STOCHASTIC MARKOV CHAIN APPROACH FOR TENNIS: MONTE CARLO SIMULATION AND MODELING

by

Kamran Aslam

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(AEROSPACE ENGINEERING)

May 2012

Dedication

This work is dedicated to my father, Javed Aslam. If he wasn't so persistent this might not have happened.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation describes the computational formulation of probability density functions (*pdf*s) that facilitate head-to-head match simulations in tennis along with ranking systems developed from their use. A background on the statistical method used to develop the *pdf*s, the Monte Carlo method, and the resulting rankings are included along with a discussion on ranking methods currently being used both in professional sports and in other applications. Using an analytical theory developed by Newton and Keller in [34] that defines a tennis player's probability of winning a game, set, match and single elimination tournament, a computational simulation has been developed in Matlab that allows further modeling not previously possible with the analytical theory alone. Such experimentation consists of the exploration of non-iid effects, considers the concept the varying importance of points in a match and allows an unlimited number of matches to be simulated between unlikely opponents. The results of these studies have provided *pdf*s that accurately model an individual tennis player's ability along with a realistic, fair and mathematically sound platform for ranking them.

Chapter 1.

Introduction

During a Grand Slam season in tennis, consisting of four high profile tournaments per year, multi-million dollar purses are awarded to the champions and top finishers at each event. Therefore, it's immediately apparent that the ranking and seed of a particular player for each tournament is of great importance not only to his reputation but to his money earning potential. To date, no collective ranking system exists for the set of Association of Tennis Professionals (ATP) Grand Slam tournaments causing skepticism regarding the true merit of a particular player's seed in a given event. Other major sports have attempted to address similar issues with respect to ranking, most notably the Bowl Championship Series in NCAA college football.

This work addresses common problems encountered by such ranking systems and suggests an alternative to established, well documented systems used in sports and elsewhere. Chapter §2 contains an overview of the Monte Carlo method, its strengths and suitable applications along with fundamental constraints. In Chapter §3, a detailed discussion is presented on the analytical approach that describes a tennis player's probability of winning a game, set, match and even a multi-player tournament based on his probability of winning a point on serve. The investigation naturally leads to the development of probability density functions that are used to mathematically

model a particular player's performance. This is documented in detail in Chapter §4. From these *pdf*s, an infinite number of experimental matches can be played between any two players. This has immediate advantages. First, any given player's computational performance can be adjusted dynamically to adequately mimic their observed real life performance. Additionally, the outcome of notional future tennis matchups can be realistically predicted based on individual performance, even when considering players with little or no previous match history. Of particular importance are the additional statistics that can be gathered from millions of these simulated matches that aren't available for compilation based on limited availability of actual results. Organizing and post-processing the outcome of these *pdf* driven matchups provide a natural foundation for a number of different tennis ranking systems.

To fully appreciate ranking theory and the state of the art, a comprehensive background on rankings for sports and other applications is provided in Chapter §5. Utilizing the Monte Carlo based *pdf* player model, two particularly interesting tennis ranking systems are presented in Chapter §6. These systems have been built by simulating millions of games, sets and matches under a variety of conditions. Ultimately, a probabilistic based ranking system such as one of these could provide unity to the ad-hoc systems currently in place in professional tennis. Therefore, it's possible that a new standard of equity could be realized in the level of competition and in the money earning potential for all professional players. On an even greater level, the concept of developing and using *pdf*s to model individual data sets could have a profound impact on how individuals receive and process (or rank) information in realms outside of sports such as stock market/financial analysis, database browsing/searching, weather/geological surveying, computer gaming and much more.

Chapter 2.

Monte Carlo Methods and Probability Density Functions

The Monte Carlo method can be defined as a method whereby any technique of statistical sampling is used to obtain an approximate solution to some quantitative problem. The problem may be inherently probabilitic such as counting the outcomes of repeatedly flipping a coin. However, this isn't a prerequisite for its successful use. For example, the solution of a parabolic differential equation has been modeled as early as 1899 with what is now considered the Monte Carlo method by Lord Rayleigh [41]. In fact, the use of statistical sampling to solve numerical problems isn't a new concept. In the mid 1700's a French mathematician was able to estimate the value of π by using a needle and a lined grid alone [42]. This experiment is carried out by repeatedly dropping a needle onto the grid and recording the total number of drops along with the total number of times the needle crosses one of the lines. Using elementary probability theory, it can be shown that the probability of this event is simply $\frac{2}{\pi}$. Therefore after many trials the value of π is easily estimated from the ratio of needle crosses to total needle drops.

The Monte Carlo method can be implemented for stochastic problems using physical processes to generate random samples. Examples of such physical processes would be card draws, dice tosses and lottery power ball sequences. Further, success or failure of a given event in a

particular experiment can be simulated via a random number generator. This greatly simplifies the process of sampling success or failure of a random process with a given probability and transforms the Monte Carlo method from a bulky and time consuming affair to a hands off, effective computational technique.

2.1 Computational Monte Carlo

To say that the advent of computers and the digital revolution have made a huge impact on the efficacy of the Monte Carlo method would be an understatement. With new tools, the task of generating random number sequences can be automated and is no longer a burdening physical task. In 1946, Stanislaw Ulam and John von Neumann developed a set of statistical algorithms to effectively simulate neutron diffusion as part of the the Manhattan Project [47]. Using the first electronic computer, ENIAC, the team of scientists at Los Alamos Scientific Laboratory quickly concluded that the Monte Carlo process produced effective results. However, a problem in developing realistic random numbers for the simulations was encountered due most directly to the computing power, or lack thereof, of the ENIAC system. Originally, the so called von Neumann "middle square digits" technique for random number sampling was used [29]. In this technique an arbitrary n-digit integer is squared creating a $2n$-digit product. A new sample integer of n-digits is extracted from this product and the cycle continues. While the shortfalls of this number generating sequence have been extensively studied, the limits of the computational power available to the scientists was far more restrictive. Though a number generating sequence can provide a robust set of random values, a computer with finite memory can only store values in finite precision as binary sequences of *0's* and *1's*. The results of those studies and the Manhanttan

Project itself are quite evident. They demonstrated on a world scale the power of the Monte Carlo method not just as a statistical curiosity but as a formal methodology.

2.2 Random Number Generation

A suitable random number generator is critical to both the accuracy and fidelity of a Monte Carlo based analysis. Thus, no discussion about the method would be complete without outlining the tools available and problems associated with random number generation [21]. An ideal generator provides numbers from the $Uniform[0,1]$ distribution that are uncorrelated, i.e. $cov(x_i, x_j) = 0$, and have a large (ideally infinite) period of repetition. Further, desirable characteristics include the ability to adjust the sequence using an initial or "seed" parameter(s), the ability to generate numbers very quicky and portability between various computer platforms [10]. Finally, a somewhat ambiguous criteria when evaluating a random number generator is that it satisfies all reasonable statistical tests of randomness. Note that any random number generator will fail some statistical test if the test is sensitive enough. Nonetheless, this standard allows the designer of a particular algorithm some flexibility as to the "level of randomness" for the particular generator. The uniform distribution is desirable for random number generation because the selection (or production) of any value within its range is equally likely. A random variable X is said to be $Uniform[a,b]$ if it has the distribution

$$F_X(x) = P(X \leq x) = \frac{x-a}{b-a} \tag{2.1}$$

Therefore the $Uniform[0,1]$ distribution for the random variable X is given by

$$F_X(x) = P(X \leq x) = x \tag{2.2}$$

which means that X can take any value in the range [0,1] with equal probability. Developing random numbers in this range is useful because from this distribution, random numbers from the uniform distribution can be obtained on any interval with ease.

Random numbers can be broken down into two main classes. The first, pseudo-random number generators, use deterministic processes intended to imitate purely random sequences. Quasi-random number generators make no pretense at being random, but they have important predefined statistical properties shared by random sequences [50]. For the purpose of most Monte Carlo based studies, a psuedo-random number generator is necessary because it provides a platform to develop "desirable" random number sequences that have the characteristics previously described. On the other hand, a quasi-random number generator incorporates logic to create sequences that are statistically "nice" for a given purpose. To create high fidelity simulations the most purely random sequence is necessary and hence psuedo-random number sequences are almost exclusively used in Monte Carlo applications.

Several families of pseudo-random number generators have evolved as a result of the emergence of Monte Carlo as a robust numerical method. A brief discussion is included to provide background on some of the more prominent ones and to characterize their strengths and weaknesses. This will ultimately provide confidence in their performance in Monte Carlo simulations.

2.2.1 Chaos Theory and the Logistic Map

The formal study of Chaos Theory was in its infancy in the 1940's when von Neumann and Ulam were just beginning their work on the ENIAC and the Manhattan Project. Nevertheless, some informal investigation had begun on the Logistic Map given by

$$x_{n+1} = r x_n (1 - x_n) \tag{2.3}$$

where r is called the *bifurcation parameter*. In particular, von Neumann had already suggested the use of the Logistic Map with bifurcation parameter $r = 4$ as an early random number generator [51]. While his thoughts were pioneering, a more general understanding beyond the oscillatory patterns of this map weren't noted until nearly a decade later. Over time, the bifurcation parameter $r_\infty \approx 3.569945672$ was obtained, providing a completely chaotic sequence from the Logistic Map (a full treatment is presented by Rasband in [40]). Therefore, in practice, this algorithm can provide a perfectly random sequence of numbers. A nice feature of this equation is that it will map a value in the interval $[0,1]$ to another perfectly random value in the same interval when using r_∞. However, when implemented computationally, these mapped values don't behave suitably random in finite precision as documented by Wagner in [48]. More recently it's been shown that use of a lattice of numbers can provide suitable randomness that can't be obtained from a single Logistic Map implementation alone. This added complexity comes at an added cost of longer computational run time and hence it's generally considered unsuitable for use in Monte Carlo based simulations. As time progresses and computational power becomes

even cheaper and readily available, further studies may be undertaken to determine the value of Logistic Map pseudo-random number generators for Monte Carlo applications.

2.2.2 Linear Congruential Generators

The linear congruential generator, or LCG, is one of the most basic algorithms for generating random number sequences. It was first proposed by Lehmer in 1949 [23] and is given by the following equation

$$x_{n+1} = (ax_n + b)\,\mathbf{mod}\,m \tag{2.4}$$

where a, b, and m are arbitrary constants chosen by the generator's designer. An LCG with these constants would then be denoted $LCG(m,a,b,x_0)$ for seed parameter x_0. A resulting number from this generator can be scaled to the $Uniform[0,1]$ distribution by $u_n = \frac{x_n}{m}$.

The LCG is a very fast algorithm but one that has serious drawbacks. Most importantly, the LCG is very sensitive to its initial parameters and can easily produce short, repeating number sequences. In other cases, a strong correlation can be seen between successive samples from this generator [10]. For example, consider $LCG(13,3,5,7)$. Table 2.1 shows the psuedo-random numbers generated from this algorithm.

n	x_n	$ax_n + b$
0	7	26
1	0	5
2	5	20
3	7	26

Table 2.1. Pseudo-random numbers from a fictional LCG

The resulting sequence of numbers demonstrates the major weakness of the LCG. Even though values 0 through 12 should appear arbitrarily in such a list, only 3 samples are needed to start a repeating chain of the sequence 7,0,5. While standalone LCGs can still be found on modern day computer systems, they're much more commonly found as part of multiple congruential generators (MCG). MCGs are effectively linear combinations of LCG based pseudo-random numbers.

2.2.3 Inversive Congruential Generators

Inversive congruential generators (ICGs) were first proposed by Eichenauer and Lehn in 1986 and are given by the equation

$$x_{n+1} = (a\bar{x}_n + b)\,\mathbf{mod}\,m \tag{2.5}$$

where $\bar{x} = x^{-1}$ such that $x\bar{x}\,\mathbf{mod}\,m = 1$ [13]. In mathematical terms, $\bar{x}$ is called the modular inverse of x. An explicit form of the ICG has also been presented by Niederreiter in [36]. For m prime, it can be shown that $\bar{x} = x^{m-2}\,\mathbf{mod}\,m$. As with the LCG, the ICG can be scaled to $Uniform[0,1]$ by $u_n = \frac{x_n}{m}$ and is denoted $ICG(m,a,b,x_0)$.

The major benefit of the ICG is that it eliminates the direct correlation between elements n and $n+1$ in the sequence. It's slightly slower in performance than the comparable LCG as it adds minor complexity with modular inversion, adding a factor of $O(log_2 m)$ to the cost of machine multiplication. However, modern computing power virtually negates this loss in large scale applications and the added performance can easily justify this cost. Another drawback, similar to LCGs, is the existence of a defined "mother-son" relationship between the parameters

a and m. Therefore, careful selection of these parameters is required for optimally tuned ICGs. Consider $ICG(13,3,5,7)$ as an analog to the sample LCG previously presented in 2.1. Table 2.2 shows the pseudo-random numbers computed from this generator.

n	x_n	$\bar{x}_n$	$a\bar{x}_n + b$
0	7	2	11
1	11	6	23
2	10	4	17
3	4	10	35
4	9	3	14
5	1	1	8
6	8	5	20
7	7	2	11

Table 2.2. Pseudo-random numbers from a fictional ICG

It's apparent that the ICG under consideration greatly outperforms the comparable LCG. While this isn't always the case, the ICG will generally outperform the LCG with appropriate selection of the initial parameters.

2.2.4 Lagged Fibonacci Generators

Another example of a common pseudo-random number generator, the Lagged Fibonacci Generator (LFG), is given by

$$x_{n+1} = (x_{n-i} + x_{n-k}) \,\mathbf{mod}\, m \tag{2.6}$$

and is named due to its similarity to the Fibonacci sequence $1,1,2,3,5,8...$ from $x_{n+1} = x_n + x_{n-1}$. The parameters i,k,m are chosen arbitrarily with i and k as the "lags". The larger lag

value sets the requirement for the number of seed values that must be predefined in the sequence. The LFG, denoted $LFG(i,k,m)$, is scaled to $Uniform[0,1]$ by $u_n = \frac{x_n}{m}$.

A major advantage of the LFG is its larger maximum period of repetition $(2^k - 1) * 2^{M-1}$ versus the conventional LCG or ICG with same modulus parameter $m \equiv 2^M$ [28]. Those generators have corresponding maximum periods of simply $m \equiv 2^M$ for appropriate choices of the parameters a and b [22]. While the period of repetition is larger for the LFG it does come at some cost. The additional values that must be stored due to the lag parameters result in a larger computational memory burden with the LFG making it slower than the LCG and ICG. In particular, k words of memory are required here as opposed to just one word as before. Also, the use of the LFG introduces an added complexity in terms of initializing the lag parameters. While several studies have been done to address this problem, fixed values are selected here to demonstrate a psuedo-random number sequence using $LFG(13,3,5)$.

The values in Table 2.3 significantly outperform both previous sample pseudo-random number sequences listed in Table 2.1 and Table 2.2, outlining the utility of the algorithm. Without preference to the selection of lag parameters, the LFG is considerably better than the LCG and ICG. It's for this reason that the LFG is being more frequently used in pseudo-random number generation packages found in commercial, off the shelf software products.

2.3 Monte Carlo as a Numerical Method

Monte Carlo methods can be loosely defined as statistical simulation methods which are markedly different from numerical discretization methods. Numerical discretization methods ordinarily describe a mathematical system modeled from some physical system or process. However, Monte

n	x_n
-4	8
-3	6
-2	8
-1	11
0	3
1	3
2	4
3	11
4	1
5	7
6	1
7	5
8	5
9	2
10	12

Table 2.3. Pseudo-random numbers from a fictional LFG

Carlo can generally be used to simulate a given physical process directly without need for differential or integral equations that model the system. As a general example, Monte Carlo can be used to approximate the value of definite integrals. The general strategy is to select a control area that encapsulates the domain of interest. Then, the area being considered can be estimated by a ratio of random points that fall within the domain of interest to the overall control area. In this case, the scalar outcome provides an estimation of area under a one or multi-dimensional function. However, for the application under consideration, the physical system isn't a simple mathematical function; instead it's a tennis match and each point played can be simulated directly with an appropriate model. The main requirement for development of this simulation, then, would be knowledge of the probability density function associated with the random variable describing the event of winning a point in tennis for a given player. Success or failure of the repeatable event,

i.e. win/loss for each point played in the tennis match, can then be determined through sampling random numbers to determine success or failure. The development of the necessary *pdf*s and execution of these individual trials leads to a natural implementation using a Monte Carlo approach. On the other hand, a numerical discretization method for this problem would require a set of physical differential or integral equations to be developed for each point played in a tennis match along with the corresponding algebraic solution, a perilous task at best.

This discussion highlights the fact that Monte Carlo may or may not be ideally suited for a given mathematical problem. It's natural to think that Monte Carlo methods would be used to simulate random processes since these can be described inherently with probability densities. In particular, the theory presented in subsequent chapters and pioneered by Newton and Keener [34] describes a stochastic process that leads to the probability of winning a game, set and match in tennis. The Monte Carlo method, therefore, provides an extremely efficient and adaptable format for characterizing player performance, exploring various second order effects and evaluating other nonlinearities that can't be achieved with the analytical model alone. A novel approach is described that presents the outcome of these simulations as *pdf*s, hence as continuous random variables. This results in *pdf* based player models that can be used as building blocks to run subsequent scenarios. In this way, the Monte Carlo tennis simulation is an adaptive stochastic model that produces detailed, well understood metrics in the field of applied mathematics and statistics. To date no other Monte Carlo analysis has sought to produce probability density functions for the sake of a statistical review into non-iid (independent and identically distributed) effects, analysis of density function (performance) trending and application towards predictive ranking systems.

Chapter 3.

The Markov Chain Model for Tennis

Here, a Monte Carlo method is described which can be used to calculate the probability that a given player wins a game, set, match and single elimination tournament in tennis. The corresponding analytical theory, documented fully by Newton and Keener in [34] is based on each player's probability of winning a point on serve (denoted p_A^R for player A). In the initial treatment, the values $p_A^R \in [0,1]$ and $p_B^R \in [0,1]$ for player A and player B are assumed constant throughout each match *and* tournament. This assumption means that points in tennis are taken as independent, identically distributed (iid) random variables with *standard deviation of zero*. In practice, these values are obtained empirically from each of the player's statistics gathered over enough matches and against different opponents so that the accumulated value of the ratio of points won on serve over total points served can be used predictively. This ratio converges quite rapidly to a nearly constant value for each player, as shown for example, in Figure 3.1. This figure shows data for Pete Sampras, the winner of the 2002 US Open men's singles event, his last tournament win before retirement. The data points show the ratio of total points won on serve over total points served through the seven rounds of the tournament. The final data point ($n = 7$) contains all information on Sampras that was accumulated through the entire tournament, hence it can be viewed

as an average value over his field of opponents for the 2002 US Open. Specifically, the ratio for Pete Sampras converges fairly rapidly to its cumulative average of 0.73. This value would then be used as the input parameter for Sampras in the analytical approach.

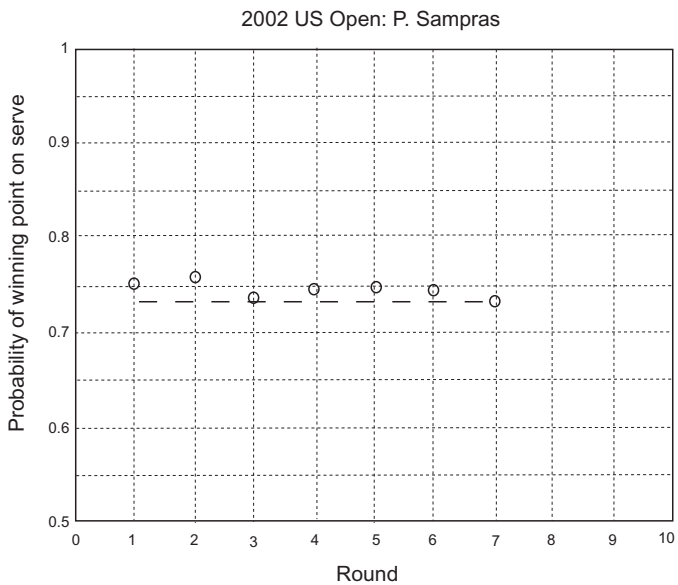


Figure 3.1. Convergence of p_A^R to the sample mean $\mu = 0.73$ for P. Sampras through the 7 rounds of the 2002 US Open

Using $p_A^R = 0.73$ for Sampras, it can be asked what the probability of defeating him would be given the full range of values for p_B^R. Figure 3.2 shows the results based on the analytical model. The curves depict the probabilities of winning a game, set and match against Sampras along with data from the 2002 US Open.

A general conclusion based on the steepness of the analytical curves throughout the typical range encountered on the professional tennis circuit ($0.60 < p_A^R, p_B^R < 0.75$) together with the fact that the curves for sets are steeper than those for games and steeper yet for matches is that *the better player usually wins in tennis* – the scoring system conspires against the weaker player.

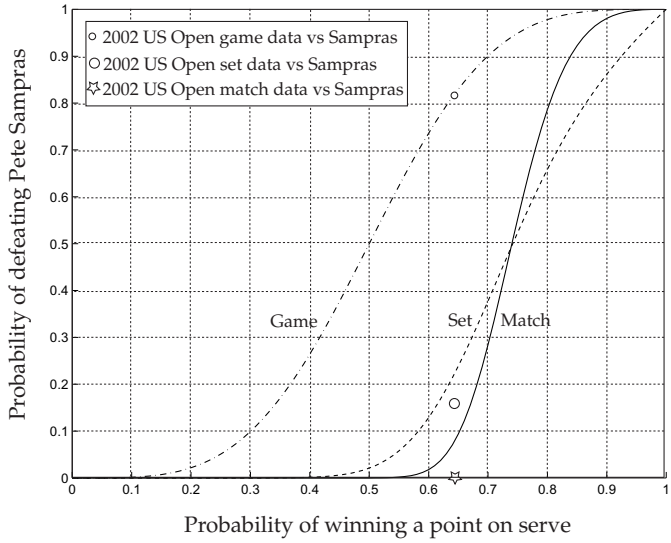


Figure 3.2. Probability of winning against P. Sampras in the 2002 US Open tournament

While a server who wins a respectable 60% of his points on serve will win well over 70% of his service games, he'll win under 15% of the sets and only around 1% of his matches against Sampras. Relatively small differences in the abilities between players are amplified relentlessly against the weaker player with the way the scoring system is constructed. In addition, because of the fact that the top players are spread throughout the tournament draws based on the seeding system, they tend to meet often (typically in the semifinals or finals) during a tournament season. These facts should, in principle, make ranking systems for tennis easier to construct than in other sports which have a more random component in matchups making upsets more common. In fact, several other popular spectator sports have top ranked teams that may never play each other during a season.

There are some important points to make. First, despite the fact that the convergence shown in Figure 3.1 is rapid, there are always fluctuations of the higher order moments around the mean.

These fluctuations are typically small (roughly 1%) compared to the difference with the average of other players. Hence, on a match-by-match basis, each player's ratio of points won on serve to points served varies somewhat from his accumulated value gathered over large numbers of matches against different opponents. In the case of Pete Sampras for the 2002 US Open, his match mean (i.e. the mean value of the seven ratios associated with each match) was $\mu = 0.7392$ with a standard deviation of $\sigma = 0.0314$. For women, the match mean tends to be lower but the variation tends to be higher. The match mean associated with Serena Williams in the 2002 US Open and Wimbledon tournaments, both of which she won, was $\mu = 0.7158$ with a standard deviation of $\sigma = 0.0762$. In practice, this means that for any given match, even with a large amount of data in hand, there's some uncertainty as to what actual value to take for p_A^R and p_B^R for each of the players. This uncertainty is addressed and quantified by introducing the concept of "the field" in Chapter §4. When examining targeted homogeneous data sets such as the Borg-McEnroe series of head-to-head matches, there's some evidence of non-iid effects creeping in, such as the so called "back-to-the-wall" and "hot-hand" effects (see the works of Jackson and Mosurski in [17] along with Klaassen and Magnus in [20] for more discussions). Other aspects of tennis that potentially introduce non-iid effects are the introduction of new balls (documented by Klaassen and Magnus in [25]) or psychological factors that could be present in the first or final sets of a match (also documented by Klaassen and Magnus in [26] and [27]). Although possible, these are thought of as second order effects here but in close matches they play prominant roles. Impacts of these effects to the analytical model are studied in further detail in Chapter §3.

It's natural to ask how accurate the analytical model is in predicting probabilities of winning *individual* matches given that it uses fixed values of p_A^R and p_B^R. In principle, the theory could

be compared with data gathered from tournaments, however this is difficult. While the number of points and games played by each player in a typical match and tournament is large enough to extract meaningful statistics (as shown by Newton and Keller in [34]) when comparing the theory with data for points and games) the number of sets and matches played aren't nearly enough for these purposes. What's worse, the analytical formulas predicting set and match probabilities are functions of both p_A^R *and* p_B^R, while the formulas for predicting tournament outcomes are functions of up to 128 variables, one for each of the players in the tournament. Hence, gathering data for an individual player, say player A, requires a look at that player's matches only against opponents with the same value of p_B^R reducing the data set even further. It might be tempting to look at data for each player over an entire season of tournaments against opponents with one value of p_B^R, but this introduces other problems as tournaments are played on several different surfaces and player characteristics can vary widely from surface to surface and match to match. For example, although Pete Sampras was dominant on grass, winning a record seven Wimbledon singles titles, he never did well at the French Open which is played on the much slower clay surface.

For all of these reasons, it's desirable to develop a Monte Carlo approach that's capable of generating large data sets quickly and reliably that would be difficult if not impossible to gather otherwise. A repeatable simulation can be used to evaluate the robustness of the iid assumption adopted by Newton and Keener and used in other analytical approaches that pre-date that work, such as those of Carter and Crews in [8] and Pollard in [39]. As discussed in more detail later, such a simulation is used as the basis for development of *pdf* based player models that address some of the limitations of the analytical approach. Namely, such *pdf* player models facilitate fair statistical comparisons between players with little or no match history. Futher, the simulation and

it's resulting player models provide the cornerstone in the development of probabilistic ranking systems that address limitations with ones currently in use such as the Bowl Championship Series (BCS) as fully described by Callahan, Mucha and Porter in [7]. The search for evidence of non-iid effects has been pursued in several sports with mixed success. In basketball, an investigation into whether or not points are iid was pioneered by Tversky and Gilovich in [46]. The analysis of consecutive basketball shots shows that, contrary to popular belief, the chances of a player making a basket are as good after a miss as after a success and thus they found no evidence of a "hot-hand effect." A similar analysis of hitting streaks in baseball by Albright in [1] and Stern and Morris in [44] also failed to detect any significant effects on the probability of getting a hit due to a player's recent history of success or failure. In tennis, the question of whether points are independent events was addressed by Klaassen and Magnus in [20] by performing a statistical analysis of $90,000$ points at Wimbledon collected over a wide range of matches. In that work there was some evidence found that winning the previous point had a positive effect on winning the next one and that "important" points were more difficult to win for the server than less important ones. Their ultimate conclusion, however, was that although points in tennis aren't exactly iid random variables, the deviation from iid is small. A recent attempt to model some of these non-iid effects in tennis can be found in the work of Jackson and Mosurski [17].

A Monte Carlo approach is introduced which can be used to computationally characterize some of the second order effects described, refine the analytical model with development of robust player models (i.e. *pdf* models with *non-zero* standard deviations) and explore probabilistic rankings that provide implicit comparative meaning instead of conventional, deterministic-based ones. This approach can be used to investigate the effects of small deviations from the iid model

and to explore some specific non-iid factors. It also characterizes the robustness of the analytical approach. Finally, the framework for utilizing *pdf* based player models to formulate probabilistic rankings is introduced.

3.1 Analytical Model

First, a brief review is presented on the analytical model that provides the basis for the Monte Carlo computational approach. In order to calculate the probability that one player, player A, wins a tennis match against another player, player B, it suffices to know the probability p_A^R that player A wins a rally when he serves and the probability p_B^R that player B wins a rally when he serves. When these two independent parameters are held constant throughout a match, explicit formulas for the probabilities of winning a game, set and match for each player can be obtained. For example, the probability of player A winning a game on serve, p_A^G, is given by

$$p_A^G = (p_A^R)^4[1+4q_A^R+10(q_A^R)^2]+20(p_A^R q_A^R)^3(p_A^R)^2[1-2p_A^R q_A^R]^{-1} \tag{3.1}$$

where $q_A^R = 1 - p_A^R$. Note that (3.1) depends only on characteristics from player A and not on player B. This simple, explicit and compact formula that encodes the game scoring system, was first documented by Carter and Crews in [8]. As previously described, its relatively steep slope in the region of interest for most players ($0.60 \leq p_A^R \leq 0.75$) highlights the amplification of small differences in player abilities, making upsets more rare in tennis than in other sports such as football and basketball.

To obtain corresponding formulas for the probability of winning a set and a match, 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, 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

$$p_A^S = \sum_{j=0}^{4} p_A^S(6,j) + p_A^S(7,5) + p_A^S(6,6)p_A^T \tag{3.2}$$

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 the terms $p_A^S(i,j)$ needed in (3.2), the following system of recursive equations is solved: For $0 \le i, j \le 6$:

$$\begin{aligned} if\, i-1+j\, is\, even: \qquad & p_A^S(i,j) = p_A^S(i-1,j)p_A^G + p_A^S(i,j-1)q_A^G \\ & omit\ i-1\, term\, if\ j=6, i \le 5 \\ & omit\ j-1\, term\, if\ i=6, j \le 5 \end{aligned} \tag{3.3}$$

$$\begin{aligned} if\, i-1+j\, is\, odd: \qquad & p_A^S(i,j) = p_A^S(i-1,j)q_B^G + p_A^S(i,j-1)p_B^G \\ & omit\ i-1\, term\, if\ j=6, i \le 5 \\ & omit\ j-1\, term\, if\ i=6, j \le 5 \end{aligned} \tag{3.4}$$

along with the initial conditions:

$$p_A^S(0,0) = 1; \quad p_A^S(i,j) = 0 \qquad if\, i < 0\, or\, j < 0 \tag{3.5}$$

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

$$p_A^S(7,5) = p_A^S(6,5)q_B^G; \quad p_A^S(5,7) = p_A^S(5,6)p_B^G \tag{3.6}$$

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

$$p_A^T = \sum_{j=0}^{5} p_A^T(7,j) + p_A^T(6,6) \sum_{n=0}^{\infty} p_A^T(n+2,n) \tag{3.7}$$

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

$$p_A^T(n+2,n) = \sum_{j=0}^{n} (p_A^R p_B^R)^j \left(q_A^R q_B^R\right)^{n-j} \frac{n!}{j!(n-j)!} p_A^R q_B^R = (p_A^R p_B^R + q_A^R q_B^R)^n p_A^R q_B^R \tag{3.8}$$

Using (3.8) in (3.7) and summing yields

$$p_A^T = \sum_{j=0}^{5} p_A^T(7,j) + p_A^T(6,6) p_A^R q_B^R \left[1 - p_A^R p_B^R - q_A^R q_B^R\right]^{-1} \tag{3.9}$$

To calculate $p_A^T(i,j)$, solve:

For $0 \leq i,j \leq 7$:

$$if\ i-1+j = \qquad 0,3,4,...,4n-1,4n,...$$

$$p_A^T(i,j) = p_A^T(i-1,j)p_A^R + p_A^T(i,j-1)q_A^R \tag{3.10}$$

$$omit\ j-1\ term\ if\ i=7, j\leq 6$$

$$omit\ i-1\ term\ if\ j=7, i\leq 6$$

$$if\ i-1+j= \quad 1,2,5,6,...,4n+1,4n+2,...$$

$$p_A^T(i,j)=p_A^T(i-1,j)q_B^R+p_A^T(i,j-1)p_B^R \tag{3.11}$$

$$omit\ j-1\ term\ if\ i=7, j\leq 6$$

$$omit\ i-1\ term\ if\ j=7, i\leq 6$$

with the initial conditions:

$$p_A^T(0,0)=1; \quad p_A^T(i,j)=0 \qquad if\ i<0\ or\ j<0 \tag{3.12}$$

p_A^T is then calculated using the solution of (3.10) - (3.12) in (3.9). This allows p_A^S to be calculated using the solution of (3.3) - (3.6) with the result for p_A^T in (3.2). More details along with all the solutions of the recursion formulas are documented by Newton and Keller in [34]. Note that the formulas for winning a tiebreaker, set and match for each player are functions of *both* p_A^R and p_B^R in contrast to the formula for winning a game in (3.1).

A byproduct of the formulas described here is an explicit proof that tennis scoring is "service neutral". As long as the iid assumption is used, the probability of winning a set or a match is independent of which player serves first. This rather surprising result is proven more directly (but less explicitly) by Newton and Pollard in [35]. A statistical analysis examining this result is described by Magnus and Klaassen in [27]. In addition, Newton and Keller have documented the

process of "handicapping" a tournament. For example, at each round all players' probabilities of winning their next match can be computed. This concept can be carried forward to calculating probabilities that a given player ultimately becomes the tournament champion. As a computational analog, corresponding Monte Carlo simulated predictions for the 2002 US Open and 2002 Wimbledon singles events are shown in Section §3.5.

3.2 Description of the Monte Carlo Approach

The starting point for the Monte Carlo tennis simulation is a random number generator capable of generating values for success or failure based on p_A^R and p_B^R sampled from a uniform distribution on the interval $[0,1]$. When player A is serving, for each point a value on the unit interval is sampled. If the value lies in the range $[0, p_A^R]$, player A wins that point. Otherwise player B wins the point. Similarly, when player B is serving, for each point a value is sampled on the unit interval. If the value lies in the range $[0, p_B^R]$, player B wins the point. Otherwise player A wins the point. The point-by-point simulation proceeds in this way governed by the scoring rules of tennis and statistics are gathered to show the number of points, games, sets and matches won by each player. For this purpose, the pseudo-random number generator algorithm RAND in Matlab is suitable, and the statistics generated from a sequence of trials are discussed below.

3.3 Results of Simulations

3.3.1 Convergence to Analytical Theory

Figure 3.3 shows the convergence results for games as a function of trials. The analytical curve (also depicted in Figure 3.2) is shown together with the statistics based on 1000 realizations of games in Figure 3.3(a). Figure 3.3(b) shows the Gaussian convergence for one of the data points ($p_A^R = 0.5$) which converges most slowly to the analytical value ($p_A^G = 0.5$) for 10, 100 and 1000 trials. Figure 3.3(c) shows the standard deviation as a function of the number of games plotted on a log-log scale.

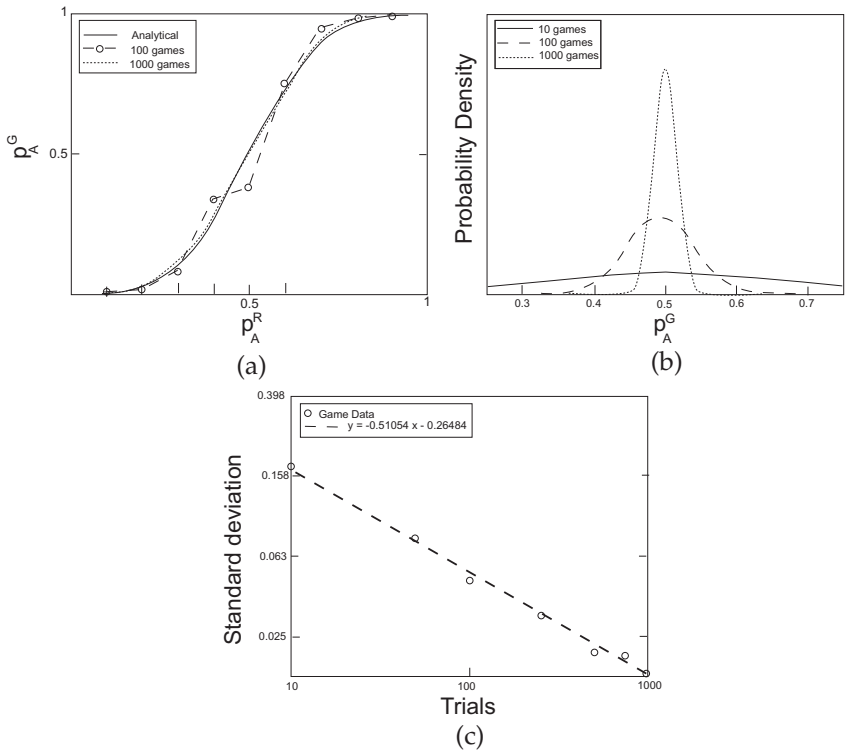


Figure 3.3. Convergence of the Monte Carlo simulation to the analytical curve for games

These data show power law convergence $\sigma \sim \alpha n^{-\beta}$ with power law exponent $\beta \sim 0.511$. Figures 3.4 and 3.5 show the corresponding results for sets and matches against a neutral server $p_B^R = 0.5$. The power law exponent for sets, as seen in Figure 3.4(c), is slightly larger than games, $\beta \sim 0.611$, whereas the convergence rate for matches is slightly slower, with exponent $\beta \sim 0.475$.

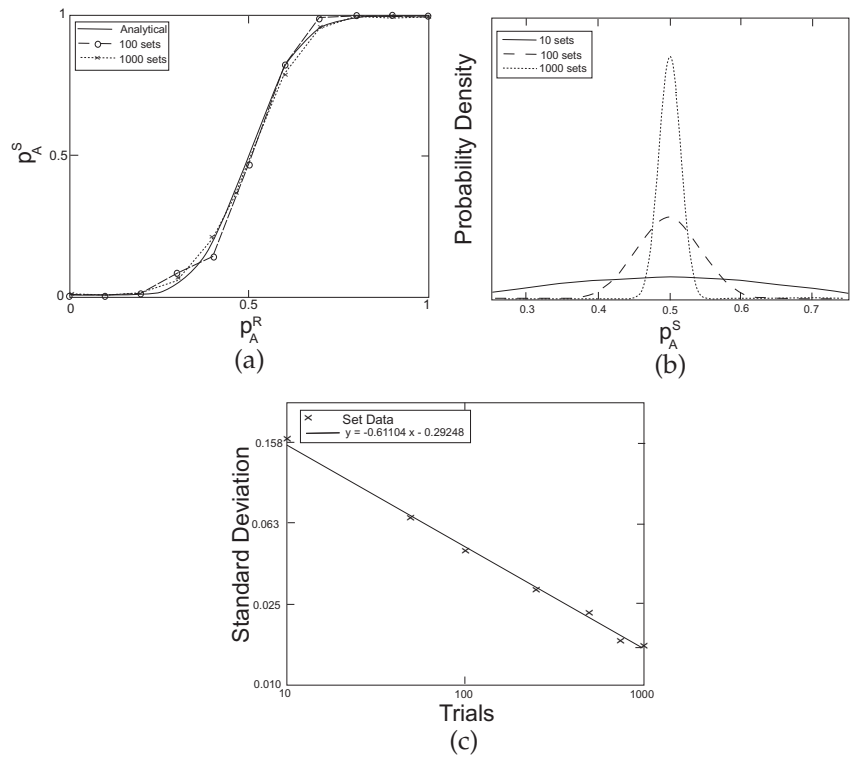


Figure 3.4. Convergence of the Monte Carlo simulation to the analytical curves against a neutral server $p_B^R = 0.5$ for sets

This is expected as match realizations are built upon multiple set and game realizations. In all cases, after 1000 realizations, convergence is sufficiently close to the analytical curves and is uniform throughout the entire range of values of p_A^R. Thus, in practice, 1000 realizations can be used with confidence to ensure computational convergence.

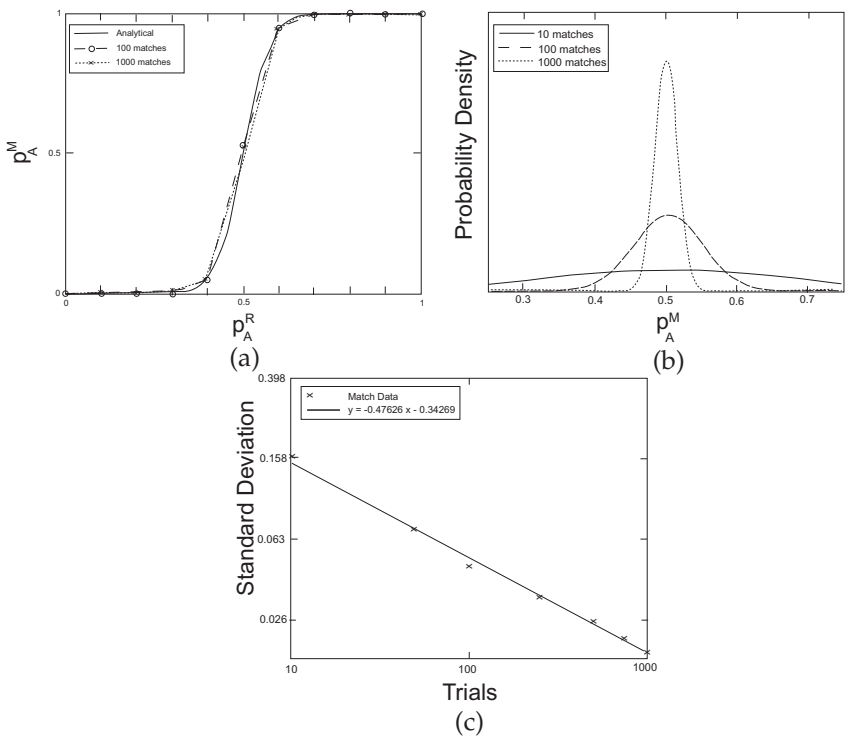


Figure 3.5. Convergence of the Monte Carlo simulation to the analytical curves against a neutral server $p_B^R = 0.5$ for matches

Figure 3.6(a) shows the results of p_A^S as a function of p_A^R for an entire spectrum of opponents, hence $p_B^R = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Again, for 1000 realizations the convergence to the analytical curves is uniform throughout the full range. The same is true of Figure 3.6(b) which shows results for p_A^M (3 out of 5 set format) as a function of p_A^R against the full spectrum of opponents. The main conclusion from these test runs is that 1000 trials is sufficient to ensure that the statistics are accurate throughout the entire range of both parameters p_A^R and p_B^R.

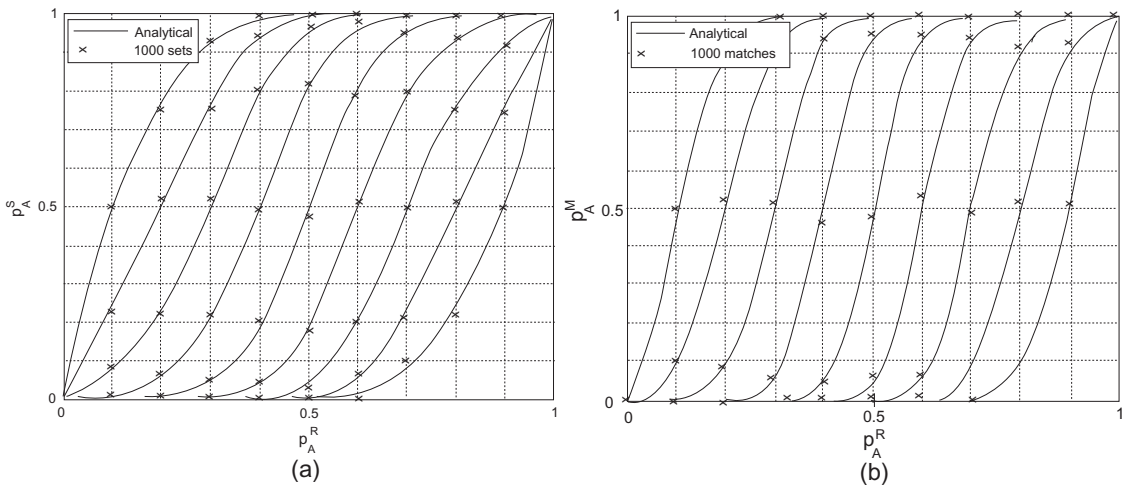


Figure 3.6. Convergence of the Monte Carlo simulation to the analytical curves against a spectrum of servers $0.0 < p_B^R < 1.0$ for sets and matches

3.3.2 Robustness of the Analytical Model

To test the robustness of the analytical model based on the iid assumption, the values of p_A^R and p_B^R can be perturbed to see the effect this has on the shapes of the curves. Symmetric perturbations are used first, taken from a uniform distribution centered at the nominal values of p_A^R. In Figure 3.7, the curve for p_A^G which results from whitenoise perturbations added after each point is shown

with amplitudes as high as 20% off the nominal value of p_A^R. The figure shows relatively quick convergence to the analytical curve throughout the full range of values. Similar convergence properties hold when perturbing both p_A^R and p_B^R in the case of the curves for p_A^S and p_A^M.

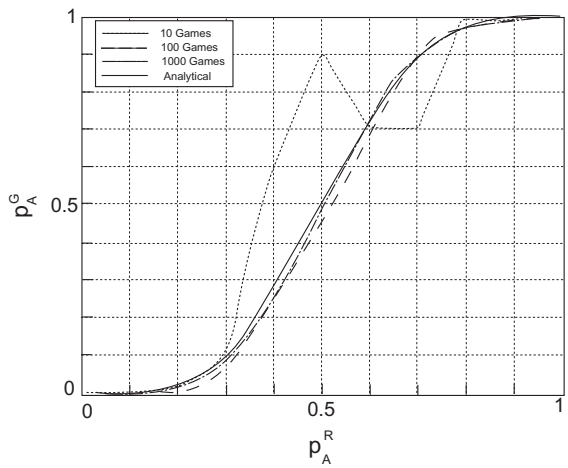


Figure 3.7. Effects of symmetric unbiased whitenoise perturbations with 20% amplitude

3.4 Non-iid Models

3.4.1 Importance

Numerous theories exist claiming that points in a tennis match aren't equally important to determining its outcome. Although this is generally assumed by most professional players and commentators, there isn't uniform agreement on which points are most important. Some argue that the first point of a game is the most important as it's crucial for players to get off to a good start, while probably the most frequently specifically cited points are 15-30 and 30-15 as most important. What is uniformly agreed upon is that great players adjust their efforts according to

which points, games and sets they feel are the most crucial towards winning a match. In fact, one of the qualities frequently cited as a sign of a great champion is the ability of that player to focus on key points and to be able to raise his level of play accordingly. An example of a player who seemed to have this ability was Pete Sampras, but Bjorn Borg and Chris Evert are also frequently cited as well.

If players do change their level of effort according to the point, game or set then not all points, games or sets would be identical and the iid assumption on which Newton and Keller's theory in [34] is based would need some modification. To carry out this modification, a method for quantifying the importance of a point, game or set is described based on the original formulation of Morris in [30]. He defines the *importance* of a point to a game as the difference between two conditional probabilities: the probability that the server wins the game given that he wins the point and the probability that he wins the game given that he loses the point. If I_{ij}^{P} denotes the importance of point i for the server and j for the receiver for winning the game, then

$$I_{ij}^{P} = P_{i+1,j}^{G} - P_{i,j+1}^{G} \tag{3.13}$$

where P_{ij}^{G} is defined as the probability that the server will win the game given that the score is i points for the server and j points for the receiver. In a similar way, the importance of a given game toward winning a set can be defined as

$$I_{ij}^{G} = P_{i+1,j}^{S} - P_{i,j+1}^{S} \tag{3.14}$$

where P^S_{ij} denotes the probability that the first server will win the set given that the score is i games for the first server and j games for the first receiver. The importance of each point towards winning a tiebreaker, I^T_{ij} is defined as

$$I^T_{ij} = P^T_{i+1,j} - P^T_{i,j+1} \tag{3.15}$$

where P^T_{ij} is defined as the probability that the first server will win the tiebreaker given that the score is i points for the first server and j points for the first receiver. Finally, the importance of each set towards winning a match is defined as

$$I^S_{ij} = P^M_{i+1,j} - P^M_{i,j+1} \tag{3.16}$$

where P^M_{ij} denotes the probability that the first server will win the match given that the score is i sets for the first server and j sets for the first receiver. The importance of each point, game and set in terms of the variables p^R_A and p^R_B can then be calculated. To obtain the terms for the importance of each point in a game, the following system must be hierarchically expanded

$$P^G_{ij} = p^R_A P^G_{i+1\,j} + q^R_A P^G_{ij+1} \quad i,j = 0,1,2 \tag{3.17}$$

and when $i = 3$ or $j = 3$ use

$$\begin{aligned} P^G_{31} &= p^R_A + q^R_A P^G_{32} \\ P^G_{13} &= p^R_A P^G_{23} \end{aligned}$$

$$
\begin{aligned}
p_{30}^{G} &= p_A^R + q_A^R P_{31}^G \\
p_{03}^{G} &= p_A^R P_{13}^G
\end{aligned}
\tag{3.18}
$$

Since the probability of winning a game when the score is 0 to 0 is simply p_A^G, the following initial condition exists using (3.1)

$$
P_{00}^G \equiv p_A^G = (p_A^R)^4[1+4q_A^R+10(q_A^R)^2]+20(p_A^R q_A^R)^3(p_A^R)^2[1-2p_A^R q_A^R]^{-1} \tag{3.19}
$$

For example, to compute $I_{2,2}^P$ the formulas for $P_{3,2}^G$ and $P_{2,3}^G$ are needed.

$$
\begin{aligned}
P_{3,2}^{G} &= p_A^R + q_A^R P_{3,3}^G \\
P_{2,3}^{G} &= p_A^R P_{3,3}^G
\end{aligned}
\tag{3.20}
$$

Then

$$
I_{2,2}^P = p_A^R + q_A^R P_{3,3}^G - p_A^R P_{3,3}^G \tag{3.21}
$$

The formula for $P_{3,3}^G$ is obtained from [34] as

$$
P_{3,3}^G = (p_A^R)^2[1-2p_A^R q_A^R]^{-1} \tag{3.22}
$$

The results for the importance of each point in a game are shown in Figure 3.8. Each subfigure shows the importance curves as a function of p_A^R with the server having 0, 1, 2 or 3 points, respectively.

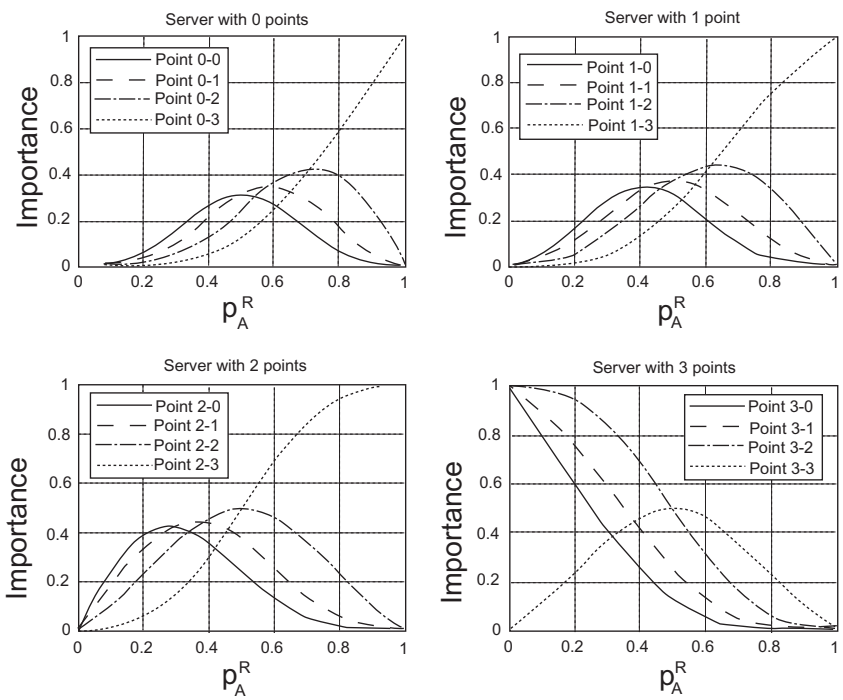


Figure 3.8. Importance of each point in a game as a function of p_A^R

The curves reveal interesting trends for both strong servers ($p_A^R > 0.5$) and weak servers($p_A^R < 0.5$). Figure 3.9 shows the point with maximum and minimum importance as a function of p_A^R.

In the region $0 \leq p_A^R \leq 0.5$, the most important point is 40-30 and the least important point is 0-40, whereas in the region $0.5 \leq p_A^R \leq 1.0$, the typical range for professional players, the most important point is 30-40 while the least important point is 40-0. These mathematically based results are at odds with the commonly suggested points of 15-30 and 30-15 as being most important.

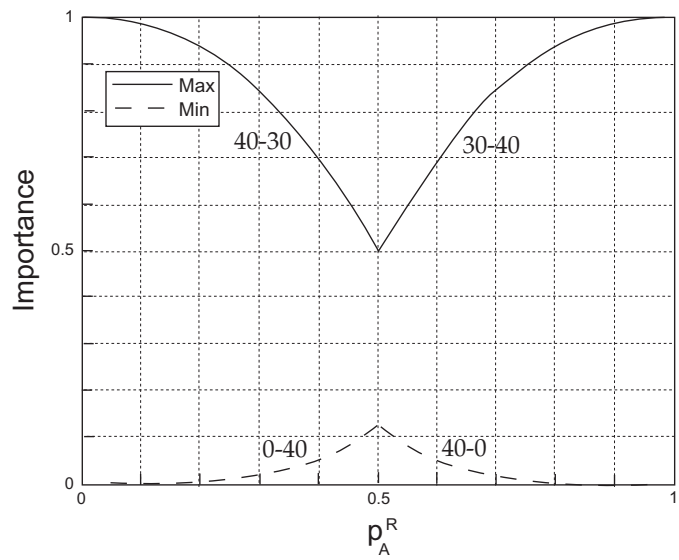


Figure 3.9. Maximum and minimal importance of a game as a function of p_A^R

The importance of each point in a tiebreaker, set and match can also be obtained by expanding similar hierarchical systems of equations, taking care to keep track of the alternating service games between players A and B throughout the set as well as the serving order in a tiebreaker. For example, if playing against Pete Sampras, the importance of each point in a tiebreaker and the importance of each game towards winning a set as a function of p_A^R are shown in Figure 3.10 and Figure 3.11, respectively. The gray zones in these figures denote typical values of p_A^R at the professional level; $0.6 < p_A^R < 0.75$.

Again, the detailed values of the importance of points in a tiebreaker and games in a set depends on the value of both p_A^R and p_B^R. This information can be used to test the effects of a non-iid model based on the following idea. Suppose player A adjusts his level of play according to the importance of each point. In the simplest case, adjust player A's nominal value of p_A^R by 20%, increasing the value on the most important point of the game (40-30 or 30-40) and decreasing the

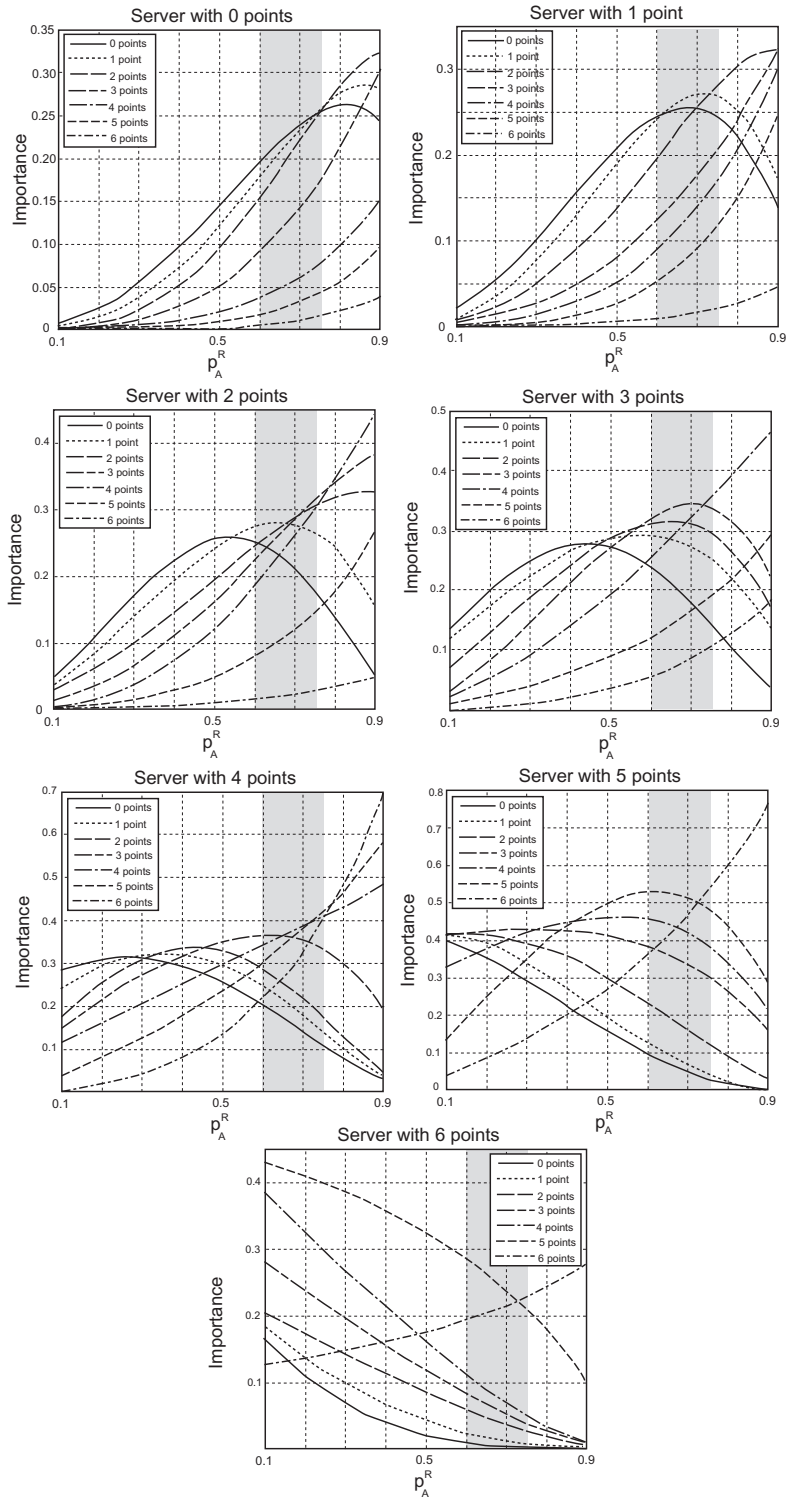


Figure 3.10. Importance of points in a tiebreaker against P. Sampras ($p_B^R = 0.73$)

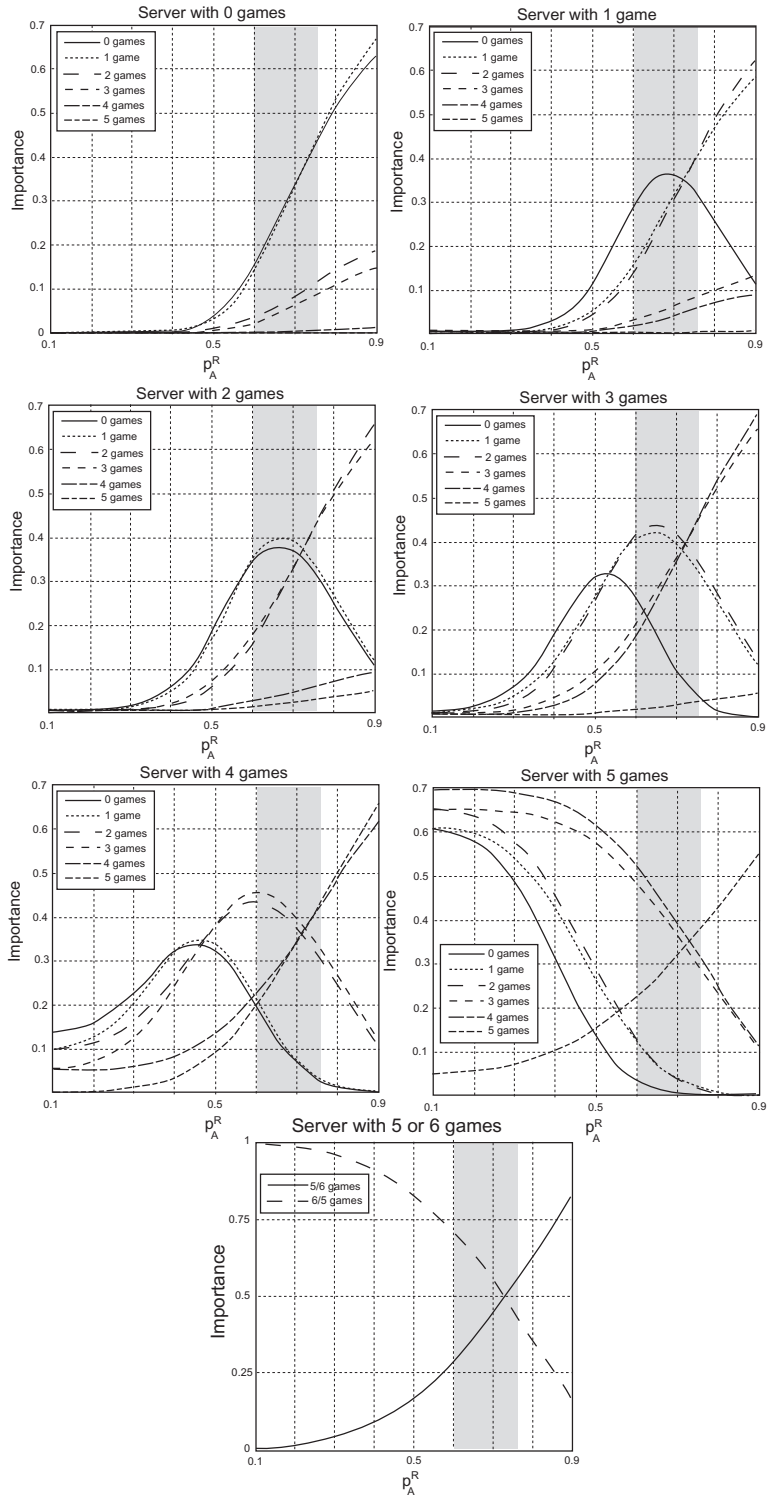


Figure 3.11. Importance of games in a set against P. Sampras ($p_B^R = 0.73$)

value by the same amount on the least important point (40-0 or 0-40). Note that the least important point in a game can only occur once per game whereas the most important point can occur many times (advantage-in during a deuce is equivalent to 40-30 and advantage-out is equivalent to 30-40). The effect of this model is shown in Figure 3.12 for the values $p_B^R = \{0.2, 0.4, 0.6, 0.8\}$ and for the full range of values of p_A^R.

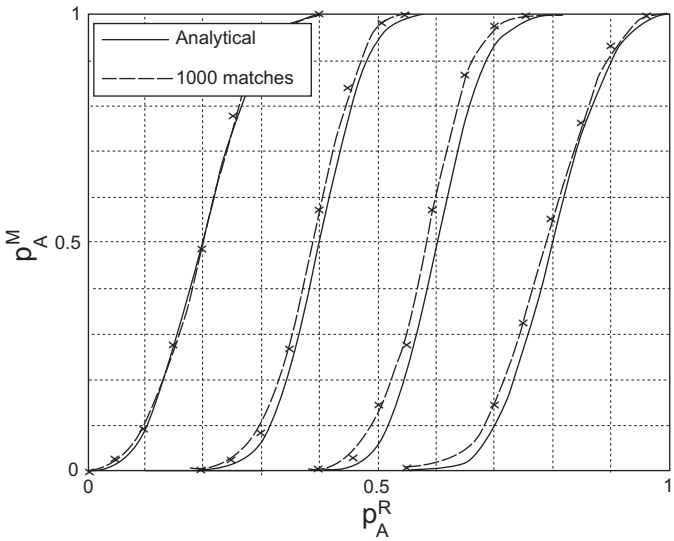


Figure 3.12. Non-iid effects based on adjusting play according to the importance of points for a range of values of p_B^R

The overall effect is that the curves are shifted slightly up from the baseline iid theory, i.e player A's probability of winning is systematically increased using this approach. One key reason is that the adjustment *up* on the most important point occurs more frequently than the adjustment *down* on the least important point, hence the effective value for p_A^R is slightly higher than that in the nominal iid case. In addition, players exhibit an increase in performance by focusing their

effort adjustments solely on points that will result in the greatest impact to their probability of winning the match.

3.4.2 The Hot-Hand Effect: Does Success Breed Success?

To model the hot-hand effect, a perturbation is made to each player's value of p_A^R or p_B^R on the one point immediately following each point won. Figure 3.13 shows the result of perturbations with 20% amplitude taken from a uniform distribution evenly distributed around the analytical curve.

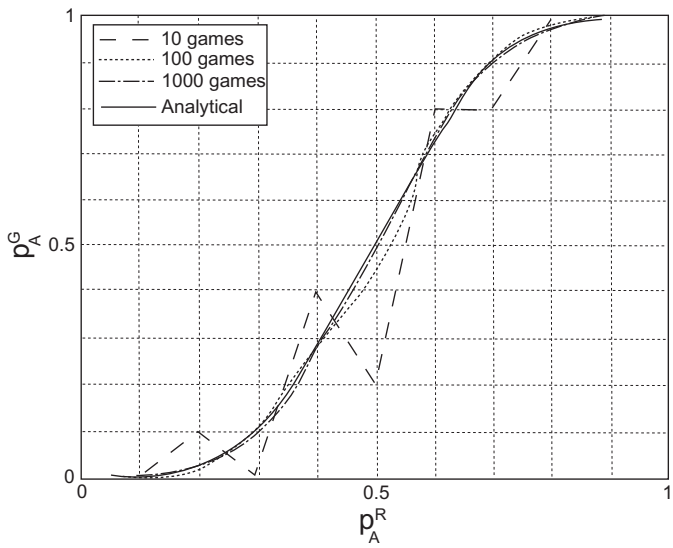


Figure 3.13. Random whitenoise hot-hand perturbations with 20% amplitude

The figure shows relatively rapid convergence to the iid curves for 10, 100 and 1000 trials. After 1000 trials, the convergence to the analytical curves is uniform throughout the range. This is somewhat surprising given the size of the perturbations and their random nature. However, because they're taken from a uniform distribution symmetric about the analytical curves, the increased bumps and decreased bumps of p_A^R effectively cancel each other after a sufficiently

large number of trials. Contrast this with Figure 3.14 which shows the results from hot-hand perturbations that are still large (20% size amplitudes) but aren't symmetric about the analytical curve.

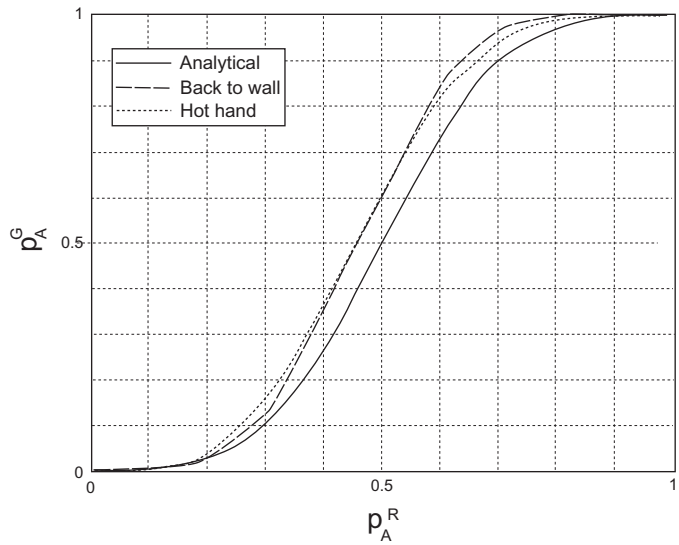


Figure 3.14. Curves showing the effect of hot-hand perturbations and back-to-the-wall perturbations with 20% amplitude

Here, an *increase* to the nominal value of p_A^R occurs on each point after a point is won. The result, after 1000 trials, doesn't collapse back to the iid curves but instead shows a systematic shift upwards as one would expect. Hence, the server's probability of winning a game is increased in a quantifiable way over what it would be from the pure iid theory.

3.4.3 Back-To-The-Wall Effect: Does Failure Breed Success?

To model the back-to-the-wall effect, a perturbation is made to p_A^R by a fixed percentage on the one point immediately following each point that the player loses. Figure 3.15 shows the analog

of Figure 3.13 with perturbations of 20% amplitude taken from a uniform distribution evenly centered around the analytical curve.

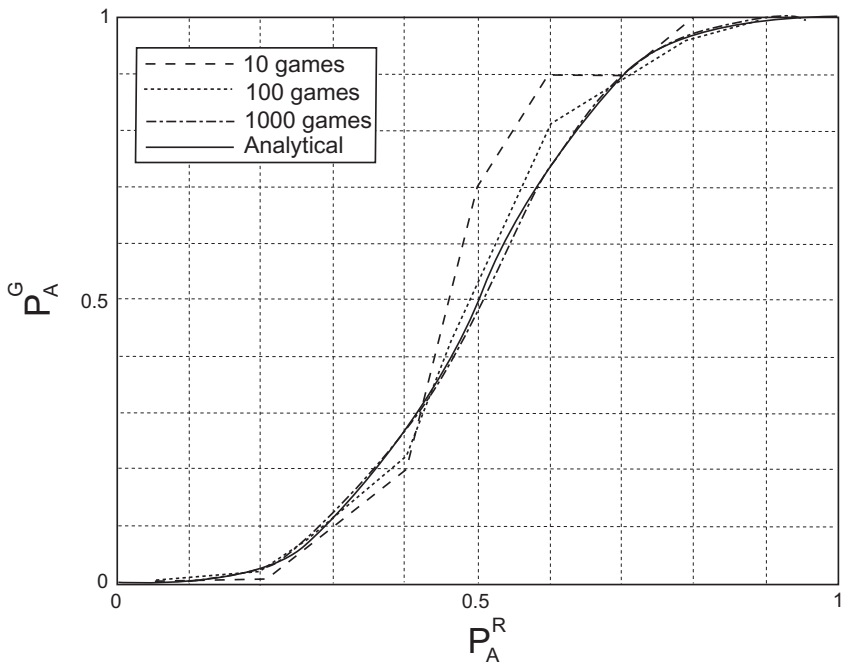


Figure 3.15. Random whitenoise back-to-the-wall perturbations with 20% amplitude

The figure again shows relatively rapid convergence to the iid curves for 10, 100 and 1000 trials. After 1000 trials, the convergence to the analytical curves is uniform throughout the range. Refer back to Figure 3.14 showing the results from back-to-the-wall perturbations that are are still large (20% amplitude) but aren't symmetric about the analytical curve. Here, an *increase* to the nominal value of p_A^R occurs on each point after a point is lost. Again, the result after 1000 trials doesn't converge to the iid curves but shows a systematic shift upwards. As with the hot-hand perturbations, the server's probability of winning a game is increased over what it would be from the pure iid theory.

3.5 Multi-Variable Tournament Realizations

To fully illustrate the computational capability of the Monte Carlo simulation, tournament simulation realizations were executed to compare predicted outcomes with actual, observed performance. Figure 3.16 shows the 2002 US Open Mens draw from the semifinal round onward, with the values of p_A^R and p_B^R collected for each of the four players over the previous rounds of the tournament.

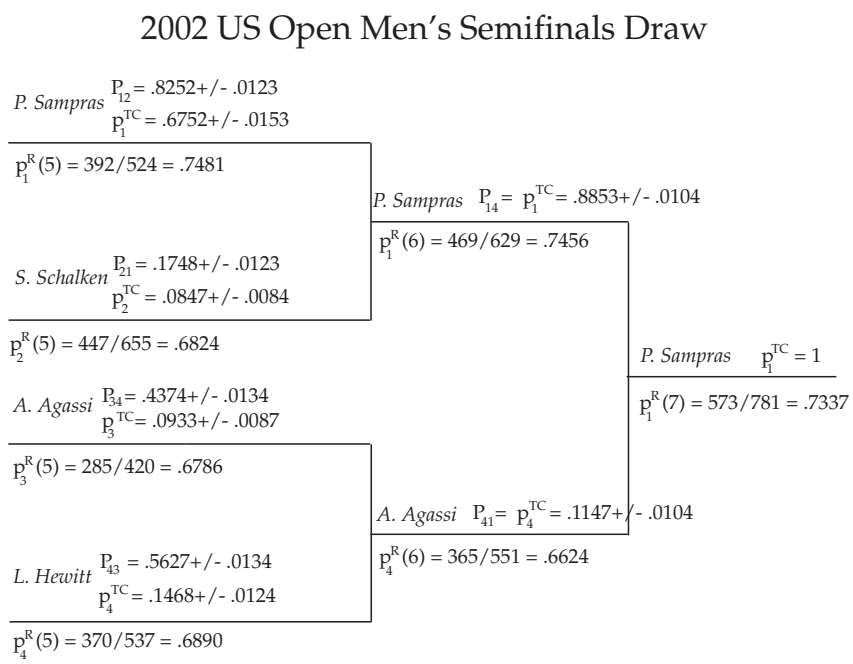


Figure 3.16. Tournament predictions based on Monte Carlo simulations of the 2002 US Open Men's draw using data through the quarterfinal round

These values are listed under each player's name as $p_i^R(n)$ where $i = \{1,2,3,4\}$ indicates the player and n indicates the number of previous rounds over which the data was collected. Using the values $p_i^R(5)$ for each of the players, 1000 simulated matches were run for each of the final two

rounds and statistics were gathered providing the probabilities for each player to advance (P_{ij}) and for each to become the tournament champion (superscript 'TC'). In each case, error values of one standard deviation (σ) are included. Analytical results for the exact same case are presented by Newton and Keller in [34] and for all computed values the agreement is outstanding. At worst, the difference between analytical and computational values is 0.13% (computational $P_{43} = 56.27\%$ versus analytical $P_{43} = 56.14\%$) which is a full order of magnitude less than the computational standard deviation obtained in that case of $\sigma_{43} = 1.34\%$. More specifically, for all comparisons the analytical predictions are within an order of magnitude less than one standard deviation of the computational values. The exercise demonstrates wonderful correlation between the Monte Carlo simulation and the fully documented analytical model even for multi-player, tournament based evaluations.

The results provide a glimpse into the power of the Monte Carlo code and to its potential usefulness and flexibility both in simulating full tournaments with fixed p_A^R and p_B^R, but also for performing non-iid simulations that vary these parameters in a prescribed way. One of the main conclusions in exercising the simulation is that varying these values in ways that might be considered reasonable from a modeling point of view for non-iid effects (such as hot-hand, back-to-the-wall or random fluctuations) doesn't dramatically alter the probabilities predicted from the pure iid assumption. While this outcome might be somewhat surprising, it's consistent with the previously cited works that indicate difficulty in detecting non-iid effects in data sets not only in tennis but in other sports. In essence, it highlights the *unreasonable effectiveness* of the iid assumption in certain situations even when it's suspected that non-iid effects are present.

Chapter 4.

Monte Carlo Simulations for *pdf*s

To build on the validated Monte Carlo tennis simulation, a stochastic Markov chain model has been developed that provides the probability density function (*pdf*) for a player to win a match. By analyzing both individual player and data for "the field" (all players lumped together) obtained from the 2007 men's Association of Tennis Professionals (ATP) circuit and beyond, it's shown that a player's probability of winning a point on serve and while receiving serve varies from match to match and that both can be modeled as Gaussian distributed random variables. Hence, this expansion of the simulation uses four input parameters for each player; an increase of three over the baseline model. The first two parameters are the sample means associated with each player's probability of winning a point on serve and while receiving serve. The third and fourth parameter for each player are the standard deviations around these means which measure a player's consistency from match to match and from one surface to another (e.g. grass, hard courts or clay). Based on these Gaussian distributed random variables, Monte Carlo simulations are used to determine the probability density functions for either player to win a match. By using input data for each of the players against the entire field, the outcome of simulations based on head-to-head matches is described starting with four top players from the men's 2007 ATP

season. Full tournament simulations are also run of the four Grand Slam events and statistics are gathered for each of these four player's frequency of winning each of the events. The stochastic model improves and corrects various deficiencies inherent in the original treatment described by the analytical model.

4.1 Background

The Markov chain model for tennis, also referred to as the analytical model, was developed by Newton and Keller in [34] and summarized in Chapter §3. As described, it's based on a single input parameter for each player like the previous models of Carter and Crews in [8] and Pollard in [39]. This parameter is the player's probability of winning a point on serve designated as p_A^R. Recall that it's the total number of points won on serve divided by total points served for a player taken over many matches against a variety of opponents. By solving the hierarchical system of recursion equations that links points to games, games to sets and sets to matches, analytical formulas were worked out for each player's probability of winning a game, set, match, and tournament under the scoring rules imposed in tennis. Because the input parameter is held constant throughout each match (interpreted as an assumption that points in tennis are independent, identically distributed random variables) the results, in a sense, model the way the scoring system in tennis effects outcomes as much as modeling individual player characteristics. More recently, Newton and Aslam [32] explored non-iid effects by analyzing ensembles of Monte Carlo simulations of tournaments, with player parameters varying from point to point based on the notion of a point's "importance" as introduced by Morris in [30] and documented in Section §3.4.1. Additionally, there have been several recent attempts at predicting outcomes of tennis matches

based on player data. For example, the work of Barnett and Clarke [4] combines player data in an attempt to predict outcomes while the earlier work of Clarke and Dyte [9] uses the official rating system as input towards a simulation of tournaments. Walker and Wooders in [49] use minimax theory from econometrics to analyze tournament data. Most recently, the work of O'Malley in [38] builds on the treatment of Newton and Keller. However, no treatment previously presented has attempted to characterize and model individual players against "the field" of opponents, thus this work presents a novel approach with solid analytical foundation that can be used to fairly compare players with little or no previous match history.

Here, the development of a *stochastic* Markov chain model is documented (with an introduction to these techniques available from Asmussen and Glynn in [3]) based on the realization that a player's probability of winning a point on serve isn't constant between matches or throughout a tournament (and hence the full season). This key parameter varies from match to match and is affected by a player's opponent. Therefore it's better modeled as a random variable whose *pdf* closely resembles a Gaussian around the sample mean. The Gaussian (normal) distribution function for a random variable X is given by

$$F_X(x) = \frac{1}{2}[1 + erf(\frac{x-\mu}{\sqrt{2\sigma^2}})] \tag{4.1}$$

with corresponding probability density function $f(x) = \frac{d}{dx}F(x)$ of

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{4.2}$$

which is characterized by its bell shaped curve. Part of the reason a player's probability of winning a point on serve varies from match to match is because it depends on his opponent's probability of winning a point when receiving serve (the two must sum to one), and return of serve ability varies from player to player. Thus, development of this model using four input parameters for each player is undertaken (also see Newton and Aslam in [33]). Again, the first two are the sample means associated with each player's probability of winning a point on serve and while receiving serve. The third and fourth parameter for each player are the standard deviations around these means which measure the player's consistency on serve and on return of serve from match to match and on the various surfaces (i.e. grass, hard courts and clay). Physical characteristics of the playing surfaces are discussed by Cross in [12]. Suffice it to say that it's widely believed (and supported by the data shown subsequently in Figure 4.2 that the fastest surface (grass) favors dominant servers (offensive play) whereas the slowest surface (clay) favors receivers and consistency (defensive play).

Using the analytical formulas obtained by solving the Markov chain system in conjunction with Monte Carlo simulations using data from "the field" it's possible to obtain the *pdf* for a player to win a point on serve which, in the parameter range of interest, has an approximately normal distribution. Targeted Monte Carlo simulations are also run for head-to-head player matches to obtain the *pdf*s for a player to win a match. A nice introduction to the properties and use of probability density functions can be found in [5] by Bendat and Piersol. Full tournament simulations have also been run using Gaussian distributed input data for 128 players tournament draws to obtain the statistical frequency of each player winning a Grand Slam event as well as their statistical frequency of winning n rounds. Then, the sample means obtained from these simulations can be

used as entries to a preference matrix (see [18] by Keener, subsequently described in full detail in Section §5.2) whose dominant eigenvector provides a ranking of the players. It's believed that this is the first stochastic player based model for tennis that has the ability to yield rankings with a natural probabilistic interpretation.

The following section describes this approach in detail. First, data obtained from the 2007 men's ATP circuit is analyzed for "the field" (all player data lumped together) for the four Grand Slam events along with four top players: Roger Federer, Rafael Nadal, Andy Roddick and James Blake. This sets the stage for viewing a player's probability of winning a point on serve or while receiving serve as (truncated) Gaussian distributed random variables. Markov chain equations are described whose solution links the Gaussian distributed inputs to the *pdf* for each player winning a match. The manner in which a player's return of serve ability is used in the model is also detailed along with the way the outputs depend on each player's standard deviation around the mean. Finally, statistics on the results of Monte Carlo simulations based on the full stochastic Markov chain model are presented. The focus is placed on individual player matchups between the four top players under consideration and the results of full Grand Slam simulations using extensive player data from the 2007 ATP season and later.

4.2 The Model

The starting point for the model are the *pdf*s for each player to win a point on serve, denoted $p_A^s(x)$ (for player A) and for each to win a point while receiving serve, denoted $p_A^r(x)$. These variables are interpreted as the player's probability as measured against "the field" of players as opposed to his probability as measured against any one particular opponent. When referring later

to player A's probability of winning a point on serve against a specific opponent, say player B, the notation $p^s_{A|B}(x)$ and $p^r_{A|B}(x)$ will be used to indicate conditional probabilities. It's shown here that the *pdf*s can be taken as truncated Gaussian distributions with support in the interval $[0,1]$ (their tails are sufficiently far from 0 and 1) that are completely characterized by their respective means, μ^s_A, μ^r_A and standard deviations σ^s_A, σ^r_A.

4.2.1 Data

For each match played on the men's ATP circuit in the 2007 season, the percentage of points won on serve for each player and the percentage of points won while receiving serve has been obtained. Figure 4.1 shows all of the combined player data, called "the field", gathered by lumping together the information for 330 players over 59 tournaments on all three surfaces (grass, clay and hard courts) over the entire season.

The *pdf*s associated with field data are denoted by $p^s_f(x)$ (*pdf* for the field to win a point on serve) and $p^r_f(x)$ (*pdf* for the field to win a point receiving serve). The corresponding means are denoted μ^s_f, μ^r_f with standard deviations denoted σ^s_f, σ^r_f. The figure shows histogrammed data for points won on serve and points won while receiving serve (scaled to have unit area) together with the associated truncated Gaussian distributions with sample mean $\mu^s_f = 0.63316$ and sample standard deviation $\sigma^s_f = 0.094779$, and $\mu^r_f = 0.36684$ and $\sigma^r_f = 0.094779$. Note that $\mu^s_f + \mu^r_f = 1$, and $\sigma^s_f = \sigma^r_f$. A chi-square goodness-of-fit test for normality was performed with a sample size $N = 5245$, resulting in $\chi^2 = 11.45$, using values $\chi^2 \leq \chi^2_{9;0.10} = 14.68$ from Table A.3 in [5]. In this case, the hypothesis of normality is accepted at the $\alpha = 0.10$ level of significance. This

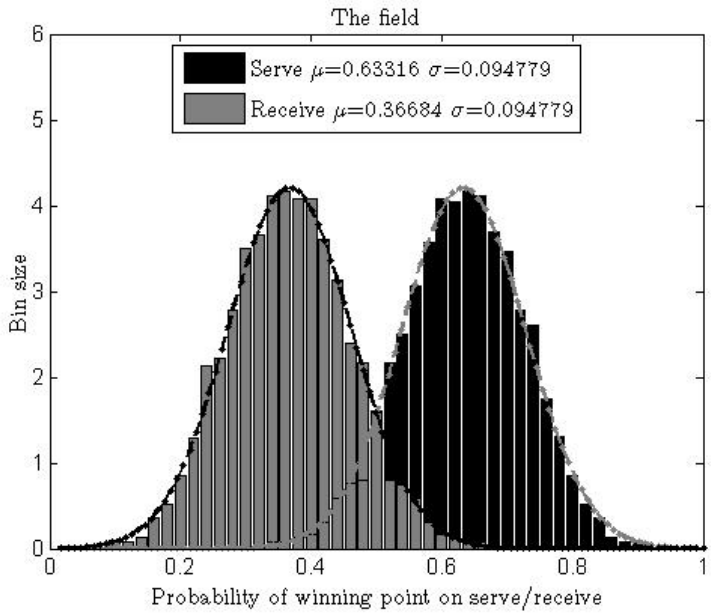


Figure 4.1. Data for "the field" in the 2007 ATP season

means that the probability density functions for "the field" can effectively be viewed as truncated normally distributed random variables, where:

$$
\begin{aligned}
\tilde{p}_f^s(x) &= (\sigma_f^s\sqrt{2\pi})^{-1}\exp\left(-\frac{(x-\mu_f^s)^2}{2(\sigma_f^s)^2}\right), \quad (-\infty < x < \infty) \\
p_f^s(x) &= C\tilde{p}_f^s(x), \quad (0 \le x \le 1), \quad \textit{otherwise} \quad 0 \\
\int_{-\infty}^{\infty} p_f^s(x)dx &= C\int_0^1 \tilde{p}_f^s(x)dx = 1, \qquad (4.3)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{p}_f^r(x) &= (\sigma_f^r\sqrt{2\pi})^{-1}\exp\left(-\frac{(x-\mu_f^r)^2}{2(\sigma_f^r)^2}\right), \quad (-\infty < x < \infty) \\
p_f^r(x) &= C\tilde{p}_f^r(x), \quad (0 \le x \le 1), \quad \textit{otherwise} \quad 0 \\
\int_{-\infty}^{\infty} p_f^r(x)dx &= C\int_0^1 \tilde{p}_f^r(x)dx = 1. \qquad (4.4)
\end{aligned}
$$

The four defining parameters for the probability density functions associated with the field are $\mu_f^s, \mu_f^r, \sigma_f^s$ and σ_f^r.

Breaking down the data further, field data for each of the four Grand Slam tournaments is shown in Figure 4.2 in the order in which they're played. Note that the histograms in this figure now only reflect sample size $N = 127$, which is only about 2.5% of the data used to construct Figure 4.1. Figure 4.2(a) shows data from the 2007 Australian Open (hard courts) histogrammed together with the Gaussian fit. Figure 4.2(b) shows data from the 2007 French Open (clay), Figure 4.2(c) shows data from the 2007 Wimbledon (grass) and Figure 4.2(d) shows data from the 2007 US Open (hard courts).

Because the data for each tournament are much more sparse than for the full season, the histograms in Figure 4.2 aren't as filled out as those in Figure 4.1, yet the Gaussian distributions still model the density functions quite accurately. This graphically illustrates the challenges associated with extracting meaningful information from smaller data sets. For Figures 4.2(a)-(d), using the chi-square goodness-of-fit test for normality, acceptance is again found at the $\alpha = 0.10$ level of significance. These data provide the values and baseline confidence level for $\mu_f^s, \mu_f^r, \sigma_f^s$ and σ_f^r. It should be emphasized that this presentation of data is itself a breakthrough in the realization that,

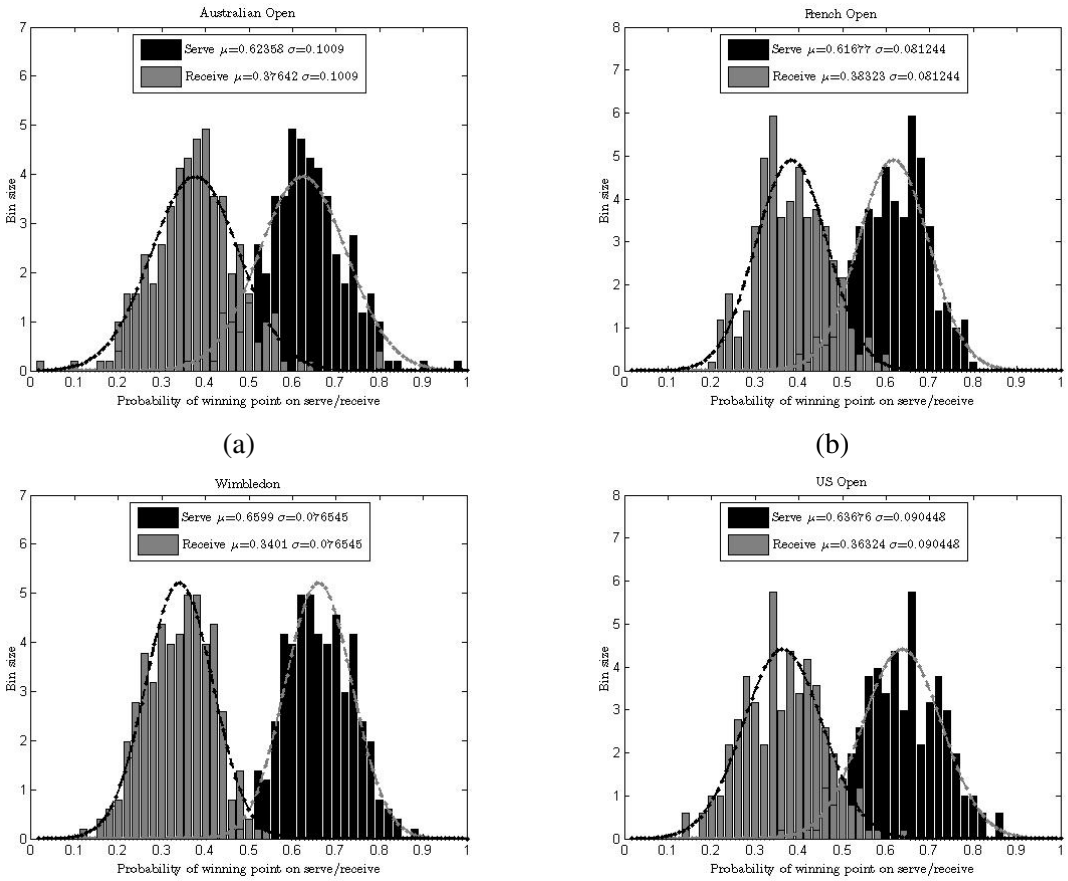


Figure 4.2. Histograms and Gaussian fit data for the field in each of the four Grand Slam tournaments of the 2007 season data using sample means and standard deviations

from large scale to small scale (i.e. season data to tournament data to player data) the modeling of player performance can be accurately described as being normally distributed. Specifically, it wouldn't be possible to reach this conclusion by analyzing individual player data alone.

The data for the four Grand Slam events are summarized in Table 4.1. The sample mean for points won on serve (receiving serve) are highest (lowest) on grass which is the fastest surface and favors players with dominant serves. The opposite is true for the slowest surface (clay). These statistical conclusions corroborate the physical characteristics of the respective surfaces with respect to the way the ball bounces as discussed by Cross in [12].

Event	μ^s	σ^s	μ^r	σ^r
Australian Open (hard)	0.62358	0.100900	0.37642	0.100900
French Open (clay)	0.61677	0.081244	0.38323	0.081244
Wimbledon (grass)	0.65990	0.076545	0.34010	0.076545
US Open (hard)	0.63676	0.090448	0.36324	0.090448

Table 4.1. Summary of field data broken down for each of the four 2007 Grand Slam events

This provides further merit to the use of the fundamental, four parameter modeling approach used by stochastic Markov chain Monte Carlo simulation. Figure 4.3 shows individual player data taken over the full 2007 season for four top players of interest: Roger Federer (#1 2007 year end ranking), Rafael Nadal (#2 2007 year end ranking), Andy Roddick (#6 2007 year end ranking) and James Blake (#13 2007 year end ranking). The notation is to underscore the variable using the initials of the player, hence Roger Federer's mean value for points won on serve is denoted μ^s_{RF} (as measured against "the field"). Once again, although the data are much more sparse than that used in Figure 4.1, it's still concluded that the histograms would fill out Gaussian distributions if more data were available. For Figure 4.3, a chi-square goodness-of-fit test for normality was

performed and it was found that the hypothesis of normality is accepted at the $\alpha = 0.10$ level of significance. For Figure 4.3(b) (receive) and Figure 4.3(d) (serve), acceptance is at the $\alpha = 0.05$ level of significance, while for Figure 4.3(d) (receive) acceptance is at the $\alpha = 0.01$ level of significance.

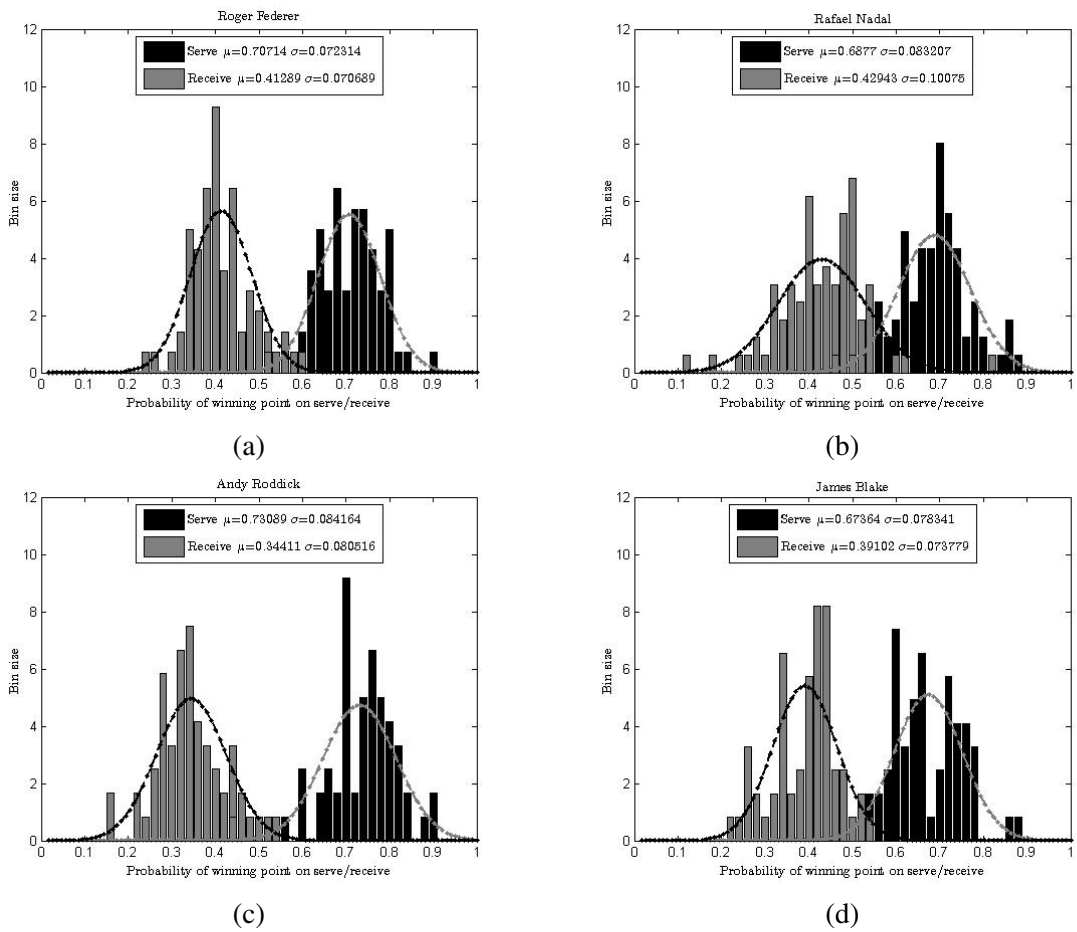


Figure 4.3. Histograms and Gaussian distributions of individual player data on serve and receive of serve for full 2007 season

4.2.2 Probability Density Functions

In the context of the current model, the probability that player A wins a point on serve, p_A^s, should be interpreted as the sample mean μ_A^s associated with the truncated Gaussian distribution for winning points on serve, while p_A^G should be interpreted as the sample mean associated with the player's probability of winning a game on serve. The two are related to each other via equation (3.1) which is plotted in Figure 4.4 together with the tangent line approximation to the curve at the sample mean for the field $p_A = 0.50$.

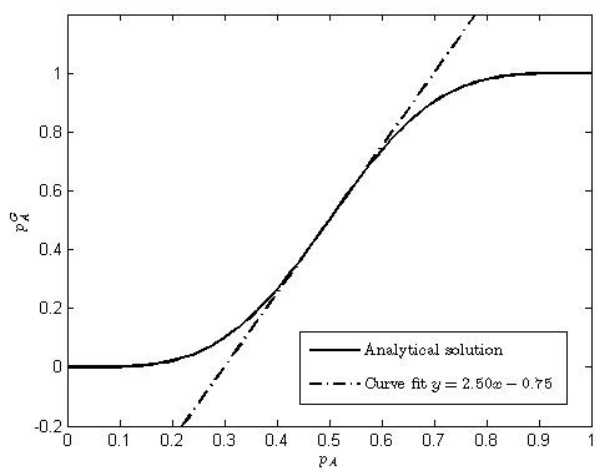


Figure 4.4. p_A^G vs p_A^s as given in equation (3.1) which links points to games, together with the tangent line approximation taken at $p_A^S = 0.5$

Note that in this region the graph is highly linear. Because of this, the probability density function governing games won on serve should be linearly related (i.e. proportional) to the *pdf* governing points won on serve (in the approximate range $0.3 < p_A^s < 0.7$). This means that since $p_A^s(x)$ is taken to be Gaussian distributed, so must $p_A^G(x)$. To obtain the *pdf* $p_A^G(x)$, Monte Carlo simulations were run to execute an ensemble of $10,000$ matches between two players with

Gaussian distributed input density functions $p_A^s(x)$ and $p_B^s(x)$. Figure 4.5 and Figure 4.6 show the results of these simulations. Figure 4.5 is a typical convergence plot (log-log scale) using inputs $\mu_A^s = 0.63, \sigma_A^s = 0$ for player A and $\mu_B^s = 0.65, \sigma_B^s = 0.06$ for player B.

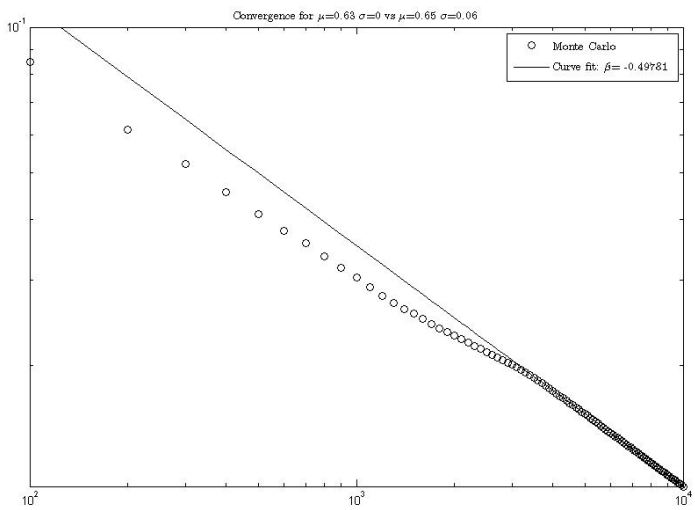
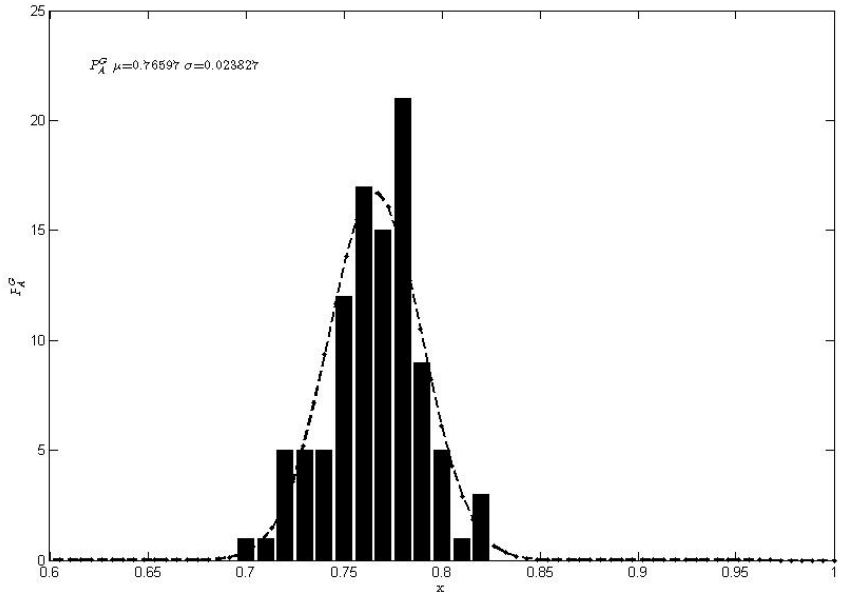


Figure 4.5. Convergence plot (log-log) showing μ_A^G vs. N with input parameters $\mu_A = 0.63$, $\sigma_A = 0$, $\mu_B = 0.65$, $\sigma_B = 0.06$

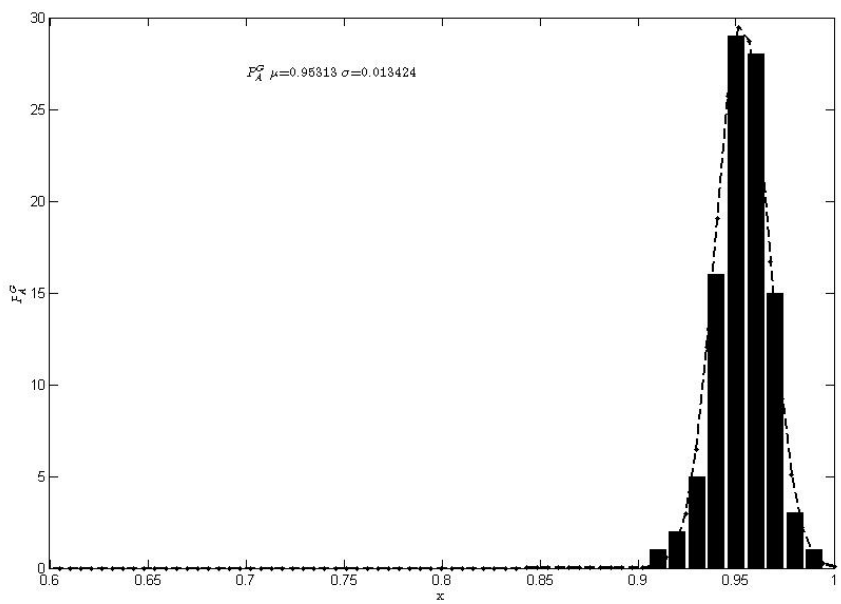
As is typical for all parameter values that were run, convergence to the sample mean has power-law form where

$$\mu_A^G(N) - \mu_A^G(\infty) \sim N^{-\beta}$$

with $\beta \approx 0.49781$. Figure 4.6 shows the *pdf* $p_A^G(x)$ for a player to win a game, using input values for $p_A^s(x)$ of $\mu_A^s = 0.63316$, with $\sigma_A = 0.094779$ in Figure 4.6(a) and $\mu_A^s = 0.8$, with $\sigma_A = 0.094779$ in Figure 4.6(b).



(a)



(b)

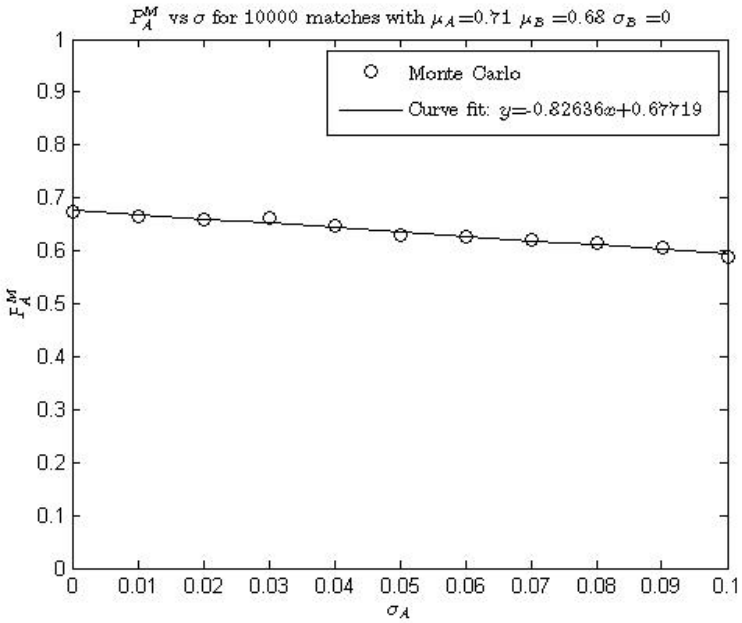
Figure 4.6. Probability density function $p_A^G(x)$ obtained from Monte Carlo simulations of $300,000$ games for $\mu_A^S = 0.63316$ and $\mu_A^S = 0.8$ with $\sigma_A^S = 0.094779$

The mean values in going from points to games shifts as predicted, by the curve shown in Figure 4.6. In particular, the case $\mu_A^s = 0.63316 = p_A^s$ results in $\mu_A^G = 0.76597 = p_A^G$ with $\sigma_A^G = 0.023827$, while the case $\mu_A^s = 0.8 = p_A^s$ results in $\mu_A^G = 0.95313 = p_A^G$ with $\sigma_A^G = 0.013424$. The main conclusion is that the distributions for $p_A^G(x)$ remain (approximately) normally distributed throughout a wide range of values and, specifically, the range of interest.

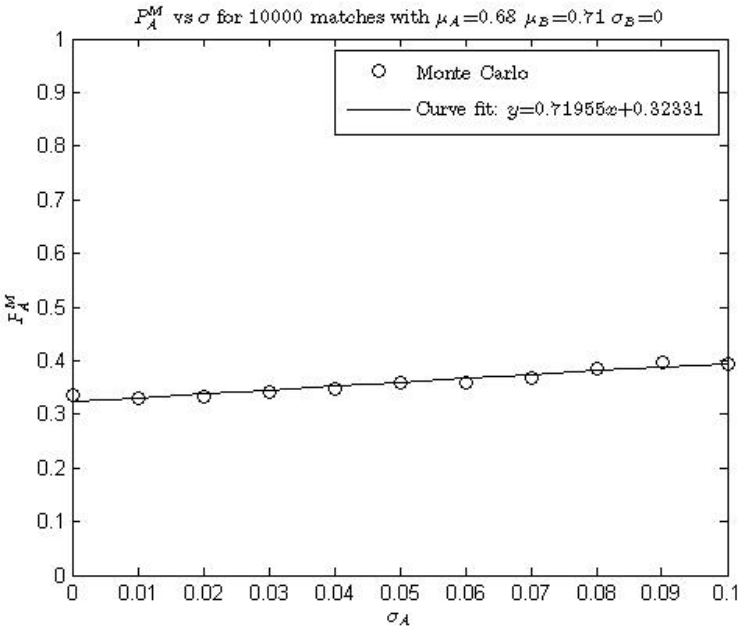
4.2.3 Consistency

A player's consistency from match to match is measured by his standard deviations on points won on serve and return of serve. To see the effect of varying this parameter, refer to Figure 4.7. In Figure 4.7(a), player A's mean value for points won on serve is taken to be $\mu_A^s = 0.71$, while his standard deviation varies $0 \leq \sigma_A \leq 0.1$. Player B is the weaker server, with mean value for points won on serve taken as $\mu_B^s = 0.68$. His standard deviation is taken to be $\sigma_B = 0$, hence he's more consistent.

The figure shows that as the standard deviation for player A increases, his probability of winning the match decreases. Thus, for the stronger server (i.e. one with higher mean value), lack of consistency (higher standard deviation) hurts his chances of winning the match. In Figure 4.7(b), parameter values are $\mu_A^s = 0.68$ for player A as his standard deviation is varied $0 \leq \sigma_A \leq 0.1$. For player B, $\mu_B^s = 0.71$ with $\sigma_B = 0$. Here, as player A's standard deviation increases so do his chances of winning the match. Interestingly, for the weaker server (the one with lower mean value for points won on serve), lack of consistency actually *increases* his chance of winning the match. Regardless, a player's standard deviation for points won on serve certainly affects his chances of winning or losing a match.



(a)



(b)

Figure 4.7. The effect of varying a player's standard deviation as a measure of his consistency from match to match

A second important and interesting point regarding a player's standard deviation can be seen in Figure 4.6. The standard deviations associated with the *pdf*s at the point level *decrease* significantly at the game level. Thus, a player's lack of consistency in winning points on serve, as modeled by his standard deviation in $p_A^s(x)$, becomes less significant in the *pdf* $p_A^G(x)$. Finally, the connection between the size of the variance of an outcome and the length of the contest must be emphasized. It's been documented (both by Newton and Keller in [34] and more recently by O'Malley in [38]) that upsets occur less often in best out of five set contests than in best out of three set contests. Shorter contests have higher variability and thus more chance for upsets. This is also true in other sports where upsets are more frequent in individual games than in playoff series. For other recent studies of the effects of stochastic variances in paired comparison models refer to the work of Glickman in [15] and [16].

4.2.4 Implementation of the Full Model

Implementation of the full model is now described. First, a description is provided on how each player's Gaussian distributed probability of winning a point on receiving serve is used in the context of the Markov chain model. In a match between player A and player B, A either wins a point on serve (with probability $p_{A|B}^s$) or he loses a point on serve in which case his opponent wins the point on return of serve (with probability $p_{B|A}^r$). Therefore:

$$\begin{aligned} p_{A|B}^s + p_{B|A}^r &= 1, \\ p_{A|B}^r + p_{B|A}^s &= 1. \end{aligned}$$

Also, since "the field" effectively plays itself:

$$p_f^s + p_f^r = 1. \tag{4.5}$$

From these equations it's clear that there's a balance between a player's probability of winning a point on serve and his opponents probability of winning a point on return of serve. Specifically, an increase in one comes at the expense of the other. When running a Monte Carlo simulation in a head-to-head matchup between players A and B, direct head-to-head match data isn't used (as typically there isn't enough available) and instead the inputs are obtained from their results measured against "the field". For this:

$$p_{A|f}^s + p_{f|A}^r = 1, \tag{4.6}$$

$$p_{A|f}^r + p_{f|A}^s = 1, \tag{4.7}$$

and

$$p_{B|f}^s + p_{f|B}^r = 1, \tag{4.8}$$

$$p_{B|f}^r + p_{f|B}^s = 1. \tag{4.9}$$

Subtracting (4.9) from (4.6) and (4.7) from (4.8) and re-arranging gives:

$$p^s_{A|f} - p^r_{B|f} = p^s_{f|B} - p^r_{f|A}, \tag{4.10}$$

$$p^s_{B|f} - p^r_{A|f} = p^s_{f|A} - p^r_{f|B}. \tag{4.11}$$

Then, adding p^r_f to the left hand side and $1 - p^s_f$ to the right hand side of both (4.10) and (4.11) using the relationship in (4.5) gives

$$p^s_{A|f} + (p^r_f - p^r_{B|f}) \equiv p^{s,eff}_{A|B} = 1 - [p^r_{f|A} + (p^s_f - p^s_{f|B})] \equiv 1 - p^{r,eff}_{B|A}, \tag{4.12}$$

$$p^s_{B|f} + (p^r_f - p^r_{A|f}) \equiv p^{s,eff}_{B|A} = 1 - [p^r_{f|B} + (p^s_f - p^s_{f|A})] \equiv 1 - p^{r,eff}_{A|B}. \tag{4.13}$$

The second terms on the left, $(p^r_f - p^r_{B|f})$ and $(p^r_f - p^r_{A|f})$ are called "field adjusted" variables. They measure the deviation of a player's return of serve ability from that of the field. The 'tilde' notation is used to denote field adjusted variables. These are listed in (4.12) and (4.13) as

$$\tilde{p}^r_{A|f} \equiv p^r_f - p^r_{A|f}$$

$$\tilde{p}^r_{B|f} \equiv p^r_f - p^r_{B|f}.$$

Thus, eqns (4.12) and (4.13) become simply

$$p^{s,eff}_{A|B} = p^{s}_{A|f} + \tilde{p}^{r}_{B|f}, \tag{4.14}$$

$$p^{s,eff}_{B|A} = p^{s}_{B|f} + \tilde{p}^{r}_{A|f}. \tag{4.15}$$

In a head-to-head simulation between players A and B, eqn (4.14) indicates that to include the influence of player B's ability to receive serve, as measured by the field adjusted term $\tilde{p}^{r}_{B|f}$, simply adjust the serve parameter for player A, $p^{s}_{A|f}$. An increase in an opponent's return of serve performance comes at the expense of a player's ability on his serve. Likewise, to include the influence of player A's ability to receive serve, as measured by the field adjusted term $\tilde{p}^{r}_{A|f}$, adjust the serve parameter for player B, $p^{s}_{B|f}$. *In short, a player's ability to win a point on return of serve is accounted for by adjusting his opponent's probability of winning a point on serve, either up or down, depending on whether the return of serve ability is better or worse than that of the field.*

As an example, in running a simulation between Roger Federer and Rafael Nadal in 2007, Federer's sample mean (measured against "the field") for points won on serve is $\mu^{s}_{RF} = 0.70714$ while his sample mean for points won on receive of serve is $\mu^{r}_{RF} = 0.41289$. For Rafael Nadal, those values are $\mu^{s}_{RN} = 0.68770$ and $\mu^{r}_{RN} = 0.42943$. The field mean for return of serve is $\mu^{r}_{f} = 0.36684$. In practice, these $\mu_{A|f}$ and $\sigma_{A|f}$ parameters would be used to build the appropriate Gaussian distributions that describe the *pdf*s for $p^{s}_{A|f}$ and $p^{r}_{A|f}$. However, here $\mu_{A|f}$ is used directly as we take all σ parameters to be zero in this example. Next, the difference between each player's

return of serve value with that of "the field" is calculated (i.e. the field adjusted value) which is denoted with a 'tilde' as:

$$\begin{aligned}
\tilde{\mu}^r_{RF} \equiv \mu^r_f - \mu^r_{RF} &= 0.36684 - 0.41289 = -0.04605 \\
\tilde{\mu}^r_{RN} \equiv \mu^r_f - \mu^r_{RN} &= 0.36684 - 0.42943 = -0.06259.
\end{aligned}$$

Note that both Federer and Nadal have above average return of serve ability with respect to the field. Therefore, they effectively reduce each other's ability to win a point on serve as given by (4.14) and (4.15) and shown explicitly here:

$$\begin{aligned}
p^s_{RF|RN} \equiv \mu^s_{RF} + \tilde{\mu}^r_{RN} &= 0.70714 - 0.06259 = 0.64455 \\
p^s_{RN|RF} \equiv \mu^s_{RN} + \tilde{\mu}^r_{RF} &= 0.68770 - 0.04605 = 0.64165.
\end{aligned}$$

These values, $p^s_{RF|RN}$ and $p^s_{RN|RF}$, are the ones specifically used as inputs to the Markov chain model governing each player's probability winning a game, set and match on serve. Thus, each player's probability of winning a point on serve is adjusted depending on the strength of his opponent's return of serve ability. A player will win fewer points on serve playing against an opponent with a stronger return of serve than one with a weaker return of serve. In this way, return of serve ability, while playing an important role in this model, manifests itself through adjustments to each player's probability of winning a point on his own serve.

Based on the concept of using these player adjusted values as inputs to the truncated Gaussian distributions, one realization of a statistical simulation between player A and player B proceeds as follows:

1. Obtain $p^s_{A|f}$, $p^s_{B|f}$, $p^r_{A|f}$, $p^r_{B|f}$ by drawing random values with probability density function given by the truncated Gaussian distribution appropriate to each player using $\mu^s_{A|f}$ and $\sigma^s_{A|f}$ to obtain $p^s_{A|f}$ and using $\mu^r_{A|f}$ and $\sigma^r_{A|f}$ to obtain $p^r_{A|f}$.

2. Calculate $p^s_{A|B}$ and $p^s_{B|A}$ using (4.14) and (4.15) for use in the Markov chain formulas described in Section §3.1.

3. Calculate each player's probability of winning a game, set and match by solving the Markov chain formulas to obtain $p^M_{A|B}$, and $p^M_{B|A}$.

4. Repeat these two steps thousands of times (choosing a new random value for each player, for each simulation) to obtain a statistical ensemble from which the probability density functions are obtained. Sample means and standard deviations are then calculated for each of the ensembles.

Results from head-to-head simulations between players and full tournament simulations are now fully explored using this methodology.

4.2.5 Simulations

The results of Monte Carlo simulations of head-to-head matches between Roger Federer, Rafael Nadal, Andy Roddick, and James Blake are described using their player data from the 2007 ATP season as input as shown in Figure 4.3 and summarized in Table 4.2. The outcome of

full tournament simulations of each of the four Grand Slam events is also described using two different methods for obtaining ensembles. Year end ranking is shown in parenthesis for these four players in Table 4.2.

Player	μ^s	σ^s	μ^r	σ^r
J. Blake (13)	0.67364	0.078341	0.39102	0.073779
R. Federer (1)	0.70714	0.072314	0.41289	0.070689
R. Nadal (2)	0.68770	0.083207	0.42943	0.100750
A. Roddick (6)	0.73089	0.084164	0.34411	0.080516

Table 4.2. Player data compiled for R. Federer, R. Nadal, A. Roddick and J. Blake for the full 2007 ATP season

Note that Roddick has the highest percentage of points won on serve, while Nadal has the highest percentage of points won on receive of serve, yet neither ended that season with the top ranking.

4.2.6 Head-To-Head Matches

As shown in Table 4.2, Federer doesn't have the highest mean value for points won on serve or the highest mean value for points won on receive of serve, yet he finished the 2007 season with the best record and the top ranking. It should be pointed out that his consistency (standard deviation) both on serve and receiving serve was the lowest of the four players. These parameters were used in Monte Carlo simulations of $30,000$ matches between each pair of the four players. The results are histogrammed in Figure 4.8 and the sample means and standard deviations are used to define the probability density functions for each player's probability of winning a match. The output values are shown in Table 4.3.

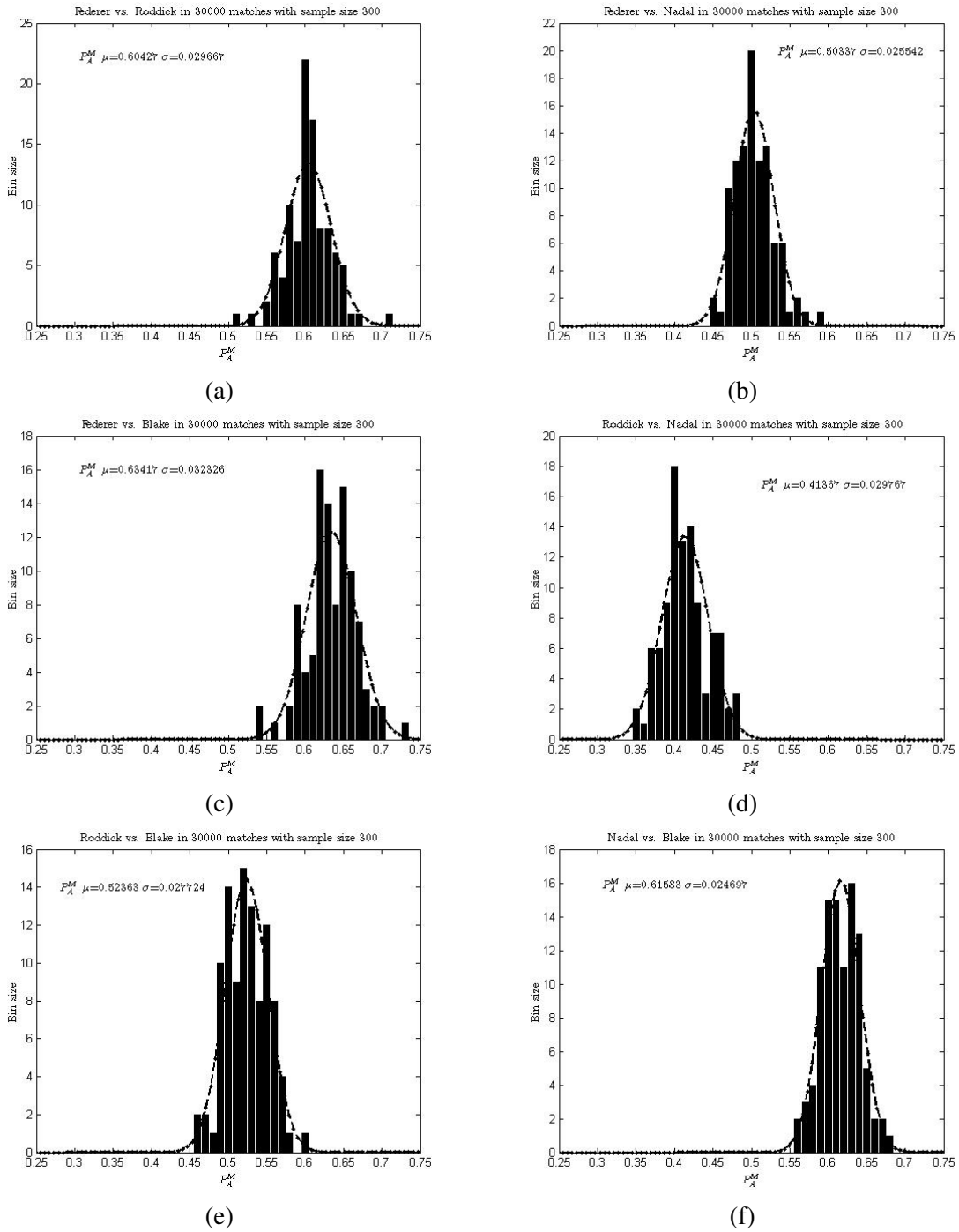


Figure 4.8. Histogram and Gaussian distribution of head-to-head matchups between R. Federer, R. Nadal, A. Roddick and J. Blake based on 30,000 simulated matches using the four parameter Monte Carlo model

	Blake	Federer	Nadal	Roddick
Blake	NA	$\mu = 0.3635$, $\sigma = 0.03043$	$\mu = 0.3869$, $\sigma = 0.02925$	$\mu = 0.4794$, $\sigma = 0.02761$
Federer	$\mu = 0.6365$, $\sigma = 0.03043$	NA	$\mu = 0.5074$, $\sigma = 0.02760$	$\mu = 0.6073$, $\sigma = 0.02780$
Nadal	$\mu = 0.6131$, $\sigma = 0.02925$	$\mu = 0.4926$, $\sigma = 0.02760$	NA	$\mu = 0.5883$, $\sigma = 0.02790$
Roddick	$\mu = 0.5206$, $\sigma = 0.02761$	$\mu = 0.3927$, $\sigma = 0.02780$	$\mu = 0.4117$, $\sigma = 0.02790$	NA

Table 4.3. Resulting means and standard deviations from the Monte Carlo simulations of head-to-head match ups

The results yield important information about how close the players are to each other in overall performance. For example, in head-to-head matches between Federer and Nadal, Federer is predicted to win a slim majority of 50.74% of their matches.

Since the head-to-head simulation method relies on data for each of the players taken against "the field", one might wonder whether using actual head-to-head data might be useful as input to a model instead. Remember that in any given year the number of actual head-to-head matches between any two players is quite small, making it very difficult to use as a basis for statistical purposes. For example, Federer defeated Nadal in three out of their five matches in 2007, consistent with his slight statistical edge shown in Table 4.3. Beyond that, however, there aren't enough matches to gather meaningful statistical information. Federer also defeated Blake in their single 2007 match and he defeated Roddick in all three of their head-to-head encounters. Nadal defeated Roddick in their single 2007 match, while Nadal never played Blake, and Roddick and Blake also never played. All of these outcomes are consistent with the statistical findings from these simulations, however the sparsity of head-to-head meetings between any two players on the tour in a given year makes it meaningless to use these data alone as input to a model.

Taking data from head-to-head encounters over several years or a full career would typically increase the amount of data available, but it would introduce other troublesome problems. It's doubtful that data from matches between Federer and Nadal in 2005 or even 2006 would be useful in directly predicting outcomes in 2007 as Nadal was in the process of rapidly improving (and altering) his game at that time. Even between two players whose career timelines match, such as Agassi and Sampras, it's hard to argue that data taken from the early stages of their career head-to-head encounters would help in predicting who would win their final encounter at the US Open in 2002. It must also be recognized that the use of individual player data taken against the field in the evaluation of head-to-head encounters possibly underestimates the significance of one player having a statistically unusually high (or low) success rate against another individual player (a 'bogey' opponent).

4.2.7 Tournament Simulations

Using the full set of 2007 tournament data for each of the 128 players in the four Grand Slam events, Monte Carlo simulations were carried out for each tournament draw. The ensembles were gathered in two ways. First, 1,000 fictitious tournaments were run, initializing each realization by using the actual first round matchups from the event. These are called "fixed" or "actual" draws. Then, for comparison, 1,000 simulations were run of each event using random draws chosen in the first round of each realization. The comparison of the two provides valuable insight into the effect of the actual tournament (i.e. player seedings) draw on outcomes. The results from the simulations were histogrammed for each tournament showing the number of rounds won by each player in the ensemble of 1000 simulated tournaments. Table 4.4 shows the number of tournament

wins for each of the four players under consideration out of 1,000 tournament simulations using fixed draws. Table 4.5 shows the same results using random draws.

Player	Australian Open	French Open	Wimbledon	US Open
J. Blake	27	22	28	20
R. Federer	115	120	122	94
R. Nadal	79	93	75	105
A. Roddick	27	24	39	38

Table 4.4. Number of tournament wins in 1,000 tournament simulations for each of the Grand Slam events using the actual draw in the first round

Player	Australian Open	French Open	Wimbledon	US Open
J. Blake	25	23	23	20
R. Federer	108	95	98	96
R. Nadal	70	90	67	80
A. Roddick	30	28	26	36

Table 4.5. Number of tournament wins in 1,000 tournament simulations for each of the Grand Slam events using random draws in the first round

The full histograms showing round-by-round statistics on the four players for Wimbledon and the US Open are shown in Figures 4.9 - 4.12.

A number of points are worth making. Generally speaking, the stronger players do better in the actual draw than in random draws. This is evidenced by the fact that the bins from the histograms for Federer and Nadal for the random draws decrease in height as the rounds increase (except for the finals) while those from the actual draws don't. These two stronger players have a better chance of surviving deep into the tournament with the actual draw as designed by the seeding committee. They also win more of the tournaments (summarized in Table 4.4 and Table 4.5) in the actual draw than in the random draws. This isn't true of weaker players who fair

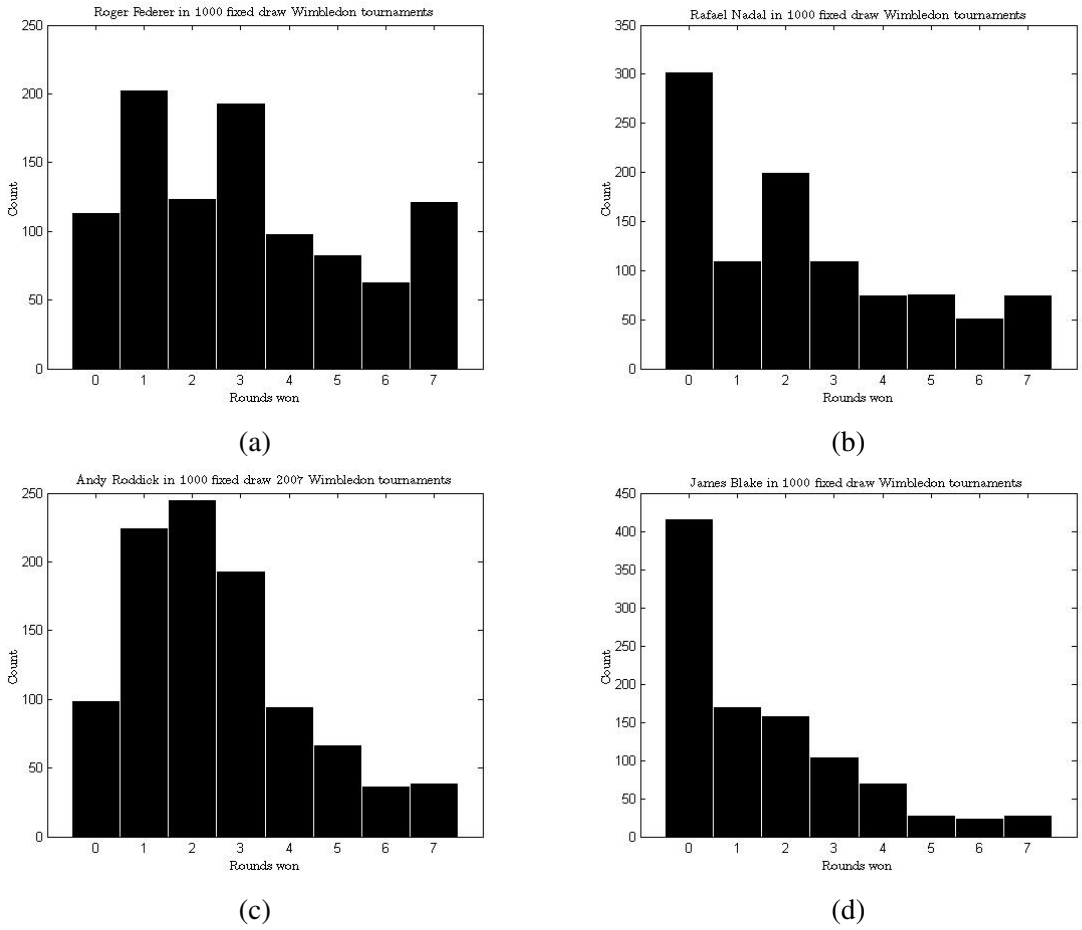


Figure 4.9. Histogram of 1,000 full tournament simulations for the 2007 Wimbledon tournament using the actual draw

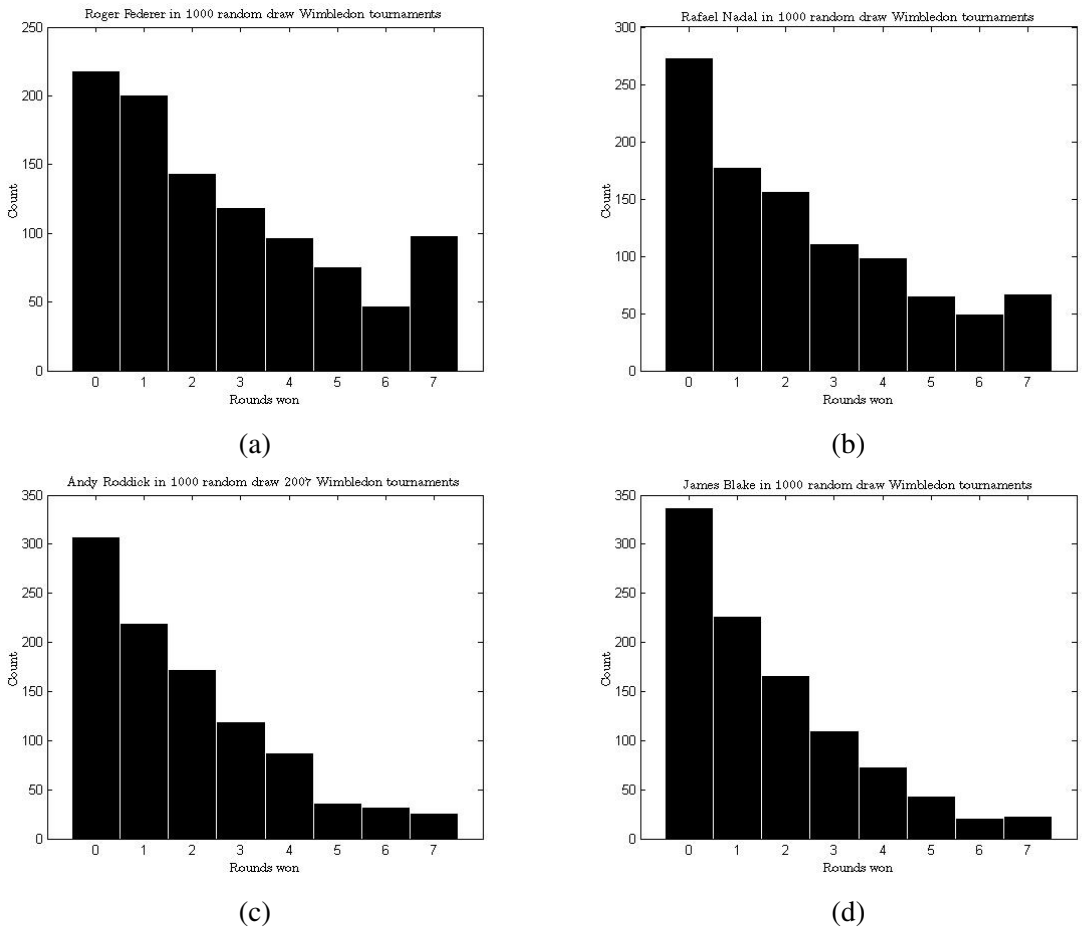


Figure 4.10. Histogram of 1,000 full tournament simulations for the 2007 Wimbledon tournament using a randomized draw

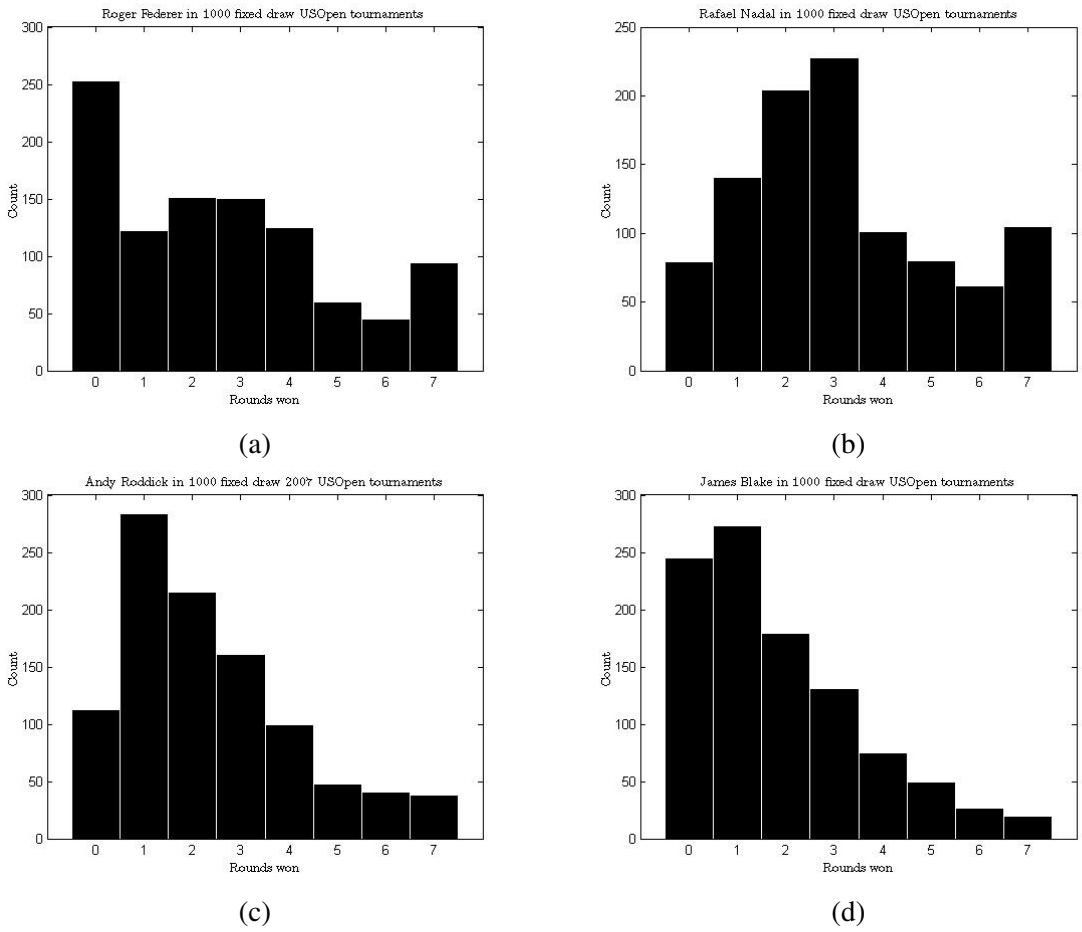


Figure 4.11. Histogram of 1,000 full tournament simulations for the 2007 US Open tournament using the actual draw

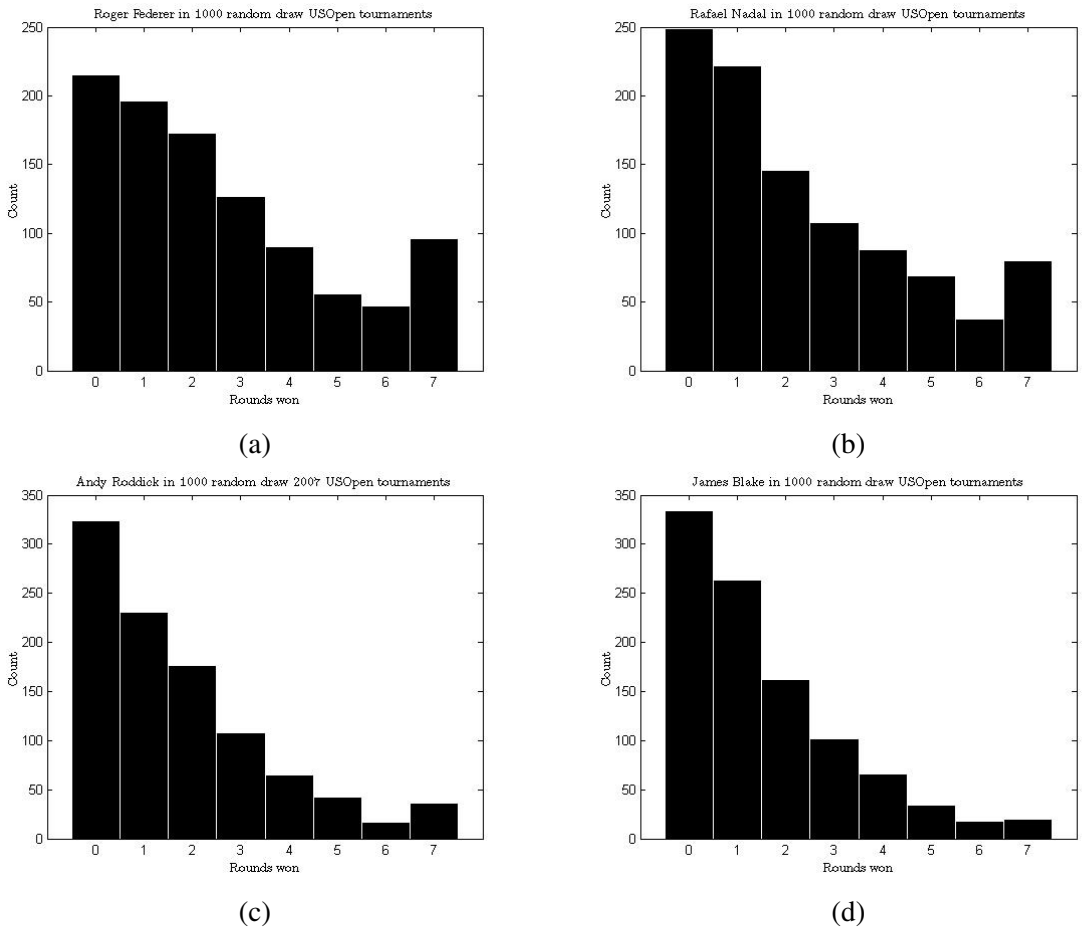


Figure 4.12. Histogram of 1,000 full tournament simulations for the 2007 US Open tournament using a randomized draw

better in a random draw than in the actual draw where they're forced to play top players in the early rounds. Again, note that the incidence of rounds won strictly decreases as the tournament progresses until the last and final round. At first this may seem puzzling, but consider the fact that when a player gets to the final round (round 7 for the 128-player draw as in this case), the last match is effectivly played against the remainder of "the field" and therefore players that are above the field average in performance will win more final round matches than they lose. Conversely, players that are below the field average in performance will lose more final round matches than they win. Since the four players under consideration in Figure 4.10 and Figure 4.12 are above average with respect to the field, the number of final rounds won for the random draw scenarios is actually an *increase* over the number of rounds won for the previous round.

When comparing the number of times each of the four players actually won a Grand Slam event in the simulations, Federer was the most successful winning 115, 120, 122 and 94 of each of the events which in probabilistic terms translates into winning 11.5% of the Australian Open simulations, 12% of the French Open simulations, 12.2% of the Wimbledon simulations and 9.4% of the US Open simulations. Although these numbers were higher than any other player (except for Nadal's outcomes in the US Open simulations which showed he won 10.5% of those events) they're perhaps surprisingly low given the fact that Federer won three out of four of the actual Grand Slam events in the 2007 calendar year (Australian Open, Wimbledon and US Open) and made it to the finals of the French Open. It's worthwhile pointing out that the statistics from Table 4.4 and Table 4.5 support the notion that Federer has a larger advantage over Nadal in Wimbledon than in the French Open based on their respective styles of play and the two different surfaces these tournaments employ.

4.3 Discussion

Several novel features in employing the stochastic Markov chain model for developing *pdf*s have been described. Most notable is the use of data based Gaussian distributed input variables into the model measuring each player's (i) strength of serve; (ii) strength of return of serve; and (iii) consistency. The data are gathered for each player over their entire portfolio of matches played in the 2007 ATP season and beyond. Using data for "the field" allows enough information to be incorporated on the performance of each of the individual players despite the fact that most pairs of players have never physically played a head-to-head match (or at most, very few) in a given year. Thus, predictions on the probability that one player will defeat another are based on how each has performed against the same control group which in this case is the entire field of players. The calculation of probability density functions as the main output of the model, as opposed to single scalar outputs gives far more detailed information regarding a player's ability and probability of winning a match against any other player in the field. It also provides the ability to carry out realistic tournament simulations, gathering statistics for each player on a round-by-round basis. There are several ways of using this information in the development of probabilistic ranking schemes. Specifically, fully developed Monte Carlo based ranking systems that utilize the *pdf* based player models are described in Chapter §6. First, a detailed background on ranking systems is provided to fully characterize the approaches available and their associated strengths and weaknesses.

Chapter 5.

Current Ranking Systems

Several ranking systems have been developed in the recent past, most notably for NCAA Division I college football. Many of these ranking systems have gained significant notoriety due to the inception of the Bowl Championship Series (BCS) at the beginning of the 1998 season. The BCS was established to determine a national champion within the bowl based system already used in college football. Due to the high profile of the sport and its associated polls and rankings, there exists adequate literature and documentation on ranking theory for technical analysis and comparison to the tennis application. In particular, the systems comprising the BCS have been subject to sometimes harsh and pointed criticism. In addition, NCAA Division I college football is one of the only major sports that crowns its champion without some form of a multigame playoff or tournament format as noted specifically by Callahan, Mucha and Porter in [7]. Beyond college football, other ranking formats exist that provide quality information on ideal traits and desirable properties. For these reasons, an overview on such ranking systems is provided so that a solid foundation for additional use of the stochastic Markov chain *pdf* player model can be approached with confidence in similar applications.

5.1 Elo Rating

In an attempt to improve chess rating methodology circa 1960, Arpad Elo created what is now known as the Elo Rating system. The system, as described by Elo in [14], ultimately calculates relative skill levels based on the win/loss outcome of a two player or two team event. For chess, scaled ratings, R, have been arbitrarily selected such that a difference of 200 points translates to an expected score of approximately 0.75. A player is awarded the actual score of simply 1 for a win, $\frac{1}{2}$ for a draw and 0 for a loss. The expected score, however, is defined as his probability of winning plus half his probability of drawing. Using the logistic curve (equation) as a baseline and the 200 point default spread, the expected scores of two players A and B are defined

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{5.1}$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}} \tag{5.2}$$

where $E_A + E_B = 1$. A player's new rating can then be determined using actual score, expected score and scaling factor (generally called K-value and set at 16 for masters level and 32 for weaker players in chess).

$$R'_A = R_A + K(S_A - E_A) \tag{5.3}$$

An advantage of this method is that players can calculate their own updated ranking with knowledge of only their own and their opponent's rating. For example, a player with rating 1500

loses a match to an opponent with rating 1600. Using (5.1) and (5.3) and a scaling factor of 32, the expected score E_A is 0.3599, the actual score S_A is 0 and the updated rating goes from 1500 to 1489 for player A and from 1600 to 1612 for player B. This highlights the particular attention that must be given to developing the average or starting rating for new players/teams along with the inflation/deflation of the ratings between various paired sets. Modern applications of Elo ratings generally use less arbitrary scaling factors and adjustments, and the resulting systems have been found in a variety of applications from computer games to team sports. A key weakness of the Elo rating system is the aforementioned rating inflation. An abnormally good (bad) player being rated in such a system can continue to build (lose) rating score without bound. Because of this rating "creep", the Elo system is considered unsuited for use in a system where probabilistic rankings are desired. Additionally, the ability to calculate a future rating based on the outcome of a single event can be seen to exclude the impact of rating effects of other participants in the field.

5.2 Matrix Based Rankings for Uneven Paired Competitions

5.2.1 A Direct Method

One of the earlier matrix based ranking systems was developed by Keener in [18]. His motivation was the lack of objectivity observed for the voting in the college football ranking polls and the general lack of understanding for some of the preemptive mathematically based ranking models. An outcome (or preference) matrix is established that assigns values (in an arguably ad-hoc fashion at times) to teams based on their interactions. Start by assuming there's a vector of rankings, $\vec{r}$, consisting of positive rankings r_j. A score is defined for team i as

$$s_i = \frac{1}{n_i} \sum_{j=1}^{N} a_{ij} r_j \tag{5.4}$$

where a_{ij} are the elements from the so called outcome (preference) matrix and N is the total number of teams with n_i as the number of games played by team i. A common and trivial way to define a_{ij} is to assign a value of 1 for a win, $\frac{1}{2}$ for a tie and 0 for a loss in each competition. The concept of establishing this preference matrix isn't new. Kendall and Smith documented a similar type of matrix methodology as far back as 1939 in [19]. It had been previously used in similar evaluations for paired comparisons. Paired comparisons consist of several competing elements that are evaluated over a series of equal matchups. An example of a paired matchup would be a round robin tournament like the preliminary round of a World Cup soccer tournament or the ATP World Tour Finals played at the end of each ATP tournament season. Note that the matchups in a college football season and in an ATP tennis season result in *unpaired* competitions.

A proposition is made that the rank of a particular team should be proportional to its score such that

$$A\vec{r} = \lambda\vec{r} \tag{5.5}$$

where the ranking vector, $\vec{r}$, is a positive eigenvector of the outcome matrix A. Here, the outcome matrix A is simply normalized from $\frac{a_{ij}}{n_i}$. The Perron-Frobenius Theorem describes the situation for which a solution exists to this eigenvalue problem.

PERRON-FROBENIUS THEOREM. *If the nontrivial matrix A has nonnegative entries then there exists an eigenvector $\vec{r}$ with nonnegative entries, corresponding to a*

positive eigenvalue λ. Furthermore, if the matrix A is irreducible, the eigenvector $\vec{r}$ has strictly positive entries, is unique and simple, and the corresponding eigenvalue is the largest eigenvalue of A in absolute value (i.e. is equal to the spectral radius of A).

Use of trivial 1, $\frac{1}{2}$ and 0 assignments for the matrix A results in a nonnegative matrix of the type described by this theorem. Further, it can be shown that in paired competition, A can only be irreducible when there are no winless teams. To extract the ranking vector from the outcome matrix, the power method can be used since $\vec{r}$ is a simple eigenvector corresponding to the largest eigenvalue of A.

$$\vec{r} = \lim_{n\to\infty} \frac{A^n \vec{r}_0}{|A^n \vec{r}_0|} \tag{5.6}$$

Using these basic assignments for the preference matrix provides some meaningful relationships with the initial ranking vector. If all elements of this initial vector, $\vec{r}_0$, are set to 1, then the *ith* component of $A\vec{r}_0$ corresponds to the winning percentage for team i. Further, the *ith* component of $A^2\vec{r}_0$ provides the average winning percentage for all opponents that team i defeated. Keener argues that, in a sense, this matrix based approach includes strength of schedule information for each particular team. While some critics argue that strength of schedule should be a very strong aspect, if not the emphasis of a ranking system, this observation highlights a major limitation on this school of thought. From a mathematical standpoint, (5.6) demonstrates that the unique positive eigenvector is obtained from $A^n\vec{r}_0$ as $n \to \infty$. This means that the ranking vector would become more predetermined and less outcome based from results of individual competitions.

A refinement is made to this relatively simple approach by redistributing the points assigned into the elements a_{ij} of the preference matrix. In a sport where a particular team may not play another many times (or even once) during a given season, the default 1, $\frac{1}{2}$ or 0 distribution doesn't provide enough information for calculating a reasonable ranking vector as it results in a sparse preference matrix. Therefore, an effort is made to distribute this one point between the two teams in a given competition not specifically based on win/loss but on other factors such as the game score. A simple ratio of points scored can be used such that $a_{ij} = S_{ij}/(S_{ij} + S_{ji})$ where S_{ij} is the number of points scored by team i during competition with team j. On one hand this provides some improvement for higher scoring games, but in defensive games that result in little or no scoring, individual elements of the preference matrix aren't partitioned fairly with this adjustment. In addition, a team shouldn't be rewarded for running up the score against a weaker opponent merely to pad its ranking. Therefore Keener developed a nonlinear modification to account for each end of this spectrum while partitioning the point in the preference matrix based on the results of the competition between both teams.

5.2.2 Nonlinear Scheme

Several adjustments can be made to Keener's direct method which result in a similar but nonlinear approach to obtaining the ranking vector. In particular, a new concept to calculate the ranking of a team is presented as

$$r_i = \frac{1}{n_i} \sum_{j=1}^{N} f(e_{ij} r_j) \tag{5.7}$$

where e_{ij} is a value that is determined from the outcome of a competition between teams i and j and f is some continuous monotone increasing function. Two ideal properties of f, namely $f(0) = 0$ and $f(\infty) = 1$, provide limitations on the assigned score a team earns in a given match. Now a team can earn only a maximum of one point for each game it plays either by playing well against a highly ranked opponent or by trouncing a weak one. Keener shows in [18] that a positive ranking vector still exists. In essence, this result with (5.7) can be viewed as a nonlinear generalization of the Perron-Frobenius theorem.

From experimentation, suitable functions were developed for f and e_{ij} so that an adequate sense of comparison could be maintained between the mathematical model and the ranking polls in college football. The equations are

$$f(x) = \frac{0.05x + x^2}{2 + 0.05x + x^2} \tag{5.8}$$

and

$$e_{ij} = \frac{5 + S_{ij} + S_{ij}^{2/3}}{5 + S_{ji} + S_{ij}^{2/3}} \tag{5.9}$$

The subjective nature in choosing these functions outlines the difficulty in using ranking systems based on these ideas. Some justification on their selection must be provided before significant merit can be placed in their meaning. Whether the source for this justification comes from existing ranking polls or elsewhere, there's little doubt that this fact alone is a weaknesses of such

a system. The other methods presented by Keener are less subjective because they have an improved theoretical basis. Nonetheless, the concepts here are significant in that there's a *technical* basis for assessing limitations in recording the invidividual outcomes in unpaired competitions.

5.2.3 Probabilistic Interpretations

The next step attempts to formulate the elements of $\vec{r}$ so that they might provide information on the probability that a team i beats another team j. A simple approach to define this probability is

$$\pi_{ij} = \frac{r_i}{r_i + r_j} \tag{5.10}$$

where

$$\pi_{ji} r_i - \pi_{ij} r_j = 0 \tag{5.11}$$

The relationship (5.11) holds because $\pi_{ij} + \pi_{ji} = 1$. However, since π_{ij} isn't known, there's no way to calculate $\vec{r}$. The problem of obtaining the ranking vector $\vec{r}$ from (5.11) is handled by a class of linear models having the form $\pi_{ij} = \Pi(v_i - v_j)$ where $\vec{v}$ is analagous to the ranking vector as discussed by Stob in [45]. The function Π is chosen as the best linear model and can be chosen analytically or by some statistical means. For example, a statistical least squares approach has been documented by Mosteller in [31]. The most natural choice to obtain π is to use game scores such that

$$\pi_{ij} = \frac{S_{ij}}{S_{ij} + S_{ji}} \tag{5.12}$$

using the same convention as before. Using (5.11) this relationship can be rewritten as

$$S_{ji}\, r_i - S_{ij}\, r_j = 0 \tag{5.13}$$

Use of (5.13) produces an overdetermined system of equations because there are many more games (and hence game scores) than there are teams in a given college football season. To calculate a "best fit" ranking vector, the least squares solution was selected. Using the constraint $|\vec{r}| = 1$, this method results in the need to solve the eigenvalue problem

$$B\vec{r} = \mu\vec{r} \tag{5.14}$$

with

$$b_{ii} = \sum_k S_{ij}^2$$
$$b_{ij} = -S_{ij}\, S_{ji} \quad i \neq j$$

The mathematical formulation for (5.14) and subsequent definitions can be found in [18]. Since B has positive diagonal but negative off diagonal elements, the ranking vector is calculated most efficiently using the inverse power method.

5.2.4 Application

Consider the following results of the lifetime competition between four professional tennis players: Andre Agassi, Andy Roddick, Lleyton Hewitt and Roger Federer through 2005. Table 5.1 provides the outcome of all lifetime matches between these four players.

Player 1	Player 2	Player 1 Wins	Player 2 Wins
Agassi	Roddick	5	1
Agassi	Hewitt	4	4
Agassi	Federer	3	8
Roddick	Hewitt	2	6
Roddick	Federer	1	10
Hewitt	Federer	7	11

Table 5.1. Lifetime head-to-head match results for A. Agassi, A. Roddick, L. Hewitt and R. Federer

Note that the results are *not* in paired competition format. The final standings for each player are as shown in Table 5.2.

Player	Wins	Losses
Federer	29	11
Hewitt	17	17
Agassi	12	13
Roddick	4	21

Table 5.2. Lifetime standings for A. Agassi, A. Roddick, L. Hewitt and R. Federer

The preference matrix for these matches as described by Keener's direct method is listed in (5.15).

$$A = \begin{pmatrix} 0 & \frac{11}{40} & \frac{8}{40} & \frac{10}{40} \\ \frac{7}{34} & 0 & \frac{4}{34} & \frac{6}{34} \\ \frac{3}{25} & \frac{4}{25} & 0 & \frac{5}{25} \\ \frac{1}{25} & \frac{2}{25} & \frac{1}{25} & 0 \end{pmatrix} \tag{5.15}$$

where the rows correspond to the same order players have been listed in Table 5.2. Note, again, that this preference matrix is irreducible. Since A is a nonnegative matrix, it's known from the Perron-Frobenius Theorem that at least one positive eigenvalue exists. The eigenvalues of A are 0.4340, -0.2328, -0.1127 and -0.0886. Thus, the ranking vector is the eigenvector corresponding to $\lambda = 0.4340$. This eigenvector, $\vec{r}$, is [0.6721, 0.5303, 0.4751, 0.2035]. Therefore, Keener's direct method suggests the rankings of these players as listed in Table 5.3.

Rank	Player	r_i
1	Federer	0.6721
2	Hewitt	0.5303
3	Agassi	0.4751
4	Roddick	0.2035

Table 5.3. Keener rankings for A. Agassi, A. Roddick, L. Hewitt and R. Federer based on their lifetime head-to-head competitions

Keener's direct method provides a reasonable result. The player with the highest winning percentage, Federer, ends up with the highest ranking and conversely the player that won the least games, Roddick, has the lowest rank. The rankings result in a very small difference between Agassi and Hewitt which can be justified. Both of these players have similar winning percentages with Agassi just slightly below 0.500. As a result, Agassi's rating is slightly less than Hewitt's. Federer's results clearly indicate that he dominated the other players under consideration and

therefore he ends up with the highest ranking. Also note that meaningful information can be extracted from relative comparisons of the individual elements of the ranking vector itself. This will become more important when probabilistic elements are used to build the preference matrix.

5.3 PageRank

Another very common ranking system that influences the lives of most people on a regular basis is the PageRank algorithm used to produce search results for the popular Internet search engine Google. Understanding the core concepts that formulate PageRank proves to be another application of standard linear algebra. Bryan and Leise have documented the mathematical, linear algebra basis for which these web searches are executed in [6]. The concept of producing a ranked list of most "important" web links (and hence websites) is the number of back links made to some specific web page by other pages. A simple approach, therefore, is to take the number of links for a specific web page k from other pages and count them as votes for page k's importance. To account for the increased importance of a vote from a more important page, a page score is calculated. The page score for page k is simply the sum of the page scores for all pages that link to k. Furthermore, to prevent a single page from contributing significantly more or less to the score of another page, a weighted score is introduced so that each page gets a total of one vote as scaled by its score. Namely,

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j} \tag{5.16}$$

where x_k is the score for page k, n_j is the number of outgoing links from page j, and L_k is the set of page k's backlinks. Ultimately, the set of linear equations resulting from (5.16) produce of system equations that can be written $A\vec{x} = \lambda\vec{x}$. The matrix A is known as the link matrix or under the guise of Keener's method it's known as the preference matrix. In fact, PageRank can be thought of as an extension of Keener's eigenvalue based ranking system where the number of *normalized backlinks* are used to formulate the preference matrix.

5.4 The Colley Bias Free Matrix Method

5.4.1 Motivation

The Colley matrix ranking method is one of the computer based methods that make up the Bowl Championship Series conglomerate in college football. In development of his ranking system, Colley has outlined a series of key features he considers to be requirements. His hope is to convince others that the Colley matrix method (as fully documented in [11])

1. has no bias towards conference, tradition, history, etc...
2. is reproducible
3. uses a minimum of assumptions
4. uses no *ad hoc* adjustments
5. adjusts for a teams strength of opponents
6. ignores large score disparities
7. produces common sense results

This motivation provides a baseline for the formulation of his method. Further, Colley argues that using only wins and losses as statistics for producing rankings provides three key elements

that support the requirements he's defined for the optimal system. First, it eliminates bias towards a team's conference or alignment, it removes the need to create an ad-hoc method to deal with scoring margins and finally it nullifies any further arbitrary adjustments one would consider making, such as game location or game surface. This is in stark contrast to Keener's nonlinear method described in Section §5.2.2 where functions have been arbitrarily created to align the final rankings in a more "reasonable" manner with those from polls. The selection of using only wins and losses for ranking parameters accomplishes nearly half of the goals previously defined by Colley.

5.4.2 Laplace's Method as a Basis for Rankings

The win/loss rating problem initially defined by Colley is related to an old problem first discussed by the famous mathematician Pierre-Simon Laplace in locating a randomly thrown dart onto a rectangular dartboard. Consider a rectangular dartboard of unit width with a vertical divider drawn somewhere on the board. The divider can't be seen by the dart thrower, but the thrower is told if the dart lands to the left or right of the divider after it hits the board. The task for the thrower is to make his best guess on the location of the divider based on the results of his throws. Colley declares the analogy to football ranking is that he must make a good guess as to a team's true rating based on only wins and losses.

The dart thrower guesses most appropriately that initially the divider is through the center of the dartboard at r=$\frac{1}{2}$. Assuming the thrower is blindfolded, it's reasonable to assume a uniform distribution of tosses on the unit interval. The average or expected value of this distribution is $\frac{1}{2}$. Calculation of the expected value requires the probability density function which can be obtained

from the probability distribution function as $f_X(x) = \frac{d}{dx}F_X(x)$. Since the distribution function is known and has been given in (2.2) the density function for $Uniform[0,1]$ is simply $f_X(x) = 1$. Now

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{5.17}$$

can be used to calculate the expected value. Therefore in the case of no darts thrown

$$r = \frac{\int_{-\infty}^{\infty} \hat{r} f(\hat{r})\, d\hat{r}}{\int_{-\infty}^{\infty} f(\hat{r})\, d\hat{r}} = \frac{\int_0^1 \hat{r}\, d\hat{r}}{\int_0^1 d\hat{r}} = \frac{1}{2} \tag{5.18}$$

with the rating random variable $\hat{r}$. Now, assuming that the first dart throw lands to the left of the divider, the probability density at the left edge of the board must be zero ($\hat{r} = 0$) since a dart can't be thrown to the left of the edge. The probability density for $\hat{r}$ simply increases linearly to the right as this corresponds with the available space between the the divider and the edge where the dart could have landed. Colley correlates this result to ranking due to the following ideas: (1) if one team has beaten another, the winner can't be the worst team after one game and (2) the number of teams that are worse than the winner increases proportionally with the rating $\hat{r}$ (see [11] for specific details). The expected value of the location of the divider (corresponding to team's rating) after one left throw (one win) is

$$r = \frac{\int_0^1 \hat{r}^2\, d\hat{r}}{\int_0^1 \hat{r}\, d\hat{r}} = \frac{2}{3} \tag{5.19}$$

since the probability densities multiply based on independence between throws (games). On the other hand, when a dart is thrown to the right of the divider, the probability density increases

from zero at the right edge. A term that grows linearly from the right is introduced as $(1-\hat{r})$. One dart that has been thrown the left and one that has been thrown to the right of the divider results in the following rank

$$r = \frac{\int_0^1 (1-\hat{r})\,\hat{r}^2\,d\hat{r}}{\int_0^1 (1-\hat{r})\,\hat{r}\,d\hat{r}} = \frac{1}{2} \tag{5.20}$$

In general, for n_w wins and n_l losses, the rank for a given team can be calculated as

$$r = \frac{\int_0^1 (1-\hat{r})^{n_l}\,\hat{r}^{n_w}\,\hat{r}\,d\hat{r}}{\int_0^1 (1-\hat{r})^{n_l}\,\hat{r}^{n_w}\,d\hat{r}} = \frac{1+n_w}{2+n_l+n_w} \tag{5.21}$$

This ranking provides several advantages over simply using a team's winning percentage. The first, and perhaps most obvious, is that all teams start off a season with an equal rating of $\frac{1}{2}$. Another advantage is that most reasonable comparisons can be made with respect to ratings. For example, consider the ratings of two teams after one game has been played. The winning team ends up with a rating of $\frac{2}{3}$ while the losing team ends up rated $\frac{1}{3}$. If winning percentage alone had been used for rating, the winner would have the rank 1 and the loser would have the rank 0. In [11], Colley declares that it's much more reasonable to think of the winning team as "twice as good" as opposed to "infinitely better than" the losing team.

5.4.3 Adjusting for Strength of Opponents

To accomplish his original motivation, Colley includes an adjustment in the ranking value calculated for the strength of a team's opponents. First, note that the number of wins by a particular team, n_w, can be rewritten as

$$n_w = \frac{n_w - n_l}{2} + \frac{n_w + n_l}{2} = \frac{n_w - n_l}{2} + \frac{n_{tot}}{2} \tag{5.22}$$

Now, the second term $\frac{n_{tot}}{2}$ can be thought of as the sum $\sum_{n_{tot}} \frac{1}{2}$. This is a sum of the ratings of all the opponents for a given team using the default rating, $r = \frac{1}{2}$. Instead, if the actual rating of each opponent is used in this sum, a new, effective "number of wins" value can be calculated as

$$n_{w,eff} = \frac{n_w - n_l}{2} + \sum_{j=1}^{n_{tot}} r_j \tag{5.23}$$

where r_j is the rating of the jth opponent of the team under consideration. This modification to a team's number of wins, n_w, comprises the adjustment for its strength of schedule.

5.4.4 The Colley Matrix

The equations (5.21) and (5.23) can be rearranged in the form

$$(2 + n_{tot,i})\, r_i - \sum_{j=1}^{n_{tot,i}} r_j^i = 1 + \frac{n_{w,i} - n_{l,i}}{2} \tag{5.24}$$

which is an N by N system of linear equations where N is the number of teams being ranked. A convenient switch is made to matrix form by rewriting (5.24) as

$$C\vec{r} = \vec{b} \tag{5.25}$$

where $\vec{r}$ is the ranking vector, $\vec{b}$ is vector of the right hand side of (5.24) and C is the self titled Colley matrix comprised of

$$b_i = 1 + \frac{n_{w,i} - n_{l,i}}{2}$$

$$c_{ii} = 2 + n_{tot,i}$$

$$c_{ij} = -n_{j,i}, \quad i \neq j$$

where $n_{j,i}$ is the number of times that team i played team j.

5.4.5 Application

To compare results between Keener's direct method and Colley's matrix method, a new Colley ranking vector is presented. Lifetime results between the four tennis players Agassi, Federer, Hewitt and Roddick through 2005 can be reviewed in Table 5.1 and Table 5.2. Based on these results, the Colley matrix would be

$$C = \begin{pmatrix} 42 & -18 & -11 & -11 \\ -18 & 36 & -8 & -8 \\ -11 & -8 & 27 & -6 \\ -11 & -8 & -6 & 27 \end{pmatrix} \tag{5.26}$$

with corresponding $\vec{b}$ vector

$$\vec{b} = \begin{pmatrix} 10 \\ 1 \\ 0.5 \\ -7.5 \end{pmatrix} \tag{5.27}$$

Solving (5.25) for these values of C and $\vec{b}$ produce the ranking vector $\vec{r}$ with elements 0.6751, 0.5398, 0.5138 and 0.2714. Table 5.4 provides these results from the Colley matrix method.

Rank	Player	r_i	$\bar{r}_i$
1	Federer	0.6751	0.6482
2	Hewitt	0.5398	0.5183
3	Agassi	0.5138	0.4933
4	Roddick	0.2714	0.2606

Table 5.4. Colley rankings for A. Agassi, A. Roddick, L. Hewitt and R. Federer based on their lifetime head to head competitions

The fourth column containing $\bar{r}$ is a unit normalized verson of the ranking vector $\vec{r}$. Normalization has been performed so that a more accurate comparison between Keener's direct method ranking results can be reviewed. The results in Table 5.3 and Table 5.4 show that there's very little discrepancy between rankings using Keener's direct method and Colley's matrix method in this particular example.

5.5 Random Walker Rankings

Another ranking system, developed by Callahan, Mucha and Porter in [7], uses a simply explained algorithm consisting of a set of random walkers that cast a single vote for the team that they think is the best. Each of the random walker voters routinely examines the outcome of a game between its favorite team and an opponent and decides whether to continue voting for their favorite or to switch allegiance to the other team based on the outcome. The particular game the voter chooses to examine is selected at random from the team's schedule. To more correctly model the behavior of real life voters, the random walker's preference is to go with the winner of the particular game he chooses to examine. However, this isn't a certainty which leads to the definition of the parameter p. p is defined as the probability that the random voter chooses the winner of the game. $p > \frac{1}{2}$ because on average the random walker will vote for the team that won the game. However, $p < 1$ because the voter shouldn't in all certainty choose to vote for the winner of the game. This allows the voter to argue that somehow the loser of the game is still the best overall.

An advantage of this particular ranking method is that it contains a single, precisely defined parameter p that has a meaningful interpretation when considering the mentality of a single voter in evaluating competitions. No attempt is made by the developers of the algorithm to claim it's either better or worse than other ranking algorithms including the ones previously presented. In fact, Callahan, Mucha and Porter credit Keener's work (see [18]) as his direct matrix method can be related to the linear matrix style problem that is ultimately formulated here with this approach. A significant motivation in the development of this ranking system is to show that perhaps even automated random walkers can adequately rank college football teams in comparison to the numerous other ranking methods in existence.

5.5.1 Mathematical Definition

For each team i in a set of N teams, denote the number of wins, losses and number of voters casting their single vote for team i as w_i, l_i and v_i, respectively. Additionally, the number of games played by team i is n_i and the total number of voters remains constant, $\sum_i v_i = Q$. A natural way to express the independence of random walker dynamics is through a system of ordinary differential equations (ODEs). First note that when team i beats team j, the rate at which a walker voting for team j switches allegiance to team i is proportional to p and the rate at which a walker that is already voting for team i switches his vote to team j is proportional to $(1-p)$. The rate which a voter considers a game by a given team is independent of n_i.

The system of linear ODEs that describe the expected rate of change of the number of votes cast for a particular team i can be written as

$$\vec{v}' = D\vec{v} \tag{5.28}$$

where $\mathbf{v}$ is the vector of votes v_i for team i. Elements for the square matrix D are given by

$$\begin{aligned} D_{ii} &= -p\,l_i - (1-p)w_i \\ D_{ij} &= \frac{1}{2}N_{ij} + \frac{(2p-1)}{2}A_{ij}, \quad i \neq j \end{aligned} \tag{5.29}$$

The term N_{ij} is the number of games between team i and team j. The term A_{ij} is the number of times team i defeated team j minus the number of times team i lost to team j. A full expansion

for the matrix D is available in [7]. Consider, for example, the case of two teams, x and y that play a single game where x defeats y. Equation (5.28) in this case would be

$$\underbrace{\begin{pmatrix} v_x \\ v_y \end{pmatrix}'}_{\vec{v}'} = \underbrace{\begin{pmatrix} -(1-p) & p \\ (1-p) & -p \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} v_x \\ v_y \end{pmatrix}}_{\vec{v}} \tag{5.30}$$

This describes the flow of votes between team x and y. Since team x has won a single game, this flow would encompass a rate of voters proportional to $(1-p)$ that change allegiance to y (flow out) and a rate of voters proportional to p that now believe x is the best team. The exact converse is true for team y in this example. This illustrates not only the rationale behind the constraint $\frac{1}{2} < p < 1$ but a conservation of total votes.

To determine the final expected population of voters for each team, a solution to the steady state equilibrium equation

$$D\vec{v} = 0 \tag{5.31}$$

must be obtained. This final population can then be used to directly rank the set of all teams. It can be shown that there exists a single attracting equilibrium $\vec{v}$ for $\frac{1}{2} < p < 1$ in [7]. Since the parameter p is the only input to the method, it's not surprising that the rankings can vary substantially based on the value selected. Results in [7] indicate that for college football, the *relative* ranking of the top few teams remain similar across a wide range of p.

5.5.2 Application

Further comparisons can be made with the rankings previously calculated using Keener's direct matrix method and Colley's matrix method with this random walker method. The parameter $p = 0.75$ is used to incorporate elements of both the "strength of opponent" and "winner takes all" aspects discussed for the limiting cases $p \approx \frac{1}{2}$ and $p \approx 1$, respectively. Refer once again to Table 5.1 and Table 5.2 for the results of head to head matches and final standings between Agassi, Roddick, Hewitt and Federer. The resulting rate matrix D in this case would be

$$D = \begin{pmatrix} -15.5 & 10 & 6.75 & 7.75 \\ 8 & -17 & 4 & 5 \\ 4.25 & 4 & -12.75 & 4 \\ 3.25 & 3 & 2 & -16.75 \end{pmatrix} \tag{5.32}$$

with steady state random walker voter distribution (null space) $\vec{v}$ containing elements 0.6690, 0.5064, 0.4686 and 0.2765. Random walker rankings for $p = 0.75$ are included in Table 5.5.

Rank	Player	v_i
1	Federer	0.6690
2	Hewitt	0.5064
3	Agassi	0.4686
4	Roddick	0.2765

Table 5.5. Random walker rankings for A. Agassi, A. Roddick, L. Hewitt and R. Federer based on their lifetime head to head competitions using $p = 0.75$

Table 5.6 provides results from all three ranking methods used on this particular tennis example, Keener's direct matrix method, the Colley matrix method and the random walker method

($p = 0.75$). For comparison purposes, the normalized rankings from the Colley matrix method are included here (all resulting ranking vectors have norm of 1).

Rank	Player	r_i Keener	$\bar{r}_i$ Colley	v_i random walker
1	Federer	0.6721	0.6482	0.6690
2	Hewitt	0.5303	0.5183	0.5064
3	Agassi	0.4751	0.4933	0.4686
4	Roddick	0.2035	0.2606	0.2765

Table 5.6. Comprehensive rankings for A. Agassi, A. Roddick, L. Hewitt and R. Federer based on their lifetime head to head competitions

Once again there's little discrepancy between ranking values and no discrepancy whatsoever with regard to the final rank of the players. Though this is a positive outcome, the individual matchups and final standings between these players should leave little doubt as to their resulting rank. Nevertheless, these results provide a solid baseline for the validity of the ranking algorithms particularly due to their strong correlation and will lend themselves to application of robust rankings through the use of the Monte Carlo *pdf* player models.

5.6 ATP Rankings

Of particular interest for the application of tennis is the ATP ranking method (singles) used to rank and seed professional players in most ATP sanctioned events. With a few exceptions, an ATP player's rank is determined by summing points earned for tournaments played during a sliding 52-week period. In this manner, the ATP ranking system is significantly different from those previously discussed as no attempt is made to mathematically scale results based on the type or quality of opponents played. A player's rank is determined simply by the direct, cumulative

outcome of his tournament performance. Because of this, it can be argued that the ATP ranking system is fundamentally flawed. While mandatory tournaments are stipulated in the rulebook (Grand Slam events, for example) a player can select tournaments based on the perceived quality of the other players in that field and potential points to be awarded. The events for which points are included into the rankings are

1. the four Grand Slams
2. eight mandatory ATP World Tour Masters 1000's
3. the ATP World Tour Finals (if the player qualifies)
4. best four results from ATP World Tour 500's
5. best two results from ATP World Tour 250/ATP Challenger Tour/Future Series

in the sliding 52-week period (as defined by the ATP rulebook in [37]). Points are allocated based on tournament category and are presented in simplified form in Table 5.7.

	W	**F**	**SF**	**QF**	**R16**	**R32**	**R64**	**R128**	**Q**
Grand Slams	2000	1200	720	360	180	90	45	10	25
ATP 1000	1000	600	360	180	90	45	10(25)	(10)	(1)25
ATP 500	500	300	180	90	45	(20)			(2)20
ATP 250	250	150	90	45	20	(5)			(3)12

(1) 12 points only if the main draw is larger than 56
(2) 10 points only if the main draw is larger than 32
(3) 5 points only if the main draw is larger than 32

Table 5.7. Simplified ATP ranking point allocations

The rankings are then used to seed the main draw for tournaments as described in [37]. In all cases, only the first and second ranked player in any given draw are guaranteed specific positions in the draw. Subsequent players are seeded based on random selections in various pots. The

resulting system provides a certain randomness in a given tournament draw for a given set of players. In this way, there have been circumstances in the past that have resulted in tournament mismatches where a higher ranked (seeded) player had to play another highly seeded player before a lower seeded player was knocked out with an upset. Additionally, only a small, fixed number of seeded players are alloted per tournament, creating scenarios where a top seed can play a highly ranked, unseeded player very early in the competition. In 2001, player discord over discrepancies between the ATP rankings and the ad hoc seeding system used by the All England Club in formulating the draw for Wimbledon had become high enough for the players to threaten boycotting that event. Cancellation of that tournament didn't occur as a uninamous agreement was reached to increase the number of Grand Slam seeds from 16 to 32 (with adjustments to reflect player form on different surfaces). However, concern on the fairness of the ATP rankings themselves has persisted through the years. As recently as 2011, Rafael Nadal has advocated a switch to a two-year ATP ranking that he argued would more adequately protect all players. Specifically, he cited the comeback of Juan Martin Del Potro from the number 485 world ranking (January 2011) after being injured to the number 3 position (December 2011). It's natural to question the overall fairness of the ATP rankings when scenarios such as this can occur.

5.7 Arrow's Impossibility Theorem and Probabilistic Ranking Systems

Sports ranking systems take on many different forms as reviewed by Stefani in [43] and as discussed here, but roughly speaking they can be grouped into two distinct categories. One type attempts to reward teams after a predetermined number of games are played by, for example,

crowning them as national champions or year end champions. These types of ranking systems are denoted "outcome based". The philosophy behind these methods is that there should be a player or team who *deserves* to be recognized as *the best*, and if only the correct method were found, such a team or player could be unambiguouly chosen. Examples of systems of this type are the previously described BCS system and the ATP/WTA rankings used in men's and women's professional tennis, respectively. They typically rely on methods that include information that is fielded directly (i.e. win/loss or score) from the outcome of their competitions.

With regard to this use of ranking systems, it's an underappreciated fact that Arrow's Impossibility Theorem [2] for aggregating individual preferences into social preferences holds. The theorem states that *there is no rule*, majority voting or otherwise, for establishing social preferences from arbitrary individual preferences (random walker notwithstanding - see Section §5.5). In other words, under certain conditions of rationality and equality, it's *impossible* to guarantee that a ranking of societal preferences will correspond to rankings of individual preferences when more than two individuals and alternative choices are involved.

As a simple and concrete example of how this translates into the ability of outcome based methods to unambiguously determine rankings, consider the problem of trying to decide upon the year end men's tennis champion for the 2002 ATP season in which Pete Sampras won the US Open, Andre Agassi won the Australian Open and Leyton Hewitt won Wimbledon. Suppose this is done by majority voting among three judges between each pair of players. Let one of the judges be an American who might naturally favor the US Open champion, one an Australian who favors the Australian Open champion and the third British judge who favors the Wimbledon champion. A set of preferences is said to be rational (or transitive) if when player A is ranked

higher than player B and B is ranked higher than C, then A is ranked higher than C. Certainly, this is a desirable property for any ranking system and any system that produces an outcome without this property is likely to be viewed as unfair. Consider Table 5.8 which shows rankings chosen by the three judges.

Judge	Sampras	Agassi	Hewitt
American	1	2	3
British	2	3	1
Australian	3	1	2

Table 5.8. Fictional year end voting by three judges in the 2002 men's professional tennis tour

The American judge gives Sampras his number one rating, the British judge gives Hewitt his top rating and the Australian judge gives Agassi his number one rating. The number two and three ranking for each of the judges is also shown in Table 5.8. If the final rankings are compiled by majority vote, Sampras would be ranked higher in a choice between those two since two out of three judges vote Sampras as the higher ranked player. In a choice between Agassi and Hewitt, Agassi would be ranked ahead of Hewitt since two out of three judges voted this way. Now, since Sampras was chosen ahead of Agassi, and Agassi ahead of Hewitt, logic would dictate that Sampras be ranked ahead of Hewitt, i.e. transitivity should hold. But consider the outcome of the voting among the three judges in trying to decide between Sampras and Hewitt - two out of three of the judges rank Hewitt higher than Sampras. The outcome could be viewed as irrational as transitivity does not hold.

Despite the fact that Kenneth Arrow was awarded the Nobel Prize in Economics in 1972 for his work, there continues to be a widespread belief that if the right method were discovered,

there would be a "correct" way to crown a national champion in college football or a year end champion in tennis which would eliminate irrational outcomes, settle all arguments and leave everyone satisfied. However, Arrow proved that under certain reasonable assumptions, there is *no* method for constructing social preferences (rankings) from arbitrary individual ones (votes). Such "outcome based" methods that rely on voting, like the one used to crown the national college football champion, very often produce logical inconsistencies that are the basis for arguments that can't be settled rationally. The amount of energy and effort spent on arguing over rankings of all types (particularly in college football where few games are played compared to the total number of teams involved, but also the notorious U.S. News and World Report Annual Ranking of Colleges) is an indication of the pervasiveness of Arrow's theorem.

A second and very different kind of system attempts to use rankings for the purposes of predicting outcomes. Hence, they're called "predictive methods". In a sense, they're inherently probabilistic. An example of this type of system is the Elo Rating used by the amateur and professional chess tour. Recall from the previous overview that these rankings are designed so that they can be interpreted into predictions of outcomes in head-to-head competitions (also referenced by Stefani in [43]). In tennis, it's widely recognized that this isn't possible with the current ranking format. In fact, two of the four Grand Slam tournaments (the French Open and Wimbledon) don't use the ATP rankings when it comes to seeding players in their draws. The players and spectators, therefore, are left to wonder what sort of ad hoc system is adopted for choosing the seeds as this choice can have very direct and important consequences on how far players can advance in the tournament which in turn affects their future seedings, rankings and earnings. For example, the seedings of the men's 2005 Wimbledon draw created a controversy when the number two ranked

player in the world, Lleyton Hewitt, was seeded third causing him to face the top ranked player, Roger Federer, in the semifinal round instead of the finals. His loss to Federer in the semifinals (instead of the finals) cost Hewitt hundreds of thousands of dollars in earnings and a significant loss in ATP ranking points. With regard to world rankings, recently on two separate occasions in women's professional tennis, the top ranking went to players who had never won a Grand Slam title.

One of the ways to limit the consequences of Arrow's Theorem is to represent the alternatives as elements in a spectrum of possibilities, i.e to use probabilistic based ranking systems. Then, if the preferences of the individual exhibits *single-peakedness*, the societal preferences can be constructed unambiguously. The Markov chain model has the potential to provide such *distributional* information as the *pdf* player models are built on these concepts, hence it represents an important step in the direction of developing probabilistic rankings in tennis. Consider the idea of running thousands of simulated tournaments with players randomly ordered in ficticious draws before the tournament begins and using the accumulated statistical winning distributions as the basis for seeding actual tournaments before they're played. The player who is most likely to win the tournament based on the simulations would be the number one seed in the real draw, the player with the second highest winning percentage would be seeded second and so on. Consider further, the fact that highly robust player models have been developed and described in detail in Chapter §4. This overcomes the limitation of trying to incorporate information directly from head-to-head competitions into the rankings. Rather, the individual outcomes flesh out the performance of the *pdf models* used to simulate these notional tournaments. Further, these robust player models can be enveloped to any specific period of performance to provide the most accurate description

for the application used. In other words, fair matchups can be simulated between any player from any generation by constraining player models to “peak” years. For example, realizations of matches between a 2002 Pete Sampras and a 2008 Roger Federer could be executed fairly, with meaningful results. Using the adaptive player concept along with performance based (as opposed to outcome based) inputs can provide inherently probabilistic and predicitive rankings that circumvent the inherent inconsistencies guaranteed by Arrow’s Impossibility Theorem.

Chapter 6.

Proposed Ranking Schemes

A pleasant outcome and key contribution of the Monte Carlo *pdf* based Markov chain model for tennis is its application towards addressing some of the problems documented with modern ranking methods. Specifically, it's anticipated that significant improvements to the ATP ranking methodology can be made with use of the Monte Carlo model in simply providing a probabilistic basis for ranking results. Two distinct, novel approaches are presented here that simultaneously address some of the ranking system limitations discussed in Chapter §5 while providing probabilistic elements to the results.

6.1 Tournament Simulation Method

The first ranking system is based on Monte Carlo simulated tournaments. In this case, rankings are inherently equitable as they're derived specifically from realizations of the actual events from which they're designed to seed. These rankings directly incorporate the single elimination stakes that occasionally result in major upsets with impacts to all other tournament participants. Using the same performance timeframe considered by the ATP, each player's performance (as well as "the field") is modeled using the results from the previous 52 weeks. Reference Appendix A

for an inclusive list of performance metrics for "the field" between 2007 and 2009. 1,000 16-player random draw tournaments were simulated after each Grand Slam tournament and at the end of each season (calendar year) between 2007 and 2009. 16 players were chosen for ease of comparison and discussion here, but in practice, a full 128-player Grand Slam tournament field could be used (ranked). More specifically, the top 16 ranked ATP players were chosen for comparison in these simulated tournament draws at the specific times corresponding to the ATP events considered. Points were awarded at the end of each simulated tournament as shown in Table 6.1 using a 2^n basis.

Round	Round Points	Accum Points	Scaled Accum Points
R16 (8)	0	0	0
R8 (4)	2	2	3.125
SF (2)	4	6	9.375
F (1)	8	14	21.875
W (1)	16	30	46.875

Table 6.1. Tournament simulation ranking point allocations (16-players)

The total points are scaled such that the sum of all awarded points total to 100 for each tournament iteration. Note that from the scaled accumulated points listed in Table 6.1 the tournament winner receives roughly half of the total allocation while the players that win only one match (R8) get only about 3% of the points available. This is done to highlight and emphasize the nature of the single elimination tournament competition and to provide a framework for collecting data that can be used to seed players based on their predicted performance in the competition. After 1,000 tournament simulations, a ranking can be extracted from the sorted list of average points earned

as governed by Table 6.1. To illustrate this, Table 6.2 lists the tournament simulation rankings for the top 16 players at the end of the 2009 ATP season.

Rank	Player	Points	σ
1	Andy Murray	9.87	15.19
2	Rafael Nadal	9.77	15.13
3	Roger Federer	8.04	13.43
4	Andy Roddick	7.66	13.40
5	Robin Soderling	7.17	12.83
6	Nikolay Davydenko	7.05	12.54
7	Novak Djokovic	6.89	12.53
8	Fernando Verdasco	6.74	12.38
9	Juan Martin Del Potro	5.65	11.25
10	Jo-Wilfried Tsonga	5.10	10.30
11	Marin Cilic	4.97	10.18
12	Gilles Simon	4.78	9.73
13	Tommy Robredo	4.70	9.84
14	Fernando Gonzalez	4.64	10.52
15	Radek Stepanek	3.66	7.80
16	Gael Monfils	3.32	7.63

Table 6.2. Tournament simulation rankings for 2009 ATP season

Several key additional points can be made with the presentation of these rankings. First, a random draw tournament ensures equality in the matchups between iterations *without* preference to any type of seeding. Thus, the only key inputs to the rankings are: (1) the *pdf* player models that most accurately describe any given player against "the field" for the timeframe under consideration and (2) the pyramid style competition that a player experiences in an ATP tournament or Grand Slam event. Use of the *pdf* player model mitigates the concern that insufficient data exists for a given head-to-head matchup that could be encountered in a tournament realization. From this perspective, each player under consideration starts exactly equal. The simulations, then,

distinguish players based on their modeled performance within the guise of the sudden-death tournament structure.

The ranking results provide interesting information. Most notably, the top ranking goes to Murray who finished the 2009 ATP Season ranked fourth. The ATP number one ranked player, Federer, slipped to position three in the tournament rank while Djokovic, the ATP number three player, slipped all the way down to position seven. Quantitatively, both Federer and Djokovic have scaled accumulated point totals that aren't in line with the top tier players Murray and Nadal in this example. This illustrates that without the biases of tournament seeding and head-to-head matchups, the actual rankings of Federer and Djokovic are potentially inflated by the ATP scoring system at the end of 2009. Also note that Federer's 2010 performance had seen him slip from the top ATP rank down to fourth - coincidentally in line with these results. Djokovic's actual peformance is clearly more anomalous as his 2010 performance vaulted him into the top ATP ranking even though the tournament method ranked him much lower at position seven. This could be explained to some extent by his inconsistent performance and the lack of quality of competition he faced during the 2009 season. Also, consider the fact that Djokovic was ranked as low as number 22 in the 2006-2007 timeframe also highlighting his drastic improvement in the years leading up to 2009.

To illustrate this disparity and further analyze the results, statistics have been gathered on the number of times a particular player wins 1, 2, 3 or 4 matches (winner for a 16-player draw) over the 1,000 tournament sample. Reference Figure 6.1 depicting the simulated round-by-round performance of Federer, Nadal, Murray and Roddick based on the results from Table 6.2.

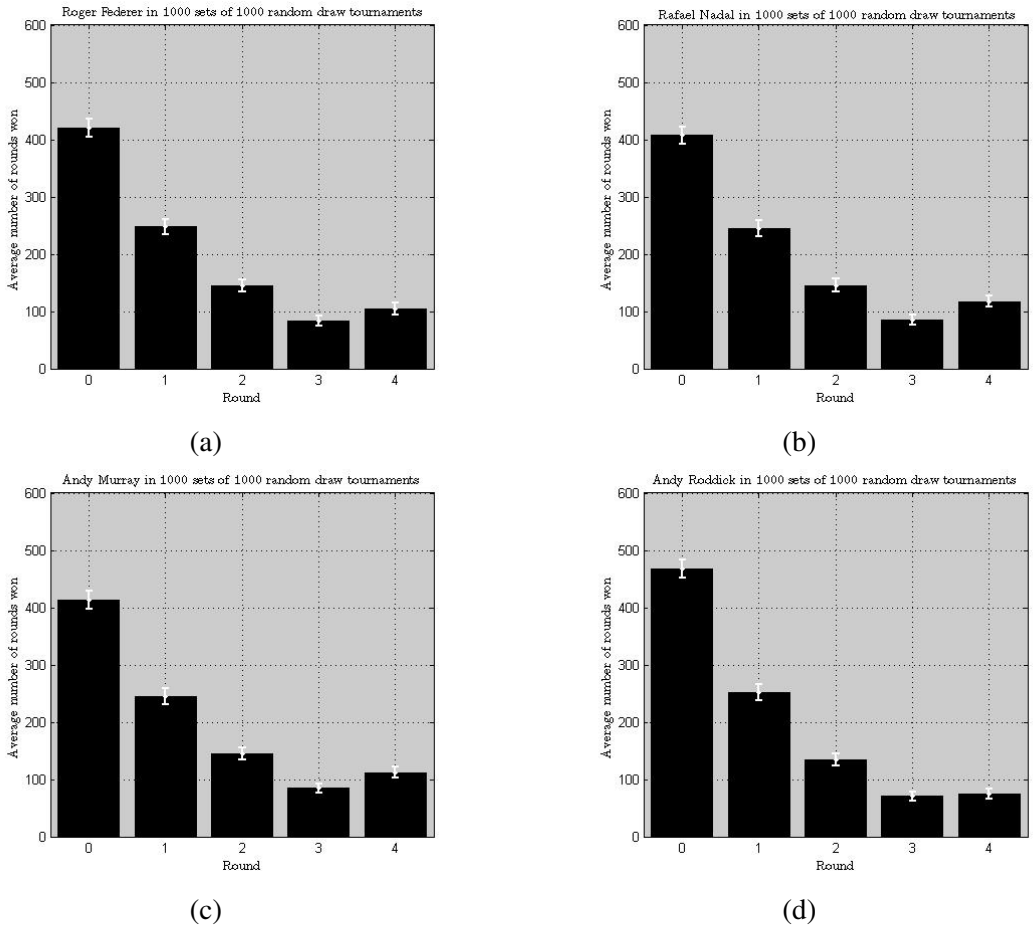


Figure 6.1. Histogram of simulated rounds won during 1000 sets of 1000 full tournament simulations for R. Federer, R. Nadal, A. Murray and A. Roddick after the 2009 ATP Season

A few key points are worth making. First, even for these top flight players, the probability of winning no matches (zero rounds won) is generally around 40%. Also note that the incidence of winning the tournament is always higher than that of losing the tournament once a player reaches the final match. Recall that this arises from the fact that these players are effectively above the field average in overall performance. The histograms shown in 6.1 exhibit remarkable similarity to those presented in Section §4.2.7 using random draws for the 2007 Wimbledon and 2007 US Open in Figure 4.10 and Figure 4.12. This demonstrates the *fundamental correlation* between the tournament rankings and the single elimination tournament scheme as baselined to the results previously presented.

In the discussion of these rankings, another key metric of interest is the standard deviation of points accumulated in the tournament simulations. Note that these values generally decrease starting from the top rank. At first glance, it may seem concerning that the average points accumulated for each player have such statistical variation. However, this is an inherent feature of the rankings as the results capture the fluctuations in outcomes that occur between tournament realizations. For example, a particular player that wins a tournament has some reasonable likelihood of being eliminated in the opening round of the next realization. The point system and outcome of these realizations result in a measurable disparity in points awarded between the individual trials. Data on statistical variation of the number of rounds won has also been gathered to quantify its effect on player performance and ultimately the rankings. Data are presented in Table 6.3 for standard deviations of rounds won on a round-by-round basis for the top four players in Table 6.2 where σ_i denotes one standard deviation of rounds won for round i on each player.

Player	σ_0	σ_1	σ_2	σ_3	σ_4
Andy Murray	15.96	14.00	10.80	8.31	9.57
Rafael Nadal	15.32	13.42	10.95	9.01	10.27
Roger Federer	15.72	13.46	11.02	8.81	9.77
Andy Roddick	16.06	14.04	10.83	8.04	8.28

Table 6.3. Standard deviation rounds won during 1000 sets of 1000 full tournament simulations for R. Federer, R. Nadal, A. Murray and A. Roddick rankings for 2009 ATP season

This can also be seen graphically by referring to the error bars included on Figure 6.1. Note that the individual values for σ_i reflect the same trend sequence seen as with the number of rounds won by each player in Figure 6.1. This is a positive characteristic that exhibits additional correlation between statistical variation in rounds won as the tournaments progress along with the histogram bins themselves. Based on these results, the output on variance of individual rounds won by each player and the resulting average accumulation of scaled points together form a meaningful composite.

These concepts accentuate the fact that use of the random draw format is ideal for formulating rankings as it characterizes a player's tournament performance fairly, without bias to a particular seed. In Chapter §4, the advantages of seeding in the tournament format have been shown for players at some of the various positions. Note that there's no clear pattern of performance that can be extracted from the individual seed positions when accounting for differences in player skill at those positions. Now, using random draws as executed here, these biases can be stripped so that the seeding "advantages" can be fairly applied to those players that earn them.

6.2 Matrix Method

Background on matrix based ranking approaches has been provided in Sections §5.2 and §5.3. Recall that a key constraint in effectively using a matrix based method is formulating a robust, ideally dense preference (or outcome) matrix. As a result, a variety of ad-hoc tricks have been developed to use this approach in cases where well defined preference matrices aren't easily formed. For the PageRank scheme, this isn't generally a limitation due to the number of hyperlinks between web pages on the Internet. However, in certain sports applications where the number of participants can outweigh the number of head-to-head matches, obtaining matrix based rankings poses some significant challenges. Keener in [18] has attempted to address some of these issues by implementing several nonlinear adjustments to the preference matrix that generally involve manipulation of game scores in lieu of win/loss outcomes. Now, the *pdf* player models can be used to overcome these limitations by providing realistic, head-to-head match data even for players that have little or possibly no actual match history. This is demonstrated on a smaller scale with the initial application which is a natural extension of simulation results presented with the full description of the stochastic Markov chain *pdf* player model in Section §4.2. A subsequent, more comprehensive analysis is then carried out to provide correlation to the tournament rankings developed in Section §6.1 and to formally document the rankings dynamics for several players of interest.

6.2.1 Initial Application

The first application in evaluating matrix based rankings for tennis builds upon the head-to-head simulations and statistical analyses discussed in Sections §4.2.5 through §4.2.7. Now that a

full background on this ranking methodology has been provided, the previous investigation can be continued based on the availability of head-to-head performance metrics. Additionaly, this implementation forms the baseline for comparing matrix based tennis ranking results to those previously presented using the tournament simulation method. The simulated match results from Section §4.2.5 are used to demonstrate another computational way to obtain rankings with a probabilistic foundation. Consider the 4×4 preference matrix, $A = [a_{ij}]$, constructed for the four players listed in Table 4.2. For this case, each entry of the preference matrix contains the sample mean associated with the probability of the (row) player defeating the (column) player as obtained from a Monte Carlo simulation of $30,000$ head-to-head matches between the two players, implemented using the *pdf* player model. In this case, since the control group only consists of four players, $30,000$ matches were used to ratchet down the statistical variance. Note that each player's score (recall that $s_i = \frac{1}{n_i}\sum a_{ij} r_j$) is the result of their interaction with all the other players (i.e. "the field") and that the score depends both on the outcome of the matches as well as the strength (rank) of the opponents. Since each of the entries a_{ij} are obtained from Monte Carlo simulations with Gaussian distributed inputs obtained from the data, the ranking vector, $\vec{r}$, inherits these desirable features, giving it a natural probabilistic interpretation. Each component of the ranking vector is itself a random variable with mean value r_j. Now, if w_{ij} is the number of wins by player i against player j of n total players, the preference matrix becomes

$$A = \begin{pmatrix} 0 & \frac{w_{ij}}{w_{ij}+w_{ji}} & \dots & \frac{w_{in}}{w_{in}+w_{ni}} \\ \frac{w_{ji}}{w_{ji}+w_{ij}} & 0 & \dots & \frac{w_{jn}}{w_{jn}+w_{nj}} \\ \dots & & & \dots \\ \frac{w_{ni}}{w_{ni}+w_{in}} & \frac{w_{nj}}{w_{nj}+w_{jn}} & \dots & 0 \end{pmatrix} \tag{6.1}$$

Note that A has nonnegative entries and is most assuredly irreducible due its dense construction, hence by the Perron-Frobenius theorem (refer specifically to Section §5.2) the largest eigenvalue of A is used to produce the ranking vector.

Based on the data shown in Table 4.3, a preference matrix is constructed with sample means as entries that indicate the probability that the row player defeats the column player. The rows and columns are as listed in that table. The preference matrix A is built using (6.1) and the associated ranking eigenvector $\vec{r}$ is given by:

$$A = \begin{pmatrix} 0 & 0.3635 & 0.3869 & 0.4794 \\ 0.6365 & 0 & 0.5074 & 0.6073 \\ 0.6131 & 0.4926 & 0 & 0.5883 \\ 0.5206 & 0.3927 & 0.4117 & 0 \end{pmatrix}; \quad \vec{r} = \begin{pmatrix} 0.4298 \\ 0.5572 \\ 0.5472 \\ 0.4550 \end{pmatrix}$$

The ranking produces: (4) Blake = 0.4298; (1) Federer = 0.5572; (2) Nadal = 0.5472; (3) Roddick = 0.4550. The final rank is in parenthesis. Despite the fact that Federer had neither the highest percentage of points won on serve or receive of serve (see data in Table 5.4) he earns the

top ranking in this format. The final ordering of the four players also agrees with their ATP year end rankings also shown in Table 5.4. The individual ranking elements add a level of probabilistic insight when compared to the random and actual draw tournament simulations carried out in Section §4.2.7. Namely, these values show that Federer and Nadal are virtually evenly matched as the two premier players while Roddick and Blake are a ways behind. Although Federer ends up with the top ranking, the values themselves are useful in weighing and separating the field. This is observed with more detail as additional players are included in the analysis.

6.2.2 Implementation of a Larger Field

Next, a larger player field is implemented using the matrix based ranking method with simulated head-to-head winning percentages again used to build the preference matrix. As with the tournament simulation approach described in Section §6.1, the analysis is undertaken for the top 16 ATP ranked players in each period of consideration (after each Grand Slam event and at the end of the ATP season between 2007 and 2009). In this case, $1,000$ matches are simulated between each two distinct players. For 16 players, 120,000 total matches are simulated to obtain a final preference matrix result. This would increase to 8,128,000 total simulated matches for a 128 player review. Applying (6.1) to the top four (16 players used in the actual analysis) players at the end of the 2009 ATP season (R. Federer, R. Nadal, N. Djokovic and A. Murray) results in the following preference matrix:

$$A = \begin{pmatrix} 0 & 0.453 & 0.569 & 0.487 \\ 0.547 & 0 & 0.561 & 0.516 \\ 0.431 & 0.439 & 0 & 0.418 \\ 0.513 & 0.484 & 0.582 & 0 \end{pmatrix} \tag{6.2}$$

Again, individual entries of the preference matrix are acquired from match simulations between players represented at each particular node. For example, a_{12} represents the winning percentage of Federer versus Nadal in $1,000$ matches: 0.453. Conversely, a_{21} is the winning percentage of Nadal against Federer: 0.547. Note that a_{12} and a_{21} sum to 1. This means that conveniently, no scaling of the preference matrix is needed to ensure that the scores ($s_i = \sum a_{ij} r_j$) aren't weighted unfairly towards any specific player. The number of effective total matches played (outcomes) are equal for all players as each participant is matched against every other player in the field. This also ensures that $||\vec{r}|| = 1$.

The winning percentages that are calculated and used for preference matrix node entries incorporate all corrections and adjustments previously described with the computational Markov chain *pdf* player model. Remember that the simulations measuring these winning percentages consider each player's performance (including "the field") in the previous 52-week sliding window. Determination of the dominant eigenvector provides weighted rankings. Due to the computational and fluctuating nature of simulating matches using the four parameter approach, minor variations are observed when calculating the rankings. This is mitigated, however, by the fact that the resulting ranking provides *relative* information between players instead of simply providing a list of ranks

Rank	Player	Value
1	Rafael Nadal	0.2947
2	Andy Murray	0.2895
3	Roger Federer	0.2859
4	Nikolay Davydenko	0.2656
5	Novak Djokovic	0.2653
6	Robin Soderling	0.2644
7	Andy Roddick	0.2636
8	Fernando Verdasco	0.2582
9	Jo-Wilfried Tsonga	0.2420
10	Juan Martin Del Potro	0.2330
11	Marin Cilic	0.2321
12	Fernando Gonzalez	0.2253
13	Gilles Simon	0.2185
14	Tommy Robredo	0.2168
15	Radek Stepanek	0.2166
16	Gael Monfils	0.2089

Table 6.4. Matrix rankings for the 2009 ATP season

as 1, 2, 3 ... etc. Listed in Table 6.4 is the output of the top 16 players from running the matrix ranking method at the end of the 2009 ATP Season.

The probabilistic interpretation of the ranking vector means that players can be grouped by rankings with values of similar magnitude. This type of information can be exploited in a variety of ways. One such example would be to break down the player rankings into “tiers” of performance. With “tiered” sets of players, a pot based system could be incorporated into the seeding system for tournament play. This is a reasonable way to de-emphasize the ranking influence that can guide players in “gaming” their tournament selections. With seeds that are guaranteed only down to a certain range (instead of specific positions) those effects could be reduced. A similar approach is used when developing groups of teams for the round robin stage of World Cup soccer competitions based on the the world rankings of the individual teams.

Player	ATP Rank	Tournament Rank	Matrix Rank
Roger Federer	1	3	3
Rafael Nadal	2	2	1
Novak Djokovic	3	7	5
Andy Murray	4	1	2
Juan Martin Del Potro	5	9	10
Nikolay Davydenko	6	6	4
Andy Roddick	7	4	7
Robin Soderling	8	5	6
Fernando Verdasco	9	8	8
Jo-Wilfried Tsonga	10	10	9
Fernando Gonzalez	11	14	12
Radek Stepanek	12	15	15
Gael Monfils	13	16	16
Marin Cilic	14	11	11
Gilles Simon	15	12	13
Tommy Robredo	16	13	14

Table 6.5. Composite rankings for 2009 ATP season

6.2.3 Rankings Dynamics

Year end rankings for the 2009 season are shown for the ATP, Monte Carlo tournament simulations and the Monte Carlo based matrix method in Table 6.5.

An interesting correlation exists between all three rankings in the table. The largest variation in rank is five positions exhibited between Del Potro's matrix rank and his ATP rank. Aside from that disparity, the average difference in all player's tournament and matrix rankings as compared to their ATP rank is three positions. On one hand then, a good relationship exists between the ranks developed using the *pdf* player model. On the other hand, however, the potential ramifications of even a single position change have to be considered due to the seeding implications in the tournament format comprising the ATP circuit. Further, both the tournament rankings and the matrix ranks provide details on the proximity of the ranks when considering the individual

elements of $\vec{r}$. For example, in Table 6.2, the points accrued by Federer at position three can be considered "out of family" with the points accrued by Murray and Nadal at positions one and two, respectively. This further accentuates the fact that in that case, Murray and Nadal are the bonafide top two players with Federer and the rest of the field trailing behind.

Composite rankings have been computed in the same fashion for each 52-week rolling period after each Grand Slam event and at the end of the ATP season from the start of 2007 to the end of 2009. These rankings are presented in Figure 6.2 - Figure 6.5 for Federer, Nadal, Murray and Roddick. The figures display the eleven unique data points for evaluation of ranking trends for these each of these methods.

Interestingly, both Federer and Nadal show consistency in all rankings between 2007 and 2009. In fact, for Federer, both tournament and matrix based rankings were identical for the entire 2007-2009 timeframe. Note that Federer moved briefly out of the top spot in 2008 before regaining number one status in the ATP while both tournament and matrix rankings showed movement in the opposite direction. This downward movement in ranking foreshadowed Federer's actual ATP drop to the number three and four spots in 2010-2011. Andy Murray is also an interesting case as all three ranking formats highlighted his improved performance up through 2009. Although he was limited by ATP rank (position four) at the end of 2009, both tournament and matrix rankings had him in a higher spot at rank two. Notably, the ATP rankings had shown Murray in the second position for a few weeks in 2009 with a slip back down to spot four by the end of the season. During that period (Oct-Dec), Murray went (9-2) with losses only to Stepanek (11-6) and Federer (7-4). This alone highlights the limitation of the point based system

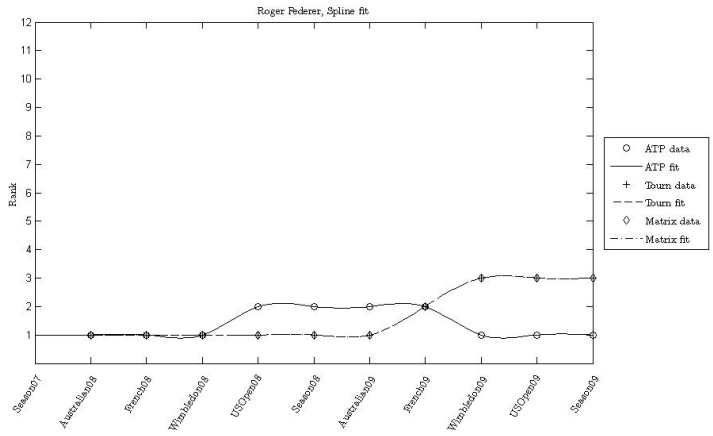


Figure 6.2. Composite 2007-2009 rankings for ATP, Tournament and Matrix based methods with spline fit for R. Federer

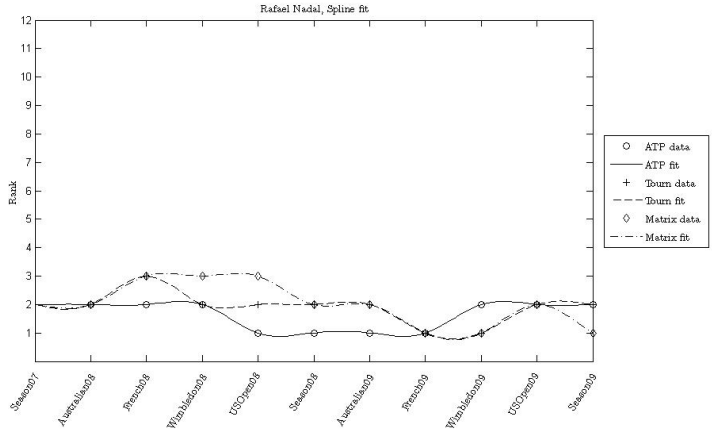


Figure 6.3. Composite 2007-2009 rankings for ATP, Tournament and Matrix based methods with spline fit for R. Nadal

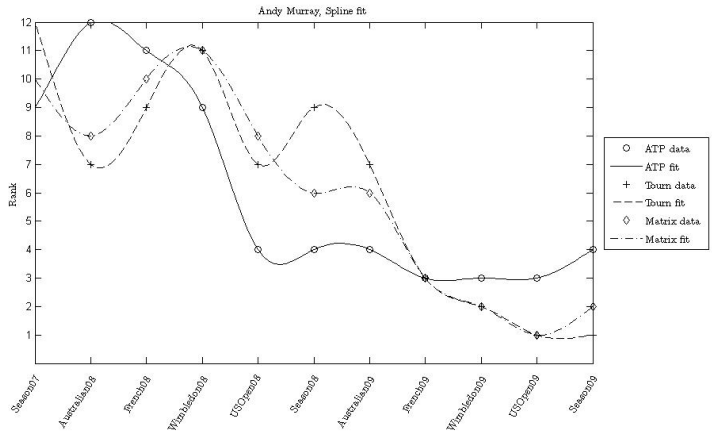


Figure 6.4. Composite 2007-2009 rankings for ATP, Tournament and Matrix based methods with spline fit for A. Murray

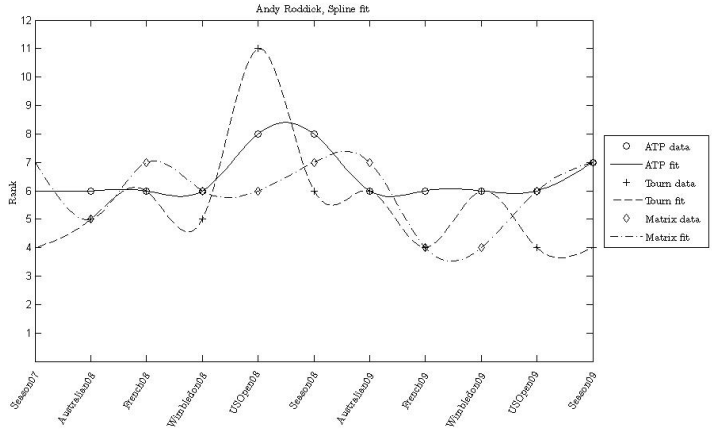


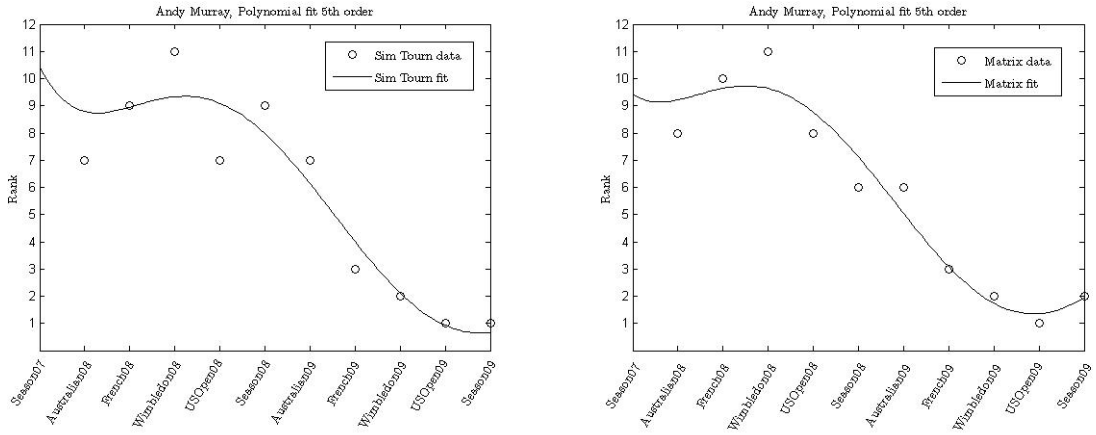
Figure 6.5. Composite 2007-2009 rankings for ATP, Tournament and Matrix based methods with spline fit for A. Roddick

utilized by the ATP as it can easily be argued that Murray's ranking is more realistically communicated by the methods presented here with their probabilistic basis and thus, their probabilistic interpretation.

Moreover, there's a variety of ways to interpret this information. One particularly interesting way is by using spline curve fitting also shown in Figure 6.2 - Figure 6.5. There are a couple of key reasons for this. First, the spline fit facilitates a meaninful overlay for the ATP, tournament and matrix rankings on the same plot. The spline fit provides a smooth graph through the individual ranking data points with visual representation of the entire period of performance, 2007-2009. This curve can be projected to provide insight into the *rank trend*. Another way to view these data is through a polynomial curve fit. With the polynomial approach, the trend line isn't constrained to the individual ranking points and the influence of outliers can be limited. Due to the dynamic nature of the trend fitting, overlay plots aren't as easily interpreted. Polynomial curve fits (5th order) are shown in Figure 6.6 for Murray and Figure 6.7 for Roddick.

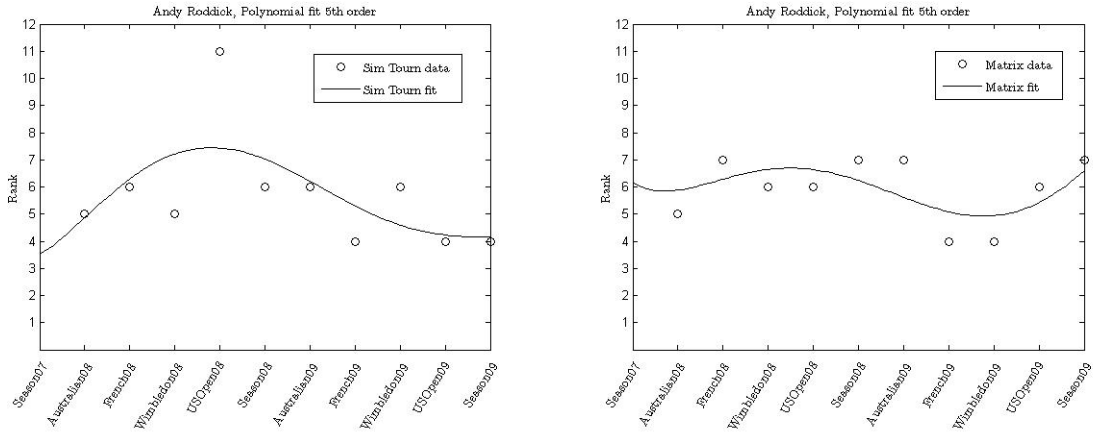
The fluctuation of tournament and matrix rankings for these two players result in the most interesting case.

Ultimately, these ranking trends offer an informative visualization in predicting a player's future performance. In the cases described, it can also provide further validation on perceived fairness of the rankings. Considering the relative movement of the trend portion of the curves and the robust head-to-head information that formulates the tournament and matrix rankings, the probabilistic basis of these methods is graphically depicted.



(a) (b)

Figure 6.6. 2007-2009 rankings for Tournament and Matrix based methods with 5th order polynomial fit for A. Murray



(a) (b)

Figure 6.7. 2007-2009 rankings for Tournament and Matrix based methods with 5th order polynomial fit for A. Roddick

6.3 Discussion

Two distinct ranking methodologies have been addressed here. The first, using a tournament simulation methodology builds upon the framework for ATP rankings and adjusts to a more equitable distribution of points based on tournament progression. This method includes a significantly higher number of match results and hence much more data are brought into consideration than that of the ad-hoc point based system in the ATP rankings. To be more specific, in a 52-week period, the ATP system only includes the information from roughly 3,000 matches worth of data, whereas here, the results incorporate approximately a thousand times more data between any two particular players. For matrix based systems, major limitations in the sparsity of data available (or used) are addressed. The *pdf* player model has provided the foundation that bridges the gap between actual, observed match results to a complete, full set of random variables that describe the expected match outcome between any two players. In fact, the *pdf* player model has been correlated so well to analytical and actual real world results that these random variables can be developed with great confidence even for players that have never played a head-to-head match. Ultimately, these two ranking systems build upon the stochastic Markov chain *pdf* player model in unique ways. While great detail has been provided on the ranking outcomes, it's important to note that the specific ranking methodology used isn't the prime result. Rather, the predictive framework that builds upon the mathematically sound, robust player models is the key contribution. Subsequent manipulation of the nuances provided can supply a variety of other meaningful ranking systems. However, the most important facet to providing that predictive framework is the effective use of the Markov chain *pdf* player models.

Chapter 7.

Conclusion

The use of Monte Carlo analyses has risen exponentially with modern day computing power. Using these techniques, computational analysis is possible in new fields of study that were previously limited to analytical approaches. A full treatment on the Monte Carlo method and its implementation has been provided. Specifically, when applied to tennis and its associated ranking system, the Monte Carlo method has revealed *fundamental limitations* in the way tennis players are evaluated and ultimately compared. Most notably, the lack of head-to-head match data for the majority of the players invariably leads to complicated questions on interpretations of player standings with respect to the entire field.

The concept of addressing weaknesses in traditional methods of providing valuations is relatively new but highly successful for a wide array of applications. The critical issue is prioritization of the information that's most meaningful for a given function. One example of this is documented in the book (and subsequently the film) *Moneyball* by Lewis [24] that characterizes the flaws and preconceptions under which Major League Baseball players had been evaluated by the majority of scouts and managers. Using an unbiased, objective and mathematically driven

approach for defining player performance, the Oakland Athletics were able to field very successful teams that could compete with larger market teams like the New York Yankees and Chicago Cubs for less than a third of the cost in payroll. This approach led to what many consider a new era in the sport of baseball where sabermetrics, the analysis of baseball through objective statistical evidence, has redefined the way professional teams are fielded. The general concept that arguably revolutionized baseball can potentially be understood and exploited in other fields such as economic financial analysis, internet searching and even in the fields of biology and medicine.

Similary, the stochastic Markov chain *pdf* player model has led to a more accurate and mathematically sound way of modeling and evaluating tennis players through the use of probability density functions. The *pdf* player models have the distinct advantage of allowing individual player performance to be *baselined* against the "the field" of all other players. This means that a fair comparison between a player with 10 or 100 career matches can essentially be made with another who has played several thousand. In clearly delimited steps, the *pdf* player model has first been exhaustively linked to the analytical model of Newton and Keller (see [34]). Subsequently, a slew of actual ATP player match and tournament results were found to correlate extremely well to the model. Next, several second order effects were explored including the impacts of non-iid probabilistic modeling. Finally, development of rankings with a *probabilistic* basis were undertaken. In all facets, the stochastic Markov chain *pdf* player model has demonstrated its capability as a robust, expandable tool for evaluating player performance from individual matchups all the way through full ranking comparisons encompassing the entire field.

References

[1] S. C. Albright. A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association*, 88(424):1175–1183, Dec 1993.

[2] K. J. Arrow. *Social Choice and Individual Values*. Yale University Press, 2nd edition, 1970.

[3] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer-Verlag, 2007.

[4] T. Barnett and S. R. Clarke. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16(2):113–120, 2005.

[5] J. S. Bendat and A. G. Piersol. *Random Data: Analysis & Measurement Procedures*. John Wiley & Sons, 2nd edition, 1986.

[6] K. Bryan and T. Leise. The $25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM Review*, 48(3):569–581, 2006.

[7] T. Callahan, P. J. Mucha, and M. A. Porter. The Bowl Championship Series: A mathematical review. *Notices of the American Mathematical Society*, 51:887–893, Sep 2004.

[8] W. H. Carter and S. L. Crews. An analysis of the game of tennis. *The American Statistician*, 28(4):130–134, Nov 1974.

[9] S. R. Clarke and D. Dyte. Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, 7(6):585–594, 2000.

[10] P. D. Coddinton. Analysis of random number generators using Monte Carlo simulation. *International Journal of Modern Physics*, 547(5), 1994.

[11] W. Colley. Colley's bias free college football ranking method: The Colley matrix explained. *Ph.D dissertation*, 2002.

[12] R. Cross. Tennis physics, anyone? *Physics Today*, pages 84–85, Sept 2008.

[13] J. Eichenauer and J. Lehn. A non-linear congruential psuedo random number generator. *Statistiche Hefte*, 27:315–326, 1986.

[14] A. E. Elo. *The Rating of Chess Players, Past & Present*. Arco Pub., 1st edition, 1978.

[15] M. E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, 48:377–394, 1999.

[16] M. E. Glickman. Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28:673–689, 2001.

[17] D. Jackson and K. Mosurski. Heavy defeats in tennis: Psychological momentum or random effect? *Chance*, 10(2):27–34, 1997.

[18] J. P. Keener. The Perron-Frobenius theorem and the ranking of football teams. *SIAM Review*, 35(1):80–93, 1993.

[19] M. G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 48:303–312, 1939.

[20] 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. *Journal of the American Statistical Association*, 96(454):500–509, Jun 2001.

[21] P. L'Ecuyer. *The Handbook of Simulation*, chapter Random Number Generation, pages 93–137. Wiley, 1998.

[22] H. Leeb and S. Wegenkittl. Inversive and linear congruential psuedorandom number generators in empirical tests. *ACM Transactions on Modeling and Computer Simulation*, 7(2):272–286, 1997.

[23] D. H. Lehmer. Mathematical methods in large-scale computing units. In *Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery*, pages 141–146, Cambridge, Massachusetts, 1949. Harvard University Press.

[24] M. Lewis. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company, 1st edition, 2003.

[25] J. R. Magnus and F. J. G. M. Klaassen. The effect of new balls in tennis: Four years at Wimbledon. *The Statistician*, 48(2):239–246, 1999.

[26] J. R. Magnus and F. J. G. M. Klaassen. The final set in a tennis match: Four years at Wimbledon. *Journal of Applied Statistics*, 26(4):461–468, 1999.

[27] 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(2):247–256, 1999.

[28] M. Mascagni, S. A. Cuccaro, D. V. Pryor, and M. L. Robinson. A fast, high quality, and reproducible parallel lagged Fibonacci psuedorandom number generator. *Journal of Computational Physics*, 119(2):211–219, 1995.

[29] N. Metropolis. *From Cardinals to Chaos. Reflections on the Life and Legacy of Stanislaw Ulam*, chapter The Beginning of the Monte Carlo Method, pages 125–130. Cambridge University Press, 1984.

[30] C. Morris. *Optimal Strategies in Sports*, chapter The Most Important Points in Tennis, pages 131–140. North-Holland, 1977.

[31] F. Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:3–9, 1951.

[32] P. K. Newton and K. Aslam. Monte Carlo tennis. *SIAM Review*, 48(4):722–742, 2006.

[33] P. K. Newton and K. Aslam. Monte Carlo tennis: A stochastic Markov chain model. *Journal of Quantitative Analysis in Sports*, 5(3), 2009.

[34] P. K. Newton and J. B. Keller. The probability of winning at tennis I. Theory and data. *Studies in Applied Mathematics*, 114:241–269, 2005.

[35] P. K. Newton and G. H. Pollard. Service neutral scoring strategies for tennis. In *Proceedings of the Seventh Australasian Conference on Mathematics and Computers in Sports*, pages 221–225, Massey University, Palmerston North, New Zealand, 2004.

[36] H. Niederreiter. On a new class of pseudorandom numbers for simulation methods. *Journal of Computational and Applied Mathematics*, 56:159–167, Dec 1994.

[37] Association of Tennis Professionals. ATP world tour rulebook chapter IX. Technical report, ATP, 2009.

[38] A. J. O'Malley. Probability formulas and statistical analysis in tennis. *Journal of Quantitative Analysis in Sports*, 4(2), 2008.

[39] G. H. Pollard. An analysis of classical and tie-breaker tennis. *Austrailian Journal of Statistics*, 25:496–505, 1983.

[40] S. N. Rasband. *Chaotic Dynamics of Nonlinear Systems*. Wiley, 1st edition, Jan 1990.

[41] Lord Rayleigh. On James Bernoulli's theorem in probabilities. *Philosophical Magazine*, 47:246–251, 1899.

[42] S. M. Ross. *A First Course in Probability*. Macmillan, 2nd edition, 1976.

[43] R. T. Stefani. Survey of the major world sports rating systems. *Journal of Applied Statistics*, 24(6):635–646, Dec 1997.

[44] H. S. Stern and C. N. Morris. A statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association*, 88(424):1189–1194, Dec 1993.

[45] M. Stob. A mathematicians guide to popular sports: Supplement. *The American Mathematical Monthly*, 91(5):277–282, May 1984.

[46] A. Tversky and T. Gilovich. The cold facts about the 'hot hand' in basketball. *Chance*, 2(1):16–21, 1989.

[47] J. von Neumann and R. D. Richtmyer. Statistical methods in neutron diffusion. Technical Report LAMS-551, Los Alamos Scientific Laboratory, 1947.

[48] N. R. Wagner. The logistic equation in random number generation. In *Proceedings of the Thirtieth Annual Allerton Conference on Communications, Control and Computing*, pages 922–931, University of Illinois at Urbana-Champaign, 1993.

[49] M. Walker and J. Wooders. Minimax play at Wimbledon. *American Economic Review*, 91(5):1521–1538, 2001.

[50] T. Warnock. *From Cardinals to Chaos. Reflections on the Life and Legacy of Stanislaw Ulam*, chapter Random-Number Generators, pages 137–141. Cambridge University Press, 1984.

[51] S. Wolfram. A new kind of science. Technical Report 1098, Wolfram Media, Champaign, IL, 2002.

Appendix.

Field Performance, 2007-2009

Table A.1 below lists the full complement of performance parameters for "the field" between 2007 and 2009 after each Grand Slam tournament and at the end of the season. Performance metrics are obtained using the same 52-week sliding period of performance used by the ATP. For example, the statistics listed for the 2008 Australian Open reflects the combined performance of "the field" for the 2007 Australian Open through (but not including) that event for 2008.

	μ_f^s	μ_f^r	σ_f
2007 Season	0.63316	0.36684	0.094779
2008 Australian Open	0.63401	0.36599	0.094160
2008 French Open	0.63407	0.36593	0.100550
2008 Wimbledon	0.63294	0.36706	0.100950
2008 US Open	0.63300	0.36700	0.100370
2008 Season	0.63309	0.36691	0.100790
2009 Australian Open	0.63322	0.36678	0.100930
2009 French Open	0.63342	0.36658	0.092389
2009 Wimbledon	0.63425	0.36575	0.092525
2009 US Open	0.63485	0.36515	0.092281
2009 Season	0.63422	0.36578	0.092792
2007-2009 Seasons	0.63349	0.36651	0.096238

Table A.1. Field performance parameters for 2007-2009

Histograms that graphically illustrate the performance parameters listed in Table A.1 along with the associated truncated Gaussian curve fit are shown below.

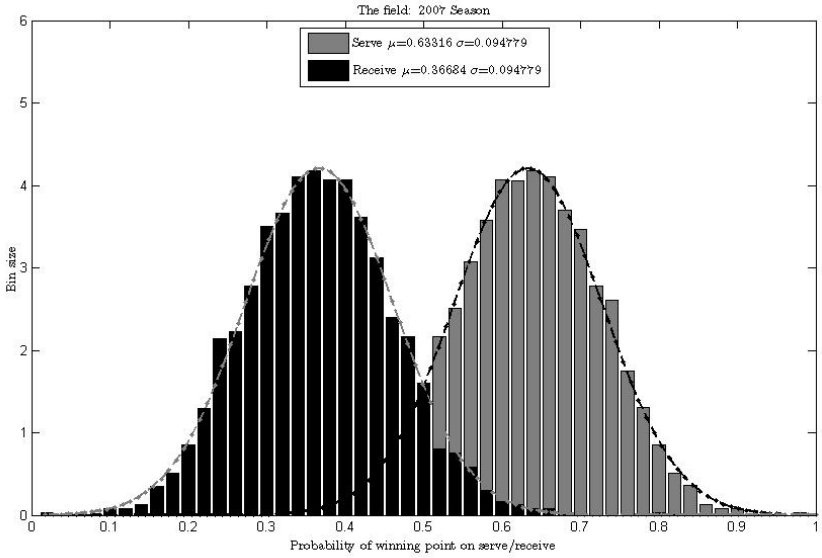


Figure A.1. Field performance: 2007 Season

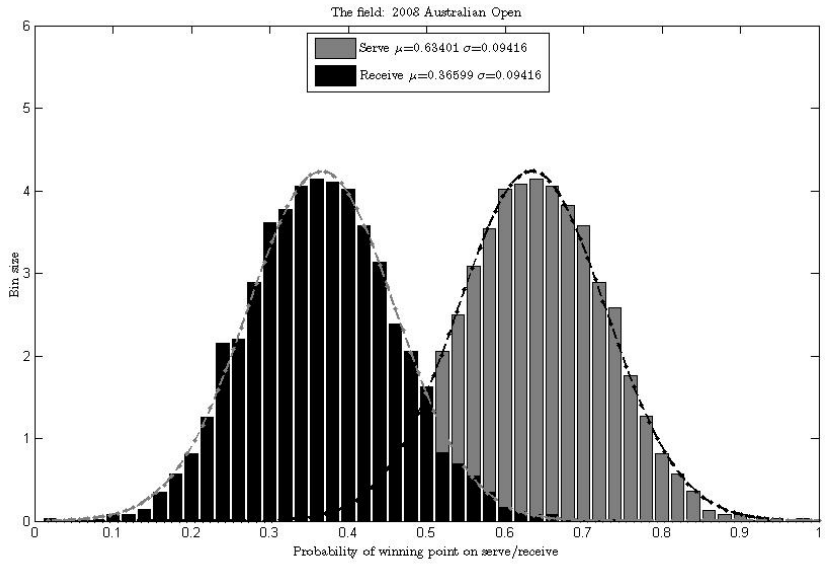


Figure A.2. Field performance: 2008 Australian Open

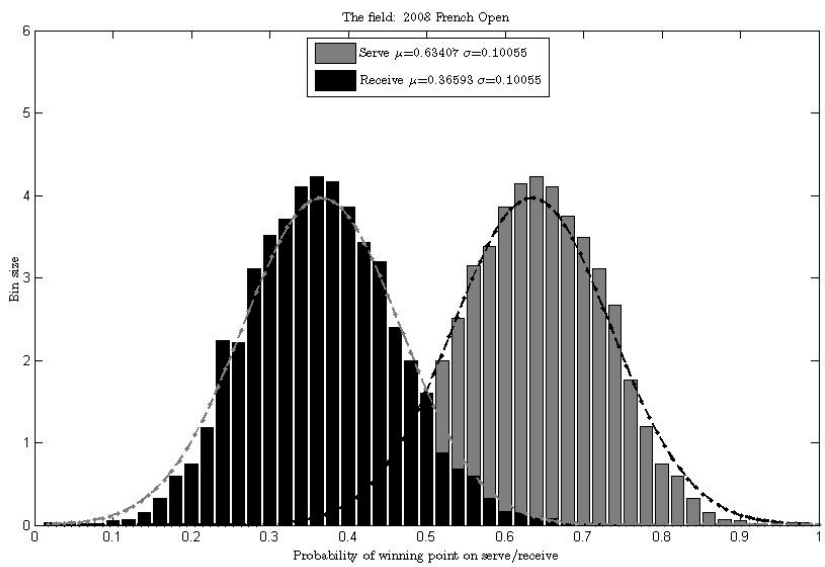


Figure A.3. Field performance: 2008 French Open

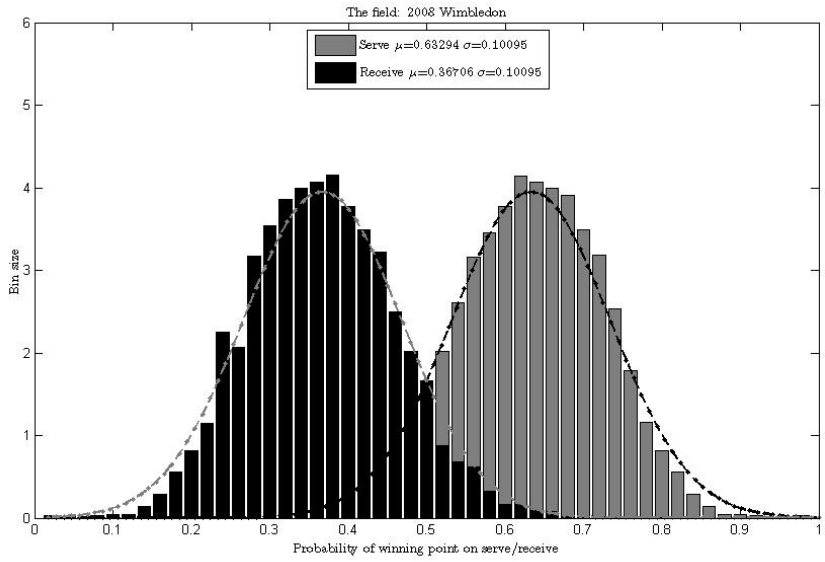


Figure A.4. Field performance: 2008 Wimbledon

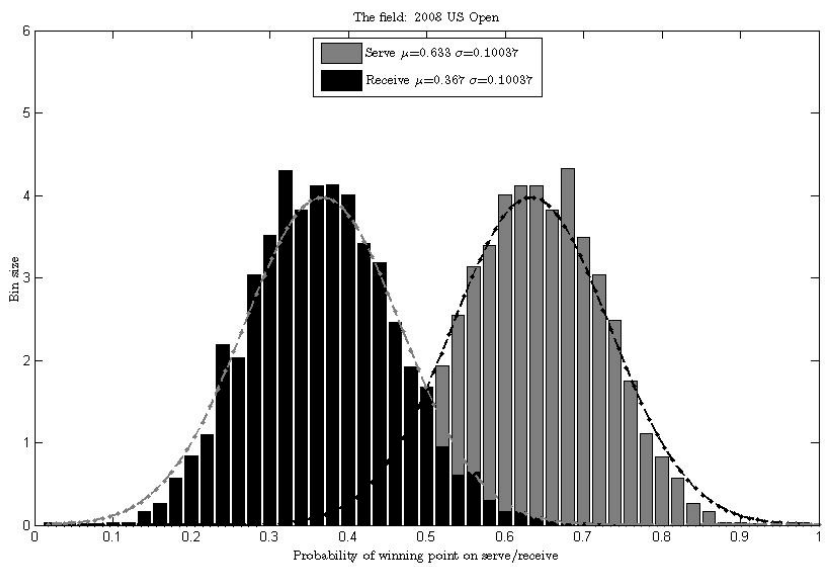


Figure A.5. Field performance: 2008 US Open

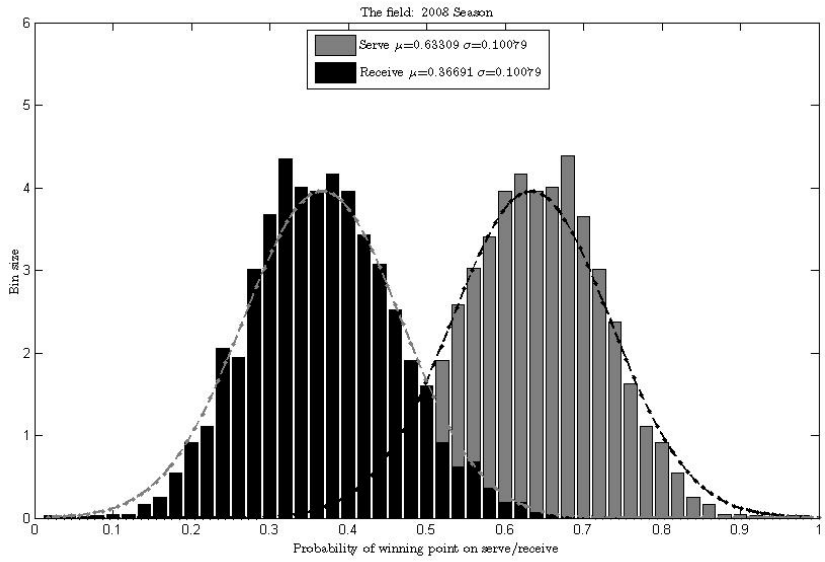


Figure A.6. Field performance: 2008 Season

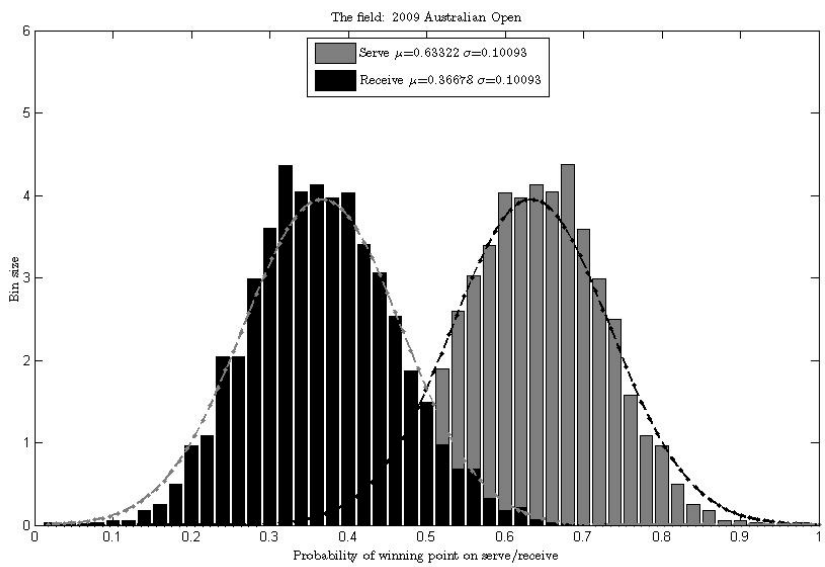


Figure A.7. Field performance: 2009 Australian Open

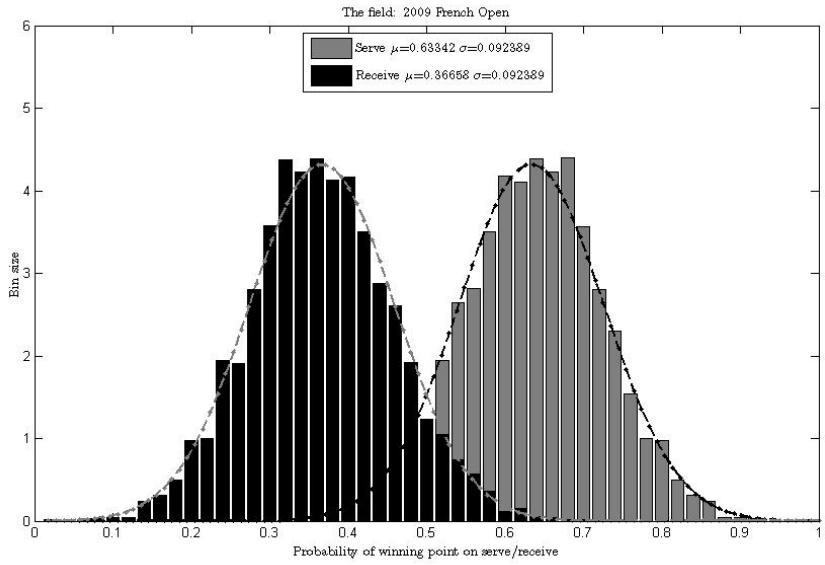


Figure A.8. Field performance: 2009 French Open

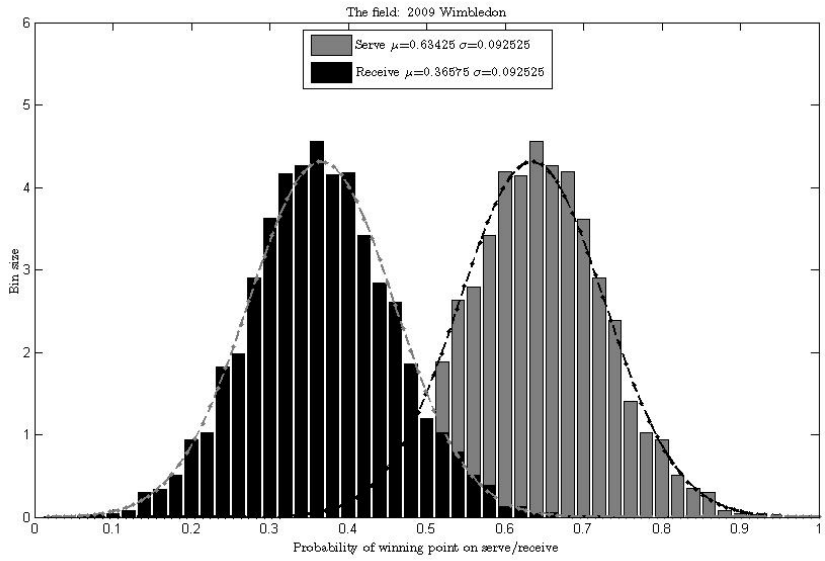


Figure A.9. Field performance: 2009 Wimbledon

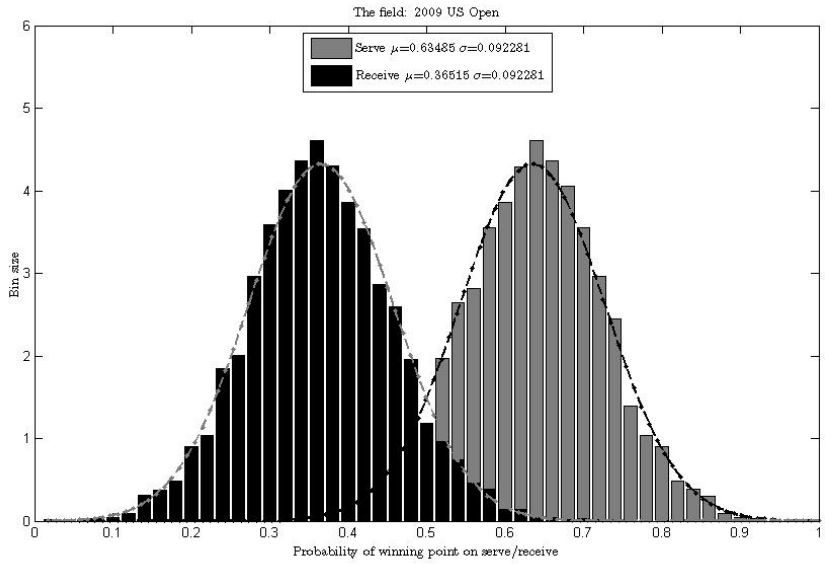


Figure A.10. Field performance: 2009 US Open

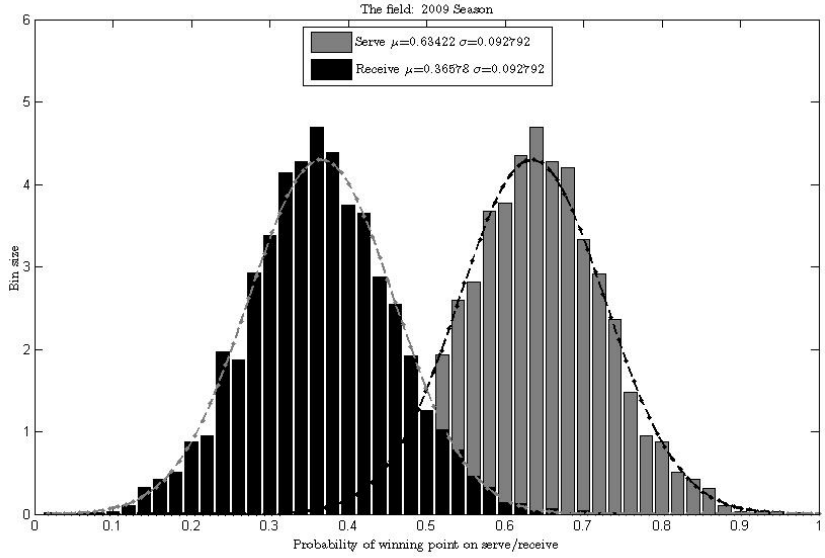


Figure A.11. Field performance: 2009 Season

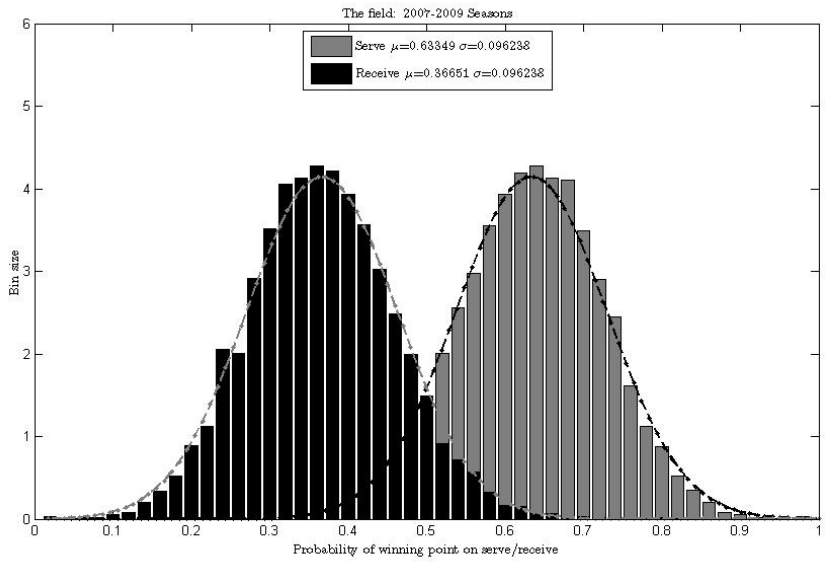


Figure A.12. Field performance: 2007-2009 Seasons