Random Walks with Memory Applied to Grand Slam Tennis Matches Modeling

Tomáš Kouřim

Institute of Information Theory and Automation, Czech Academy of Sciences Prague, Czech Republic
kourim@outlook.com

Abstract

The contribution presents a model of a random walk with varying transition probabilities implicitly depending on the entire history of the walk, which is an improvement of a model with varying step sizes. The transition probabilities are altered according to the last step of the walker using a memory parameter to either reward or punish success by increasing or decreasing its probability in the next step. This walk is applied to model Grand Slam tennis matches and fitted on their entire history since 2009. The suitability of the model is thoroughly tested on a number of real datasets. The model seems to be robust and describe well the majority of matches, making it an useful tool to produce precise *in-play* odds.

1 Introduction

Tennis is one of the most popular sports both on professional and amateur level. Millions of people pursue tennis as their leisure time activity [3] and same numbers hold also for the people following and watching the professional tennis competitions. Tennis also plays a major role in the sports betting industry, which grows rapidly and becomes more and more important part of the global economy. In the Czech Republic only, the total sales in sports betting industry reached CZK 64.5 billion (2.9 billion USD) in 2017, representing 1.3% of Czech GDP [8]. The immense size of the betting market attracts also many fraudsters. The European Sports Security Association regularly reports on suspicious betting activities, the latest report (2018) contained 267 cases of such activity, 178 (67%) in tennis [1]. It is thus obvious that a precise model describing the game of tennis has many possible uses in real life.

Tennis is also a sport more than suitable to be modeled using random walks or random processes in general, as it naturally consists of many such processes. A series of tennis matches is a random walk, the sequence of sets within a match, games within a set, points within a game or even strokes within a point can be all considered a random process and modeled using a random walk. Additionally, these walks are well described by the tennis rules and there exist lots of data describing these random processes (i.e. various tennis result databases provided by the tennis federation as well as many private subjects). In this paper, the random walk consisting of a sequence of sets within a match is studied. Matches played as a *best-of-five*, i.e. the men Grand Slam tournaments, are considered in this paper. In these matches, up to 5 steps of the random walk can be observed, making them more suitable than the *best-of-three* games, where maximum 3 steps can occur.

The matches are modeled using a new type of a recently introduced random walk with varying probabilities [6], which is a modification of a random walk with varying step size introduced by Turban [9]. It seems more than suitable to model tennis matches as the data suggest that a success in tennis yields another success, or in other words, that winning one particular part of the match increases the chances of winning the next part as well. This behavior is well described by the new random walk model.

The paper is organized as follows. Next chapter introduces the new type of random walk used for tennis modeling. Section 3 provides general description of the data used, Section 4 shows how to obtain starting probabilities. In Chapter 5 the actual model is described and its performance is evaluated. Section 6 concludes this paper.

2 Random walk with varying probability

In 2010, Turban described [9] a new version of a random walk with memory, where the memory is introduced using variable step size. This idea was further extended by Kouřim [6, 7] and an alternative version of a random walk with memory was introduced, where the memory affects the walk through varying transition probabilities.

The walk evolves in a following way. Initial step is made following the result of a Bernoulli random variable with starting probability parameter p_0, that is,

$$P(X_1 = \text{“}right\text{”}) = p_0.$$

From the second step on, the transition probability in the $t-th$ step is given by

$$X_{t-1} = \text{“}right\text{”} \implies P(X_t = \text{“}right\text{”}) = \lambda p_{t-1}$$

$$X_{t-1} = \text{“}left\text{”} \implies P(X_t = \text{“}right\text{”}) = 1 - \lambda(1 - p_{t-1})$$

for some $\lambda \in (0, 1)$. When the directions are formalized so that “$right$” ≈ 1 and “$left$” ≈ -1, the formula for the $t-th$ transition probability can be rewritten as

$$p_t = \lambda p_{t-1} + \frac{1}{2}(1 - \lambda)(1 - X_t). \tag{1}$$

This definition of a random walk means that the opposite direction is always preferred and that the walk tends to return back to the origin. Alternatively, inverse approach can be applied and the same decision can be supported. Formally, the expression for the $t-th$ transition probability is then

$$p_t = \lambda p_{t-1} + \frac{1}{2}(1 - \lambda)(1 + X_t). \tag{2}$$

For more details on the walk and its rigorous definition, see the original papers [6, 7].

3 Data description

For the purpose of this study, a database containing the results from all Grand Slam tournaments from 2009 to 2018 and corresponding Pinnacle Sports bookmaker's odds[1] was created based on the information publicly available from website www.oddsportal.com. There are 4 Grand Slam[2] tournaments each year, 40 tournaments together. Each Grand Slam has 128 participants playing in a single-elimination system (i.e. 127 games per tournament), making it a set of 5080 games together. However, the games where either one of the players retired were omitted from the dataset and so were the matches where no bookmaker's odds were available. Together there were 4255 matches with complete data available, presenting total 432 players.

[1] This bookmaker is considered leading in the sports betting industry.
[2] Australian Open, French Open, The Wimbledon and US Open.

The most active player was Novak Djokovic, who participated in 188 matches. On average, each player played 19.7 matches, with the median value of 8 matches played. The most common result was 3:0, occurring 2138 times, on the other hand, 5 sets were played only 808 times.

The order in which the players are listed is rather random. The first listed players[3] won 2201 in total, just slightly over the half. On the other hand, if the bookmaker's favorite (i.e. the player with better odds or the first listed player in case the odds are even) is considered, the situation changes significantly. The favorites won 3307 matches in total, mostly 3:0, and lost 311 times 0:3, 347 times 1:3 and 290 times 2:3. It suggests that bookmaker's odds can be used as a probability estimate, which is in accordance with previous results, for example [5].

4 Initial probability derivation

The model of a random walk with varying probabilities described in Section 2 takes two parameters, initial set winning probability p_0 and the memory coefficient λ. Finding the optimal value of λ is the main subject of this paper and is described in Section 5.

Estimating the initial set winning probability is a major task by itself and represents one of the elementary problems in tennis modeling. For the purpose of this article an estimation based on bookmaker's odds will be used. Specifically, the closing odds[4] by Pinnacle Sports bookmaker for the first set result are used to estimate the probabilities of each player winning the first set, i.e. p_0 and $1-p_0$. Such odds represent a good estimation of the underlying winning probability and are considered as a baseline in the sports betting industry. The odds, however, have to be transformed into probabilities. A method described in [4] is used to obtain probabilities, using a parameter $t \in [0, 1]$ set to the value $t = 0.5$. Obtained first set winning probabilities are then used as a given starting probability p_0 in the random walk.

5 Model description and evaluation

5.1 Model description

Original inspiration of the random walk described in Section 2 is based on intensive study of historical sport results and their development. The data suggest that the probability of success (i.e. scoring, winning a set or a point etc.) evolves according to the random walk with varying probabilities. Moreover, it follows from the data that sports can be very roughly divided into two categories. Sports played for a certain amount of time, such as soccer or ice-hockey, evolve according to the walk defined by expression 1. On the other hand, sports where there is necessary to achieve certain number of points, such as tennis or volleyball, appear to follow the pattern defined in equation 2. Therefore the later approach is used to model a tennis game.

The model is used to predict the winning probabilities of sets 2 through 5 and is constructed in a following manner. For each match, the first set winning probability of Player A[5], p_0, is given (see Section 4) and a coefficient λ is fixed for the entire dataset. In order to compute the second set winning probability[6], the result of the first set is observed and second set winning

[3] Such player/team would be normally considered as "home", however, as there are (usually) no home players on the international tournaments, the order is based on the www.oddsportal.com data and/or the respective tournament committees.

[4] Closing odds means the last odds available before the match started.

[5] The player which is listed first in the database, see Section 3 for details.

[6] Winning probability of Player A is always considered as Player B winning probability is just the complement.

Year	Optimal lambda
2010	0.8074
2011	0.8497
2012	0.8142
2013	0.9162
2014	0.8523
2015	0.8429
2016	0.8920
2017	0.8674
2018	0.8333

Table 1: Optimal values of the coefficient λ for respective years.

probability is computed using equation 2. This procedure is repeated for all remaining sets played.[7]

5.2 Model evaluation

In order to verify the model's accuracy, several tests were performed. First, the dataset was divided into training and testing sets. The division can be done naturally by the order of games played. Given a specific time, past matches constitute to a training set, future matches to a testing set. For the purpose of this paper, the split was done on a yearly basis, the data from one previous tennis season were used as a training set to predict winning probabilities in the following season, considered the testing set (i.e. 2010 was the first season used as testing data, 2017 was the last season used as training data), making it 9 training/testing splits together. Another approach to dataset splitting is to consider data from all previous years as testing data and from one future year as training data, however, previous study shows that the difference between these two approaches is negligible [5].

Next step in model verification is the estimation of parameter λ. Training sets and maximal-likelihood estimates were used for this task. The likelihood function is defined as

$$L = \prod_{i=1}^{N_{train}} (x_i p_i + (1 - x_i)(1 - p_i)),$$

where N_{train} is the number of sets 2 thru 5 played in the training dataset, p_i is Player A's winning probability in the $i-th$ set obtained using equation 2 for each match, and x_i is the result of the $i-th$ set, $x_i = 1$ if Player A won the $i-th$ set, $x_i = 0$ otherwise. For computational reasons the *log-likelihood* $L_l = log(L)$ was used, i.e. the function

$$L_l = \sum_{i=1}^{N_{train}} log(x_i p_i + (1 - x_i)(1 - p_i))$$

was maximized. Numerical methods implemented in Python library SciPy were used to obtain specific values of λ. The optimal values of the coefficient λ can be seen in Table 1.

Finally, the model was used to predict set winning probabilities of the unseen data from the training set using initial bookmaker derived odds, equation 2 and memory parameter λ

[7] There can be either 3, 4 or 5 sets played in total in a *best-of-five* tennis game.

obtained from the corresponding training set. In order to verify the quality of the model, the average theoretical set winning probability of Player A $\hat{p} = \frac{1}{n}\sum_{i=1}^{N_{test}} p_i$ and its variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{N_{test}} p_i(1-p_i)$ were computed and so was the observed Player A winning ratio $\bar{x} = \frac{1}{n}\sum_{i=1}^{N_{test}} x_i$. Using the Lyapunov variant of Central Limit Theorem [2], the resulting random variable y follows the standard normal distribution

$$y = \frac{\sqrt{N_{test}}(\bar{x} - \hat{p})}{\hat{\sigma}} \sim \mathcal{N}(0,\, 1).$$

Then ow to verify the model accuracy, the the null hypothesis that the true average Player A set winning probability $\bar{p}$ equals $\hat{p}$ against the alternative hypothesis $\bar{p} \neq \hat{p}$ was tested. Formally,

$$H_0 : \bar{p} = \hat{p}$$

$$H_1 : \bar{p} \neq \hat{p}.$$

One of the assumptions of the CLT is that the observed random variables are independent. This is obviously not true in the case when N_{test} contains all sets from the testing data. Quite the opposite, the model assumes that the winning probability of a set directly depends on the winning probability of the previous set. This can be easily solved by splitting the testing dataset into 4 subsets containing only results from single set of each match, i.e. sets 2, 3, 4 and 5 (if they were played). The matches can be considered independent from each other and so can be the $i-th$ sets of respective matches.

Using this approach, there are 36 testing sets[8] together. On a 95% confidence level, only on 2 out of the 36 available subsets provide strong enough evidence to reject the null hypothesis. On the other hand, the null hypothesis is relatively weak. It only says that the prediction is correct on average. In order to verify the quality of the predictions, more detailed tests have to be created. This can be done primarily by testing the null hypothesis on many subsets created according to some real life based criteria. The natural way how to create such subsets is dividing the matches to the 4 different tournaments. This refining yields 180 subsets[9] altogether. Using 95% confidence level, only 6 of the 180 subsets have data strong enough to reject the null hypothesis. It is worth mentioning that the size of some of the datasets regarding fifth sets is only slightly above 20 observations, which can interfere with the assumptions justifying the use of Central Limit Theorem.

To further analyze the robustness of the model it is important to realize the structure of the data. So far, the player, whose winning probability was estimated, was chosen arbitrarily based on some external (more or less random) order. As such, the observed winning probability in every subset equals approximately to $\frac{1}{2}$, see further Section 3. In such a dataset it is not very difficult to estimate the average winning probability. The situation changes if the bookmaker's favorite is considered for predictions (more details on who is the favorite and how to choose him in Section 3). Performing the same tests as described in the previous paragraph the data allows to reject the null hypothesis (at 95% confidence level) on 5 subsets containing all tournaments and 8 single tournament subsets (out of 180 subsets total).

Finally, the testing data can be divided into groups using the initial probability p_0. Such a division is based on an assumption that the matches with similar bookmaker odds should have similar development. The matches are divided into 5 groups, each containing 10 percentage points in first set winning probability. Except for the biggest favorites (with first set winning probability over 90%), this division seems reasonable. The data histogram can be seen on Figure

[8] Up to 4 sets considered in each match, 9 yearly testing datasets.
[9] 4 sets in each match evaluated, 4+1 tournaments every year, 9 years for testing.

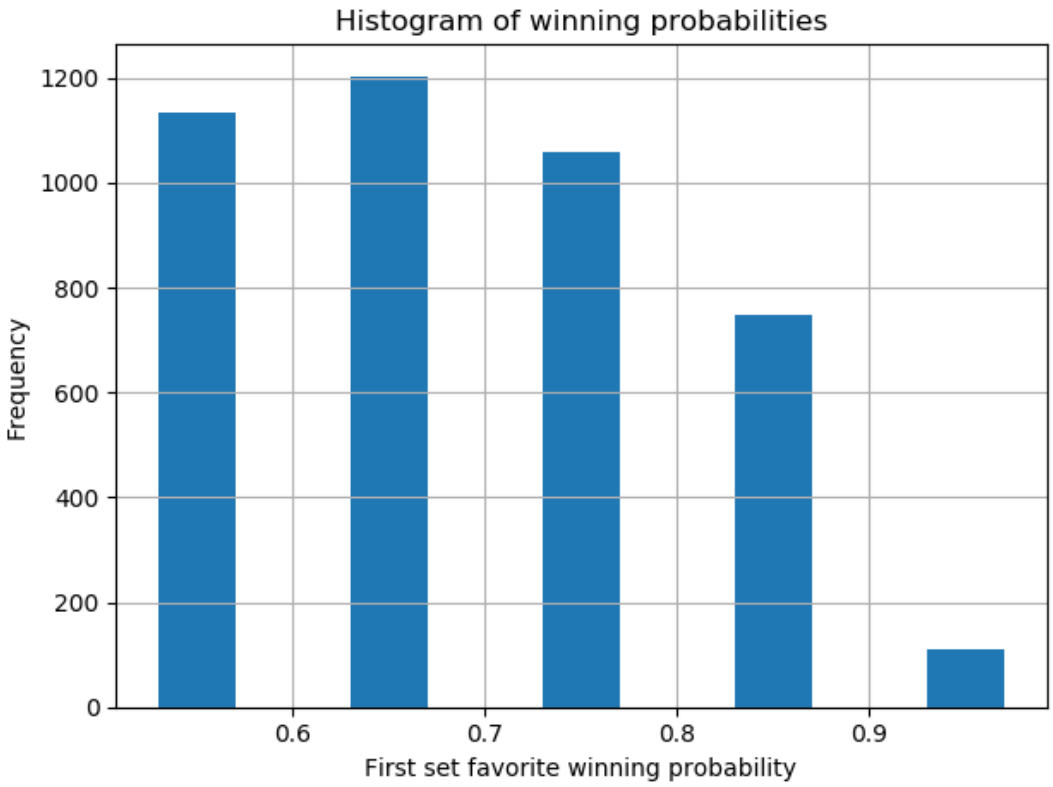

Figure 1: First set winning probability p_0 histogram.

1. Out of the 180[10] newly created odds-based subgroups, only 9 have data strong enough to reject H_0 on a 95% confidence level. The entire results of the hypothesis testing (the *p-values* of respective tests) can be seen on Figure 2.

Overall, the model was tested on 360 different subsets and only 22 of them (6.1%) provided enough evidence to reject H_0 on 95% confidence level. These subsets are distributed randomly and there is no pattern among them, indicating there is no systematic bias in the model. The random walk with varying probabilities thus seems to be a robust model which can be used to precisely predict set winning probabilities in men tennis Grand Slam matches.

6 Conclusion

This paper describes the random walk with varying probabilities and its application on Grand Slam tennis data. A model describing the development of a single match is introduced and tested on a dataset containing all matches from seasons 2009-2018. The results show that the model is robust and performs well on the absolute majority of reasonable data subsets. This suggest that the model could be used as a tool to generate precise *in-play* odds during the

[10]Further division, i.e. by tournament and odds, was not performed as the resulting datasets would not contain enough data.

Year	Set number	Australian Open	French Open	Wimbledon	US Open	0,6	0,7	0,8	0,9	1	All groups
2010	2	0,377	0,551	0,278	0,847	0,064	0,764	0,877	0,264	1,000	0,439
	3	0,981	0,629	0,302	0,703	0,738	0,091	0,927	0,644	1,000	0,561
	4	0,105	0,200	0,040	0,837	0,000	0,893	0,636	0,228	1,000	0,013
	5	0,808	0,270	0,156	0,509	0,567	0,060	0,076	0,930	1,000	0,084
2011	2	0,159	0,649	0,409	0,494	0,155	0,900	0,854	0,229	1,000	0,222
	3	0,893	0,696	0,622	0,553	0,195	0,108	0,875	0,697	1,000	0,689
	4	0,823	0,525	0,625	0,474	0,996	0,772	0,930	0,870	1,000	0,868
	5	0,496	0,329	0,014	0,427	0,144	0,130	0,622	0,206	1,000	0,127
2012	2	0,677	0,540	0,237	0,348	0,482	0,167	0,704	0,235	0,081	0,113
	3	0,752	0,304	0,264	0,621	0,245	0,852	0,440	0,313	0,928	0,138
	4	0,104	0,267	0,810	0,161	0,450	0,135	0,105	0,019	0,304	0,031
	5	0,223	0,184	0,412	0,359	0,279	0,452	0,358	0,019	0,137	0,192
2013	2	0,664	0,944	0,954	0,218	0,867	0,119	0,854	0,090	0,486	0,696
	3	0,306	0,629	0,320	0,476	0,359	0,476	0,001	0,197	0,175	0,373
	4	0,647	0,585	0,949	0,879	0,285	0,096	0,302	0,082	0,912	0,578
	5	0,488	0,501	0,510	0,385	0,357	0,907	0,377	0,161	1,000	0,579
2014	2	0,277	0,410	0,448	0,450	0,894	0,501	0,957	0,092	0,229	0,341
	3	0,244	0,612	0,511	0,987	0,908	0,181	0,404	0,894	0,885	0,253
	4	0,221	0,048	0,025	0,337	0,082	0,010	0,867	0,036	0,616	0,001
	5	0,191	0,142	0,495	0,792	0,117	0,852	0,240	0,170	1,000	0,636
2015	2	0,883	0,593	0,669	0,075	0,257	0,765	0,095	0,766	0,251	0,757
	3	0,084	0,223	0,565	0,272	0,227	0,798	0,447	0,237	0,294	0,081
	4	0,150	0,101	0,738	0,778	0,440	0,828	0,148	0,095	1,000	0,113
	5	0,025	0,316	0,428	0,454	0,063	0,694	0,520	0,907	1,000	0,089
2016	2	0,602	0,936	0,268	0,194	0,956	0,596	0,959	0,955	0,072	0,644
	3	0,563	0,021	0,697	0,341	0,411	0,929	0,867	0,169	0,986	0,442
	4	0,101	0,560	0,607	0,191	0,125	0,102	0,828	0,619	0,971	0,039
	5	0,240	0,311	0,352	0,035	0,197	0,067	0,074	0,562	0,593	0,008
2017	2	0,062	0,023	0,586	0,965	0,531	0,076	0,391	0,008	0,469	0,069
	3	0,677	0,901	0,154	0,146	0,852	0,053	0,636	0,654	0,504	0,110
	4	0,228	0,972	0,498	0,390	0,723	0,382	0,908	0,886	0,658	0,446
	5	0,526	0,381	0,542	0,465	0,783	0,613	0,729	0,466	1,000	0,915
2018	2	0,911	0,491	0,354	0,530	0,043	0,393	0,265	0,635	0,398	0,239
	3	0,765	0,233	0,404	0,165	0,481	0,730	0,473	0,965	0,444	0,320
	4	0,793	0,176	0,456	0,650	0,704	0,577	0,046	0,965	1,000	0,157
	5	0,258	0,216	0,171	0,060	0,841	0,052	0,399	0,806	1,000	0,190

Figure 2: *p-values* of hypothesis tests for different testing sets. Red are marked those allowing to reject H_0 on 99% confidence level, orange on 95% and yellow on 90% confidence level.

matches or to directly compete against the odds currently provided by the bookmakers.

7 Remarks

The source code containing all functionality mentioned in this article is freely available as open source at GitHub[11] together with a database containing all data that was used in this paper. Some results can be also obtained from the same repository.

References

[1] European Sports Security Association. Essa 2018 annual integrity report. http://www.eu-ssa.org/wp-content/uploads/ESSA-2018-Annual-Integrity-Report.pdf, 2018. Accessed: 2019-05-12.

[2] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 1995.

[3] Physical Activity Council. 2019 physical activity council's overview report on u.s. participation. www.physicalactivitycouncil.com/pdfs/current.pdf, 2019. Accessed: 2019-05-12.

[4] Tomáš Kouřim. Mathematical models of tennis matches applied on real life odds. *Doktorandské dny FJFI*, 2015. Available at http://kmwww.fjfi.cvut.cz/ddny/historie/15-sbornik.pdf.

[5] Tomáš Kouřim. Predicting tennis match outcomes using logistic regression. *Doktorandské dny FJFI*, 2016. Available at http://kmwww.fjfi.cvut.cz/ddny/historie/16-sbornik.pdf.

[6] Tomáš Kouřim. Random walks with varying transition probabilities. *Doktorandské dny FJFI*, 2017. Available at http://kmwww.fjfi.cvut.cz/ddny/historie/17-sbornik.pdf.

[7] Tomáš Kouřim. Statistical analysis, modeling and applications of random processes with memory. *PhD Thesis Study, ČVUT FJFI*, 2019.

[8] Ministry of Finance of the Czech Republic. Hazard games overview for 2017. https://www.mfcr.cz/cs/soukromy-sektor/hazardni-hry/archiv-zakon-c-202-1990-sb/vysledky-z-provozovani/2017/hodnoceni-vysledku-provozovani-loterii-2016-32211, 2018. Accessed: 2019-01-23.

[9] Loïc Turban. On a random walk with memory and its relation with markovian processes. *Journal of Physics A: Mathematical and Theoretical*, 43(28):285006, 2010.

[11] https://github.com/tomaskourim/mathsport2019