A Bayesian regression approach to handicapping tennis players based on a rating system

Timothy C. Y. Chan[1] and Raghav Singal[*,2]

[1] *Department of Mechanical and Industrial Engineering, University of Toronto*

[2] *Department of Industrial Engineering and Operations Research, Columbia University*

July 18, 2018

E-mail addresses: chan@mie.utoronto.ca (T. C. Y. Chan), rs3566@columbia.edu (R. Singal).

[*] Corresponding author.

Abstract: This paper builds on a recently developed Markov Decision Process-based (MDP) handicap system for tennis, which aims to make amateur matches more competitive. The system gives points to the weaker player based on skill difference, which is measured by the point-win probability. However, estimating point-win probabilities at the amateur level is challenging since point-level data is generally only available at the professional level. On the other hand, tennis rating systems are widely used and provide an estimate of the difference in ability between players, but a rigorous determination of handicap using rating systems is lacking. Therefore, our goal is to develop a mapping between the Universal Tennis Rating (UTR) system and the MDP-based handicaps, so that two amateur players can determine an appropriate handicap for their match based only on their UTRs. We first develop and validate an approach to extract server-independent point-win probabilities from match scores. Then, we show how to map server-independent point-win probabilities to server-specific point-win probabilities. Finally, we use the estimated probabilities to produce handicaps via the MDP model, which are regressed against UTR differences between pairs of players. We conclude with thoughts on how a handicap system could be implemented in practice.

Keywords: Tennis, Handicap, Rating systems, Bayesian models, Markov chain

Contents

1 Introduction

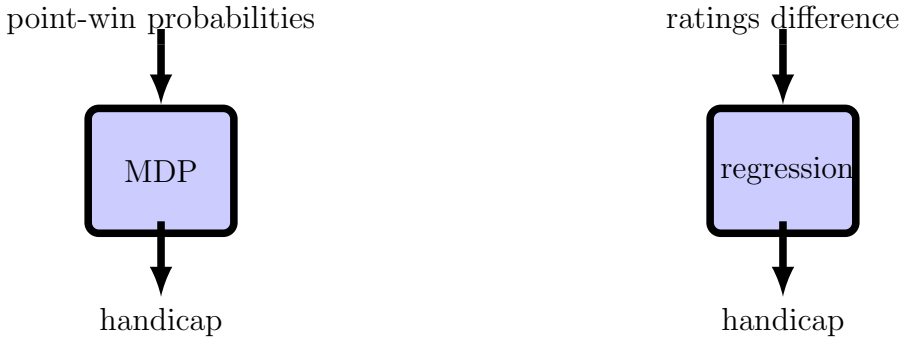
Handicap systems are used to improve fairness in competitive matches between players with different skill levels. Such systems are more commonly applied to amateur or social competitions, rather than professional competitions, in order to generate balanced and enjoyable matches. Handicapping also allows players to broaden their pool of potential opponents, which is especially important in sports where even slight skill differences can lead to unbalanced competitions. For example, in tennis, differences in skill are amplified over the course of a match, such that even small differences lead to the slightly stronger player enjoying a consistent and sustained advantage in terms of the match-win probability (Fischer, 1980).

Recently, a Markov Decision Process-based (MDP) handicap system was developed for tennis (Chan and Singal, 2016). This system gives the weaker player a budget of "credits" that can be used to claim points throughout the match in a dynamic fashion. The MDP model calculates the minimum number of credits required to generate a competitive match (that is, both players have a match-win probability ≈ 0.5) along with the optimal policy for when to dynamically use the credits throughout the match. To capture the skill difference between the players, the MDP model uses the probabilities of the weaker player winning a point on serve and return. Unfortunately, estimating point-win probabilities for amateur players is challenging because match data is usually recorded at a higher level of aggregation, unlike professional matches where point-level data is recorded. For instance, we were able to find only two sources[1] for match-level data on amateur tennis players and neither contain point-level information. Not surprisingly, the only datasets[2] we encountered that contain point-level data correspond to professional players. On the other hand, tennis rating systems are widely used by amateur players and provide an estimate of the difference in ability between players, but a rigorous determination of handicap using rating systems is lacking in

[1] https://universaltennis.com and https://tennislink.usta.com/. Note that the new domain name of https://universaltennis.com is https://myutr.com/.

[2] http://tennisabstract.com/ and https://github.com/JeffSackmann.

the literature.



(a) MDP approach **(b)** Regression-based approach

Figure 1: High-level picture of the MDP approach and the proposed regression-based approach. In the MDP approach, the input is point-win probabilites. In the regression approach, the input is ratings difference. The output of both the approaches is handicap.

The purpose of this paper is to make the MDP-based handicaps (Figure 1a) more accessible to a typical amateur tennis player. We do so by developing a data-driven, regression-based approach that leverages the MDP framework, but requires much simpler input to generate a handicap, namely player ratings (Figure 1b). In particular, we develop a mapping between the increasingly popular Universal Tennis Rating (UTR) system (UTR, 2015) and the MDP-based handicaps, so that two amateur players can determine an appropriate handicap for their match based only on their UTRs. As shown in Figure 1, our method allows amateur players to calculate appropriate handicaps without needing to estimate explicit point-win probabilities or knowledge of MDPs. Overall, such a regression-based mapping may be applicable in other sports to improve general uptake and use of handicapping methods that are based on complex mathematical models.

To develop a mapping between UTR differences and handicaps, we use a sequence of models and a granular dataset of amateur matches. Our data consists of match records of many amateur players (match scores and the UTRs of the corresponding players). For example, a match score could be 6-4, 6-4 where the losing player had a UTR of 8.3 and the winning player had a UTR of 8.8. First, we extract server-independent *game-win*

probabilities from the match scores. Note that server-specific data is generally not available at the amateur level. Second, we extract server-independent *point-win* probabilities from server-independent *game-win* probabilities using the Markov chain model for a single game of tennis (Kemeny and Snell, 1960). Third, we develop a mapping from *server-independent* point-win probabilities to *server-specific* point-win probabilities using a Bayesian model with a logistic link function. Fourth, we use the estimated server-specific point-win probabilities[3] to produce handicaps via the MDP model (Chan and Singal, 2016). These handicaps are then regressed against UTR differences between pairs of players using Bayesian linear regression. This final regression model allows us to summarize the mapping between UTR difference and handicap with an easy-to-remember formula.

Our contributions are as follows. First, using the UTR system and real match data from over 3,500 matches of amateur players played in 2015, we propose and validate a methodology for mapping a tennis ratings difference to a handicap computed using the MDP model of Chan and Singal (2016). Since point-win probabilities are difficult to obtain at the grassroots level, rigorously mapping a ratings difference to handicap has the potential to facilitate broader uptake of the handicap method. Second, we develop a novel technique to extract point-win probabilities from match score data using the classical Markov chain model for a single game of tennis. Given the lack of point-level data for amateurs, we believe estimation of point-win probabilities from match score data is important. Our approach is validated using data from over 4,000 professional matches played in 2015. Third, we show how to use and extend the Markov chain model for a single tennis game to compute the game, set, or match-win probability via the linear system of equations corresponding to the absorption probabilities of the Markov chain. We use this model to efficiently compute the match-win probability exactly given changing point-win probabilities. In addition, unlike other approaches in the literature, this approach outputs the match-win probability from *each*

[3]One can propose to use the *difference* in the server-specific point-win probabilities of the two players ("malus") instead of the probabilities themselves. However, as shown in Chan and Singal (2016), malus does not uniquely identify the MDP-based handicap and hence, such an approach would not yield exact handicaps.

match state. Fourth, we propose an original model to map server-independent point-win probability to server-specific point-win probabilities. We train our model on server-specific point-level data of over 2,500 professional matches played in 2015 and show that our model fits well to the data.

The paper is organized as follows. In Section 2, we propose and validate our models that transform match score data to server-specific point-win probabilities. In Section 3, we present the proposed mapping (UTR difference to handicap) and present the corresponding results. We provide a discussion of our results in Section 4 and some thoughts on potential implementation of a handicapping system in practice in Section 5. We conclude in Section 6.

2 Extracting point-win probabilities

In this section, we propose and validate our models that transform the match score data to server-specific point-win probabilities. Specifically, Section 2.1 discusses the mapping of match score data to server-independent point-win probabilities and Section 2.2 discusses the mapping of server-independent point-win probabilities to server-specific point-win probabilities.

2.1 Server-independent point-win probabilities

First, we describe our model to map match score data to server-independent point-win probabilities. Second, we validate it using data from 4,131 professional matches played in 2015.

Model To map the match score data to server-independent point-win probabilities (denoted as p), we first compute the server-independent game-win probabilities (denoted as q) using the match score data. For each match, we estimate q for a player as the ratio of the number of games won by that player to the total number of games played in the match.

Then, using q, we extract p by using the Markov chain model for a single game of tennis in reverse. We briefly review the Markov chain model next (Kemeny and Snell, 1960).

The state space $\mathbb{S}$ of the Markov chain consists of all possible game scores and two absorbing states (game-win state W and game-lose state L). In total, there are 17 states. Deuce is equivalent to 30-30. Advantage-in and advantage-out are equivalent to 40-30 and 30-40, respectively, assuming the player in question is serving; if the player is receiving, the assignment is reversed. For any state s, let s^+ and s^- denote the scores that result when the player wins or loses, respectively, the point at hand. We assume that the point outcomes are independent and identically distributed (iid) (Klaassen and Magnus, 2001) and denote by p the transition probability from state s to state s^+ and by $1-p$ the transition probability from state s to state s^-. By definition, the game-win probability q equals the absorption probability to state W (starting from state 0-0). Figure 2 displays the Markov chain.

Using the system of equations for absorption probabilities in a Markov chain, one can solve for q given p (Bertsekas and Tsitsiklis, 2008). Denote by a_i the game-win probability given that the current state is i. Then, as our Markov chain only has transient and absorbing states, the system of equations is:

$$\begin{aligned} a_W &= 1 \\ a_L &= 0 \\ a_s &= p a_{s^+} + (1-p) a_{s^-}, \ \forall s \in \mathbb{S} \setminus \{W, L\}. \end{aligned}$$

After solving the above system of linear equations for a_s, one can get q (which equals a_{0-0}) as a function of p:

$$q = p^4 \left(15 - 4p - \frac{10p^2}{1 - 2p(1-p)} \right). \tag{1}$$

Figure 3 shows q as a function of p. Since we are interested in mapping q to p, we simply invert the curve (numerically). Clearly, the mapping is one-to-one, which is desired.

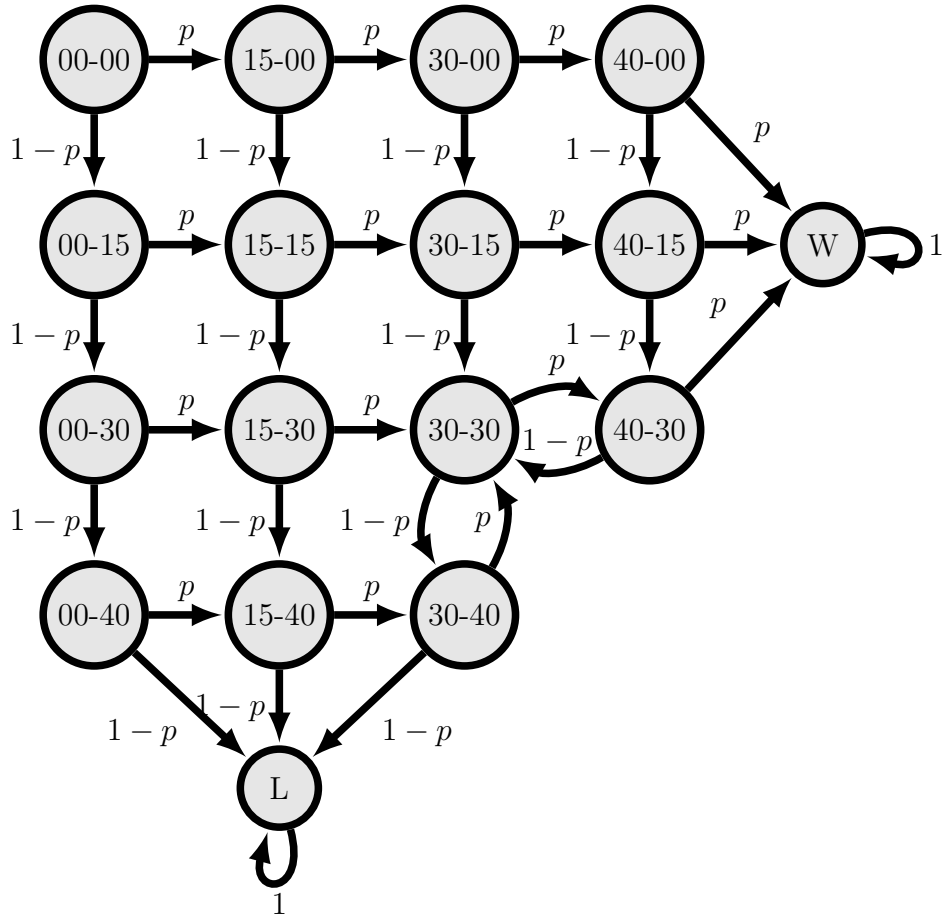


Figure 2: Markov chain for a single game of tennis with point-win probability p.

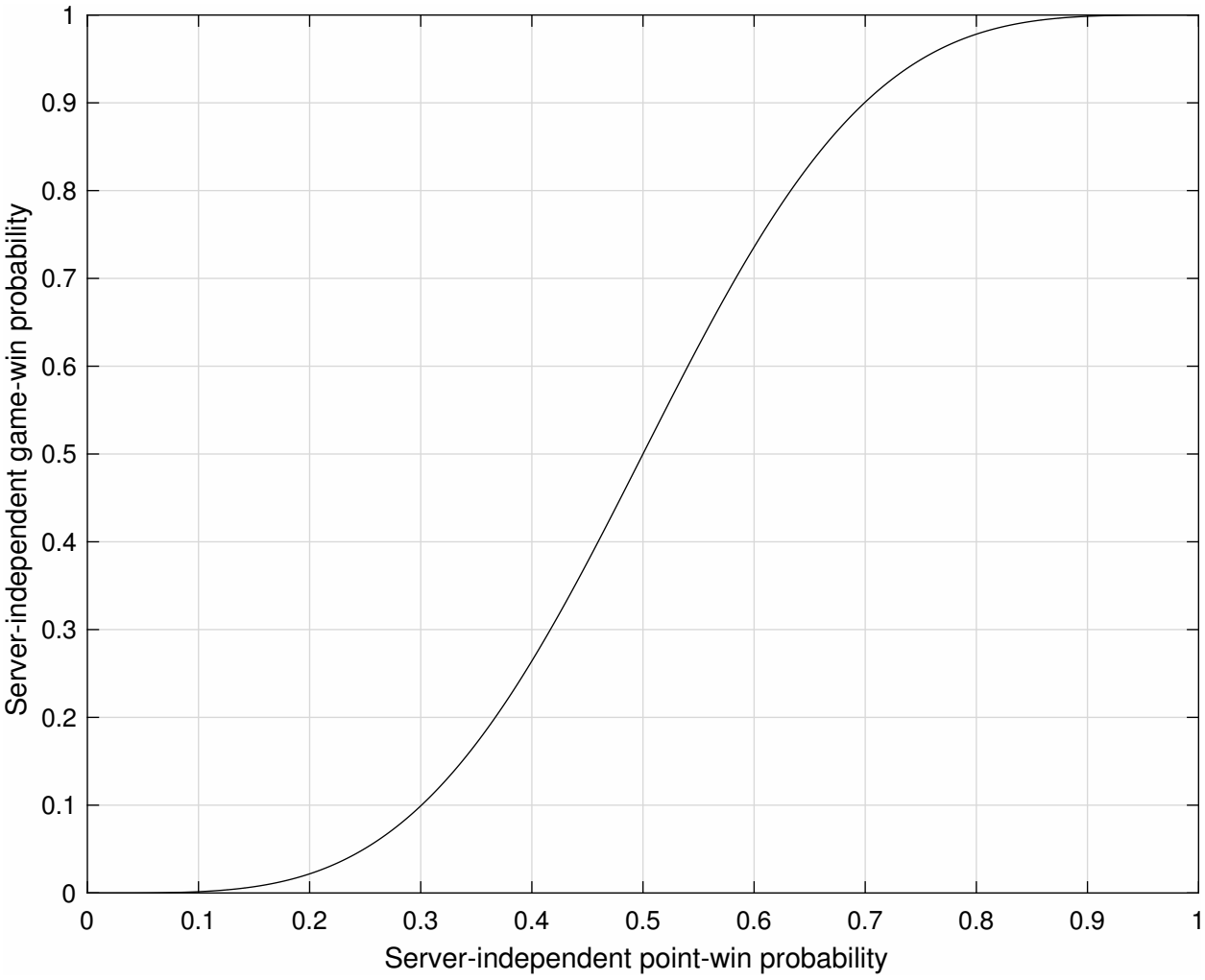


Figure 3: Relation between server-independent game-win probability and server-independent point-win probability using the Markov chain model for a single game of tennis.

Validation To validate our approach of mapping the match score data to p, we use point-level data from 4,131 professional matches played in 2015. Throughout this paper, data for professional men and women players corresponds to ATP and WTA matches data, respectively, collected from `http://www.tennisabstract.com` and `https://github.com/JeffSackmann/`. We use data from professional matches since point-level data is typically not available for amateurs. For each match, the data contains the number of points and the number of games won by both players. We estimate the server-independent point-win (game-win) probability as the ratio of the number of points (games) won by a player to the total number of points (games) played in that match.

Then, using the Markov chain approach described above, we map the estimated server-independent game-win probability to server-independent point-win probability. Figure 4 displays the scatter plot with hexagonal binning between the point-win probability computed using the Markov chain model and the point-win probability estimated from the point-level data. The Pearson correlation coefficient is 0.97, and the best-fit line through the origin has a slope of 1.00 (Eisenhauer, 2003), showing the effectiveness of our approach.

2.2 Server-specific point-win probabilities

Next, we describe our model to map server-independent point-win probabilities to server-specific point-win probabilities. Then, we validate it using the point-level data of 2,565 professional women's matches played in 2015.

Model Let p_s and p_r denote the point-win probability of the player when she is serving and receiving, respectively. To map p to (p_s, p_r), we use a Bayesian model with a logistic link function. First, we map p to p_s using the Bayesian model and then, we extract p_r by calibrating it such that the match-win probability implied by (p_s, p_r) is the same as the one implied by p. For now, we focus on the mapping from p to p_s. Later, we will comment on the calibration of p_r.

Our Bayesian model with a logistic link function is as follows. For each match i, we propose the data likelihood for p_s^i to be a truncated normal (truncated between 0 and 1), with mean equal to the logistic function $e^{\alpha+\beta p^i}/(1+e^{\alpha+\beta p^i})$ and standard deviation equal to σ, that is,

$$p_s^i \sim \mathcal{N}_{[0,1]}\left(\frac{e^{\alpha+\beta p^i}}{1+e^{\alpha+\beta p^i}}, \sigma\right), \ \forall i. \tag{2}$$

Note that p^i and p_s^i denote the server-independent point-win probability and the point-win probability when serving, respectively, of the weaker player in match i. (The player with the lower server-independent point-win probability is the weaker player.) We only consider the data for the weaker player since handicap is given to a weaker player, and hence, we

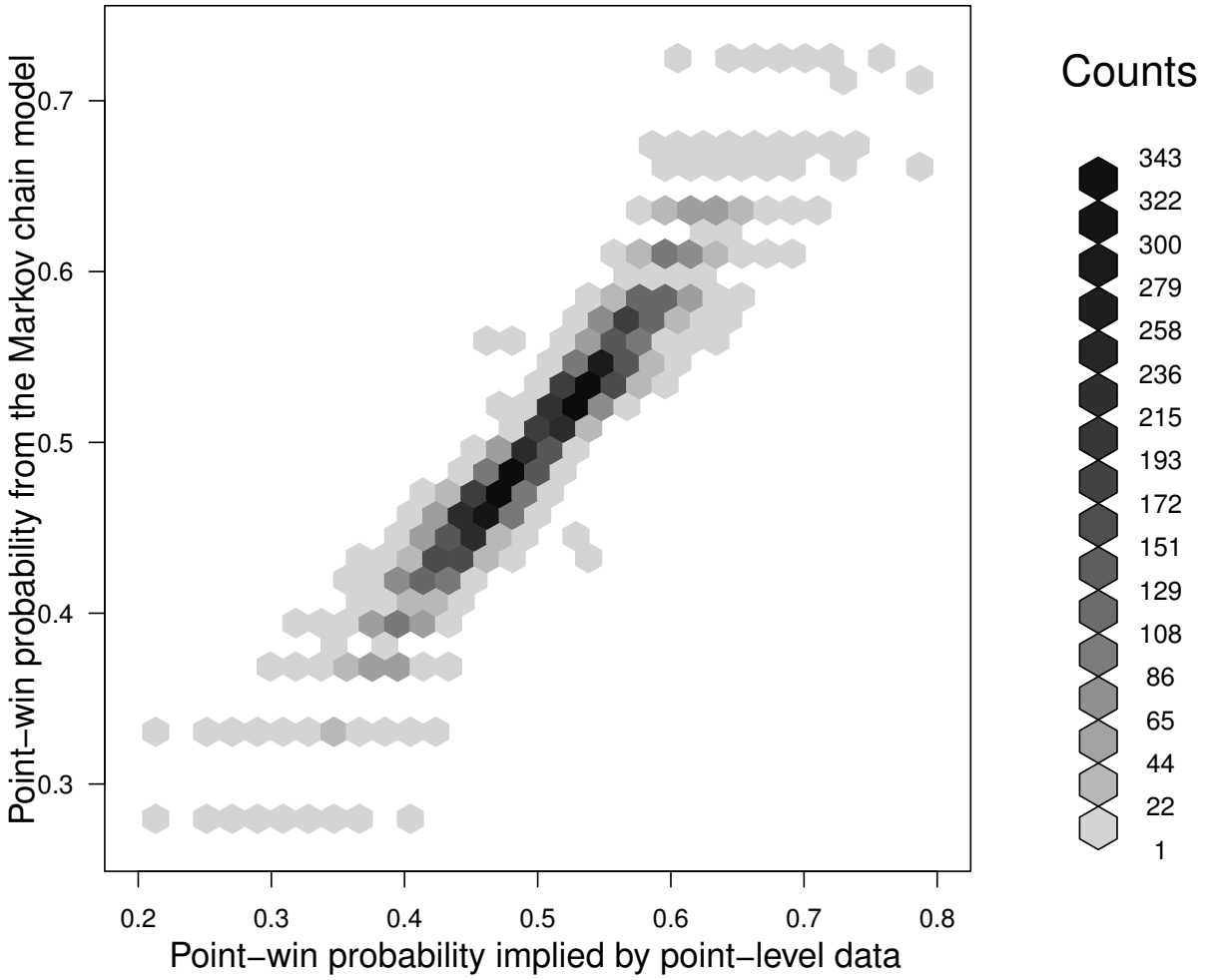


Figure 4: Benchmarking server-independent point-win probability computed from the Markov chain model against the server-independent point-win probability estimated from the point-level data. Pearson correlation coefficient is 0.97 and regression through the origin has a slope of 1.00.

only need the map for $p < 0.5$. The Bayesian parameters for which we want to obtain the posteriors of are α, β, and σ. To facilitate posterior inference, we assume that p_s^i is conditionally independent from p_s^j for all $i \neq j$ given the model parameters α, β, and σ, and data p^i. Since we have a reasonably large amount of data, we give non-informative

uniform priors[4] to all three parameters and let the inference be data-driven. The truncation is done to ensure that the output remains between 0 and 1 (since it is a probability). The intuition behind the logisitc link function comes from classical logistic regression, in which one models an outcome constrained between [0,1] using a logistic link function. Using the posterior samples of the Bayesian parameters, one can sample p_s^i. For each p^i, we sample N posterior samples of p_s^i.

Before we discuss the calibration of p_r, it will be useful to define two additional Markov chains corresponding to a tennis match with server-independent point-win probability and a tennis match with server-specific point-win probabilities, respectively. In the first Markov chain (tennis match with server-independent point-win probability), the state space consists of all possible scores (x, y, z), where x is the match score in sets, y is the set score in games, and z is the game score in points, and two absorbing states (match-win state W and match-lose state L). For any state s, let s^+ and s^- denote the scores that result when the player wins or loses, respectively, the point at hand. The server-independent point-win probability p is the transition probability from state s to state s^+ and $1-p$ is the transition probability from state s to state s^-. The second Markov chain (tennis match with server-specific point-win probabilities) is similar to the first Markov chain but the state is augmented to track who is serving and the transition probabilities are made server-specific. By construction, the absorption probability to state W equals the match-win probability in both these Markov chains.

Now, we discuss the calibration of p_r. For the n-th ($n = 1, \ldots, N$) posterior sample $p_s^{i,n}$ of match i, we obtain the corresponding $p_r^{i,n}$ as follows. First, using p^i, we compute the match-win probability for match i, denoted as r^i (via the server-independent Markov chain model for a best-of-three sets match). Then, we computationally solve for $p_r^{i,n}$ such that the match-win probability implied by $(p_s^{i,n}, p_r^{i,n})$ (in the server-specific Markov chain model for a best-of-three sets match) equals r^i.

[4]In the Bayesian modeling package (`Stan`) we used to fit this model, this translates to not specifying any prior in the code, as discussed at https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Table 1: Summary statistics of the posterior distributions of α, β, and σ.

	mean	sd	2.5%	50%	97.5%
α	-1.87	0.05	-1.96	-1.87	-1.78
β	4.27	0.10	4.07	4.27	4.48
σ	0.05	0.00	0.05	0.05	0.05

Validation To validate our approach of mapping p to (p_s, p_r), we start with server-specific point-level data of 2,565 professional women's matches played in 2015. For each match, the data contains the number of points won by each player conditioned on the server. We estimate the server-specific and server-independent point-win probabilities simply by taking the corresponding ratios ("true" values). Then, using our Bayesian model (for p_s) and the calibration approach (for p_r) described above, we map the estimated server-independent point-win probabilities to server-specific point-win probabilities. To fit the Bayesian model, we use `Stan` (Carpenter et al., 2016). We obtain posterior samples from 4 chains, each with 500 samples (we discard the first 250 samples for warm-up). The $\hat{R}$ values of all the parameters was less than 1.02, indicating convergence to the posterior (Gelman et al., 2014). The calibration of p_r is done computationally. First, we compute the best-of-three sets match-win probability for all $(p_s, p_r) \in \{0.001, 0.002, \ldots, 1\} \times \{0.001, 0.002, \ldots, 1\}$. Then, for a given match-win probability and p_s, we extract the corresponding p_r. Figure 11 in Appendix A displays the posterior distributions of the Bayesian parameters and Table 1 summarizes the posterior distributions. Figure 5 presents how well the model fits to the data. The 95% posterior intervals cover almost all the data, indicating a good fit.

3 Mapping UTR difference to handicap

In this section, we map UTR difference (denoted by d) between a given pair of amateur players to a handicap value (denoted by h) using a linear model of the form $h = \gamma d$. We do not include an intercept term in our linear model to enforce a handicap of zero for a ratings difference of zero. Our model uses the mapping from match score data to server-specific

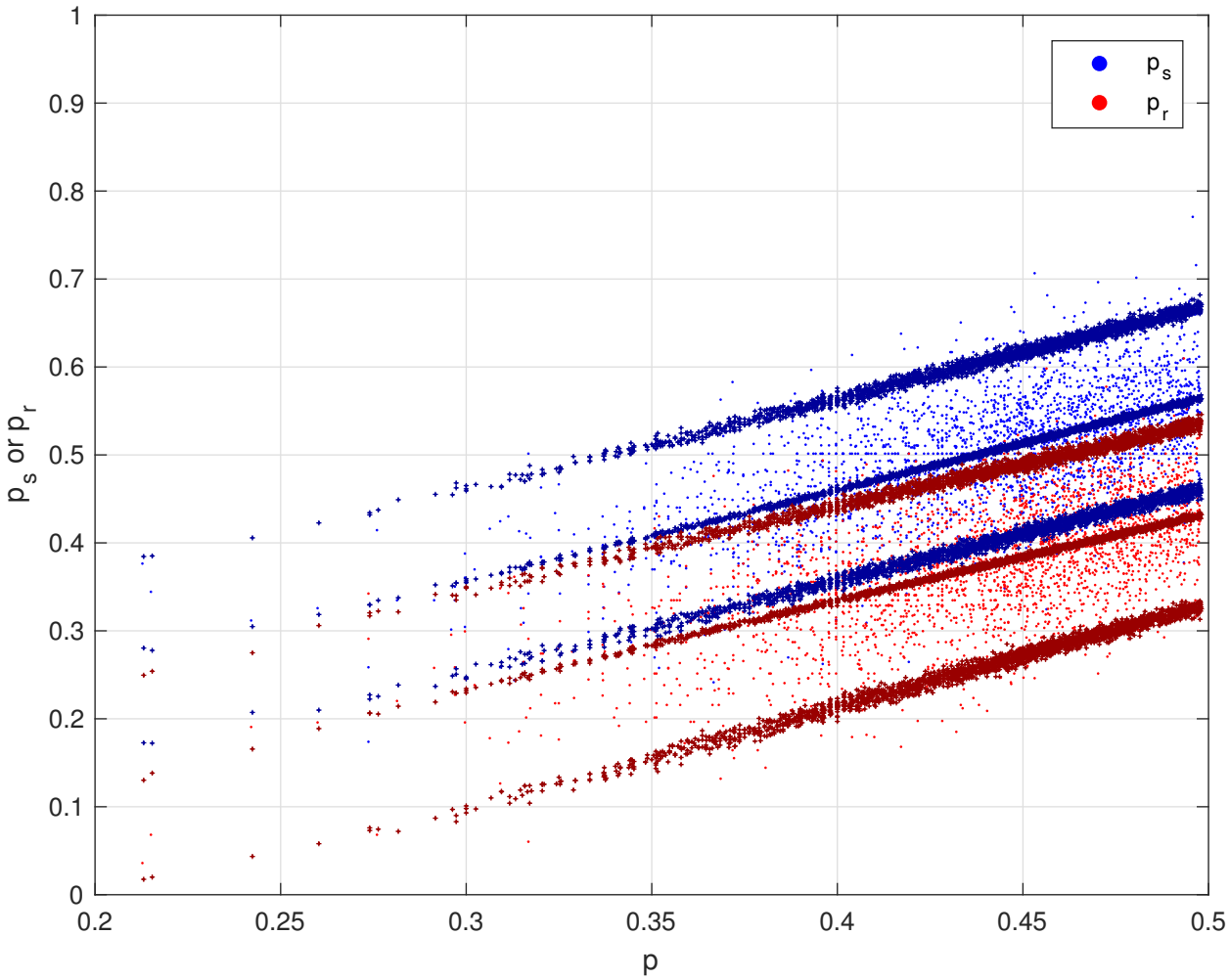


Figure 5: Predictions for p_s and p_r against the "true" values. The lighter color denotes the "true" data and the darker color denotes the prediction from the Bayesian model (p_s) and the calibration (p_r). The three darker lines (for both p_s and p_r) correspond to the mean, the 97.5th percentile and the 2.5th percentile.

point-win probabilities developed in Section 2. First, we describe our methodology (Section 3.1) and then, we present the corresponding results on a dataset consisting of 3,686 amateur matches played in 2015 (Section 3.2). Finally, observing our mapping is noisy, we discuss a denoising approach in Section 3.3.

3.1 Methodology

We start with the match records of the amateurs (match score and the UTR difference d^i for each match i). By using the models and the posterior samples of the Bayesian parameters

(α, β, and σ) from Section 2, we convert the match scores to server-specific point-win probabilities (p_s^i, p_r^i). To account for the uncertainty in the mapping from p to (p_s, p_r), we use $N = 100$ posterior samples of (p_s^i, p_r^i) for each i instead of using their point estimates (for each posterior sample n, we use a different posterior sample of α, β, and σ). Using the MDP model for a best-of-three sets match with a tie-break in all sets, we map each posterior sample $(p_s^{i,n}, p_r^{i,n})$ to the "true" linearly interpolated handicap value $h_*^{i,n}$. The linear interpolation is done to ensure match-win probability given the handicap equals 0.5. Then, we perform a Bayesian linear regression between $h_*^{i,n}$ and d^i as follows. We propose the data likelihood to be normally distributed, with mean equal to the linear function γd^i and standard deviation equal to τ, that is,

$$h^{i,n} \sim \mathcal{N}\left(\gamma d^i, \tau\right), \ \forall i \ \forall n. \tag{3}$$

The Bayesian parameters we want to obtain the posteriors of are γ and τ. To facilitate Bayesian inference, we assume conditional independence between all $h^{i,n}$ given the model parameters γ and τ, and the ratings difference data d^i. Since we have a reasonably large amount of data, we give non-informative uniform priors priors to both the parameters and let the inference be data-driven. The model is trained using the $h_*^{i,n}$ and d^i data. Figure 6 summarizes the overall process we use to map a match score to a corresponding handicap.

3.2 Results

Before presenting our numerical results, we briefly discuss the data. We use match data from 2015 for 250 amateur players (125 male, 125 female; 3,686 matches) with UTR ratings between 7.00 and 9.00 (roughly 4.0 to 5.0 on the NTRP rating scale) and a ratings reliability of 100%. The rating reliability is a proprietary measure of the UTR system that quantifies a level of confidence of the rating. Data was gathered from `http://www.universaltennis.com`[5]. In all 3,686 matches, both players are rated between UTR 7.00 and 9.00. Moreover, the ratings are uniformly spread over the range $[7.00, 9.00]$.

[5]Requires an account costing $4.95 per month to see player ratings to two decimal places

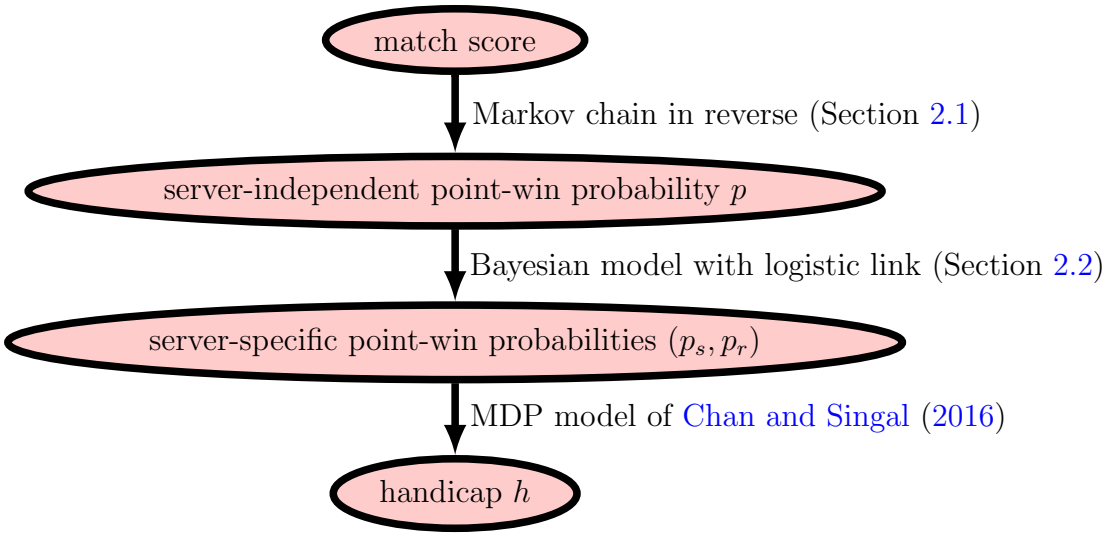


Figure 6: Mapping a match score to an appropriate handicap.

Finally, note that none of the matches were incomplete or a bagel (that is, 6-0, 6-0). We discuss the results in two parts. First, we discuss the results for combined data (data for both men and women) and then we discuss the results based on men's and women's data separately.

Results for combined data As discussed in Section 3.1, for each match, we generate $N = 100$ posterior samples of server-specific point-win probabilities, and for each posterior sample, we compute a handicap. Accordingly, for each of the 3,686 matches, we end up with 100 handicap values (expressing the uncertainty in the ratings difference to handicap mapping). Figure 7 displays the scatter plot (with hexagonal binning) showing the relationship between these 368,600 pairs of values. Expected handicap seems to increase linearly with the ratings difference, supporting our choice of a linear function for the mean in Equation (3).

We fit the Bayesian linear regression model (Equation (3)) using `Stan` (4 chains with 1000 samples per chain; first 500 samples were discarded for warm-up). All the $\hat{R}$ values were less than 1.02, indicating convergence. The results of the fitted model are summarized in Table 2 and Figure 12 (Appendix A). Table 2 summarizes the posterior distributions of the Bayesian parameters γ and τ, and Figure 12 shows a visualization of their posterior

Table 2: Summary statistics of the posterior distributions of γ and τ when fitted to the data for both men and women (3686 matces).

	mean	sd	2.5%	50%	97.5%
γ	12.51	0.03	12.46	12.51	12.57
τ	11.59	0.01	11.56	11.59	11.62

distributions. The expected value of γ equals 12.51, suggesting a handicap of 12.51 to be awarded to the weaker player for every one-point dfferential in UTR:

$$h = 12.51d. \tag{4}$$

We acknowledge that the expected value of τ is high, which is due to the fact that proportion of games won in a specific match is hard to predict and the ratings difference can only explain so much variation. Consequently, in Section 3.3, we present a method to reduce the noise. However, as seen in Figure 7, note that most of the data points lie close to the best-fit line.

Results for men's and women's data separately One of the purported benefits of the UTR system is that it uses the same scale to rate men and women (UTR, 2015), which is why we used both men's and women's matches together in the previous section. For comparison, we repeated the above process using the men's and women's data separately. The results are very similar to the results obtained for the combined data. The scatter plots with hexagonal binning are displayed in Figure 8, and the posterior distributions of γ and τ are summarized in Table 3. The expected value of γ for men (12.37) is close to that for women (12.61). This aligns with the claim the UTR scale is sex-independent. For conciseness, we do not show the posterior distributions of γ and τ but note they are unimodal and symmetric with low standard deviation (similar to the posterior distributions for the combined data shown in Figure 12).

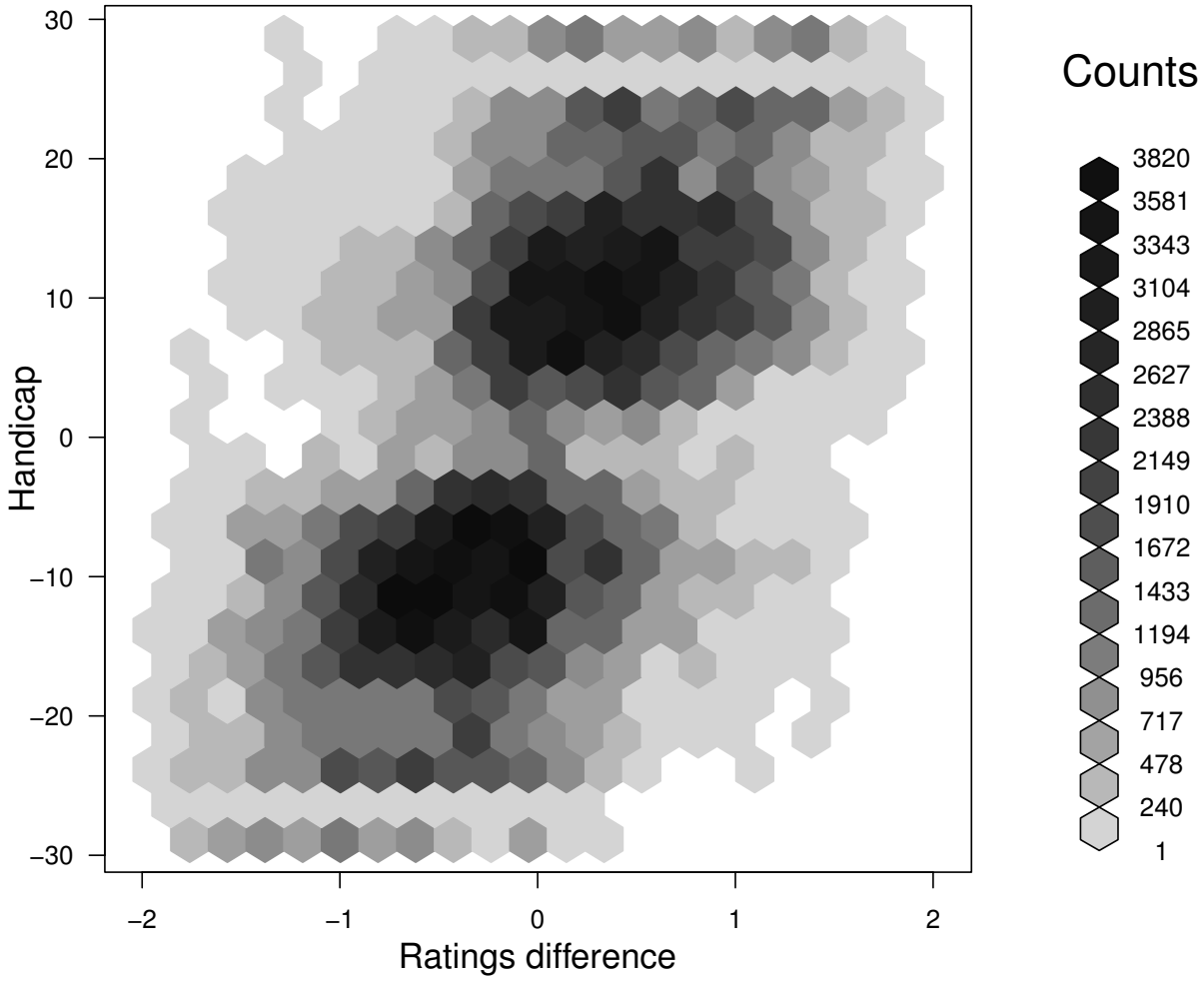
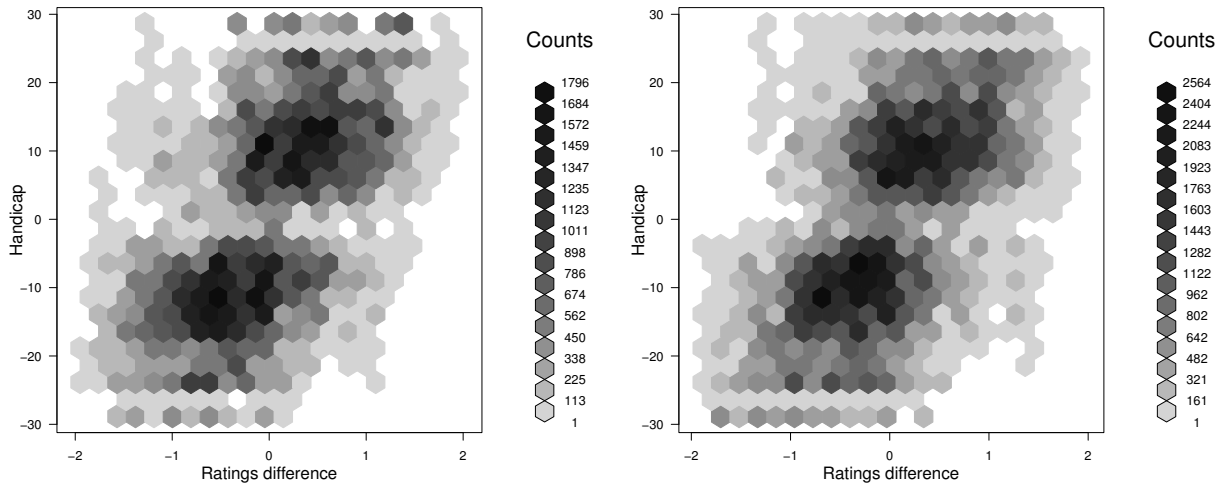


Figure 7: Scatter plot of handicap against ratings difference using hexagonal binning for the combined data.

3.3 Reducing noise in the data

As mentioned in Section 3.2, the proportion of games won in a specific match is hard to predict using UTR difference as the only predictor. In this section, we present an averaging method to reduce the noise in the mapping from UTR difference to handicap, inspired by the law of large numbers. Intuitively, predicting the proportion of games won in one



(a) Men **(b)** Women

Figure 8: Scatter plots of handicap against ratings difference using hexagonal binning for the data corresponding to men and women, respectively.

Table 3: Summary statistics of the posterior distributions of γ and τ for men (subscript m) and women (subscript w).

	mean	sd	2.5%	50%	97.5%
γ_m	12.37	0.04	12.29	12.37	12.44
τ_m	11.66	0.02	11.62	11.66	11.71
γ_w	12.61	0.03	12.54	12.61	12.67
τ_w	11.54	0.02	11.51	11.54	11.58

match is more challenging than predicting the same proportion over many matches played by that player. More formally, consider the following idealized setup. Suppose a player plays n matches and let q_i be the random variable denoting the proportion of games won by the player in match i for $i = 1, \ldots, n$. Suppose q_i is independent of q_j (given the ratings differences for matches i and j) for all $i \neq j$, and $\sigma_q^2 < \infty$ denotes the variance of q_i for all i. Then for $\bar{q} := \frac{1}{n}\sum_{i=1}^{n} q_i$,

$$\text{Var}(\bar{q}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} q_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_q^2 = \frac{\sigma_q^2}{n}.$$

Hence, the variance of $\bar{q}$ is less than the variance of each individual q_i by a factor of n, suggesting that the average proportion of games won should be easier to predict.

Accordingly, for each of the 250 amateur players in our dataset, we calculate an average server-independent game-win probability over all matches played by that player. The average game-win probability is calculated as total games won divided by total games played across all matches. We then map the average server-independent game-win probability to the average server-specific point-win probabilities using the models in Section 2. Similar to before, we use $N = 100$ posterior samples of average server-specific point-win probabilities, which are mapped to handicap values via the MDP model. For each player, we also compute an average opponent rating by taking the weighted average (weighted by total games played in each match) of the opponents' ratings. Thus, for each player, we have that player's rating, his/her average opponent rating, and 100 posterior samples of the corresponding handicap. We use this data to train the Bayesian linear regression model (Equation (3)).

We fit the Bayesian model using `Stan` (4 chains with 1000 samples per chain; first 500 samples were discarded for warm-up). All the $\hat{R}$ values were less than 1.02, indicating convergence. Figures 9 and 10 show the ratings difference versus handicap for the players considered and Tables 4 and 5 summarize the posterior distributions of γ and τ. Note that we only included players who played at least five matches. We also removed one women's player who was an obvious outlier, for example, with ratings difference of approximately 0.32 but an average handicap close to -11 (she lost all her matches, with some being quite lop-sided, and all were against lower ranked players). In total, we used 235 players.

The noise is visibly less in Figures 9 and 10 as compared to Figures 7 and 8. The mean value of τ for averaged data is around 3 (Tables 4 and 5), compared to a value of over 11 for non-averaged data (Tables 2 and 3), which aligns with our mathematical intuition presented earlier.

For the combined data (Table 4), the expected value of γ equals 10.39, which is around 2 points lower than the value obtained in non-averaged data (Table 2). Similar to before,

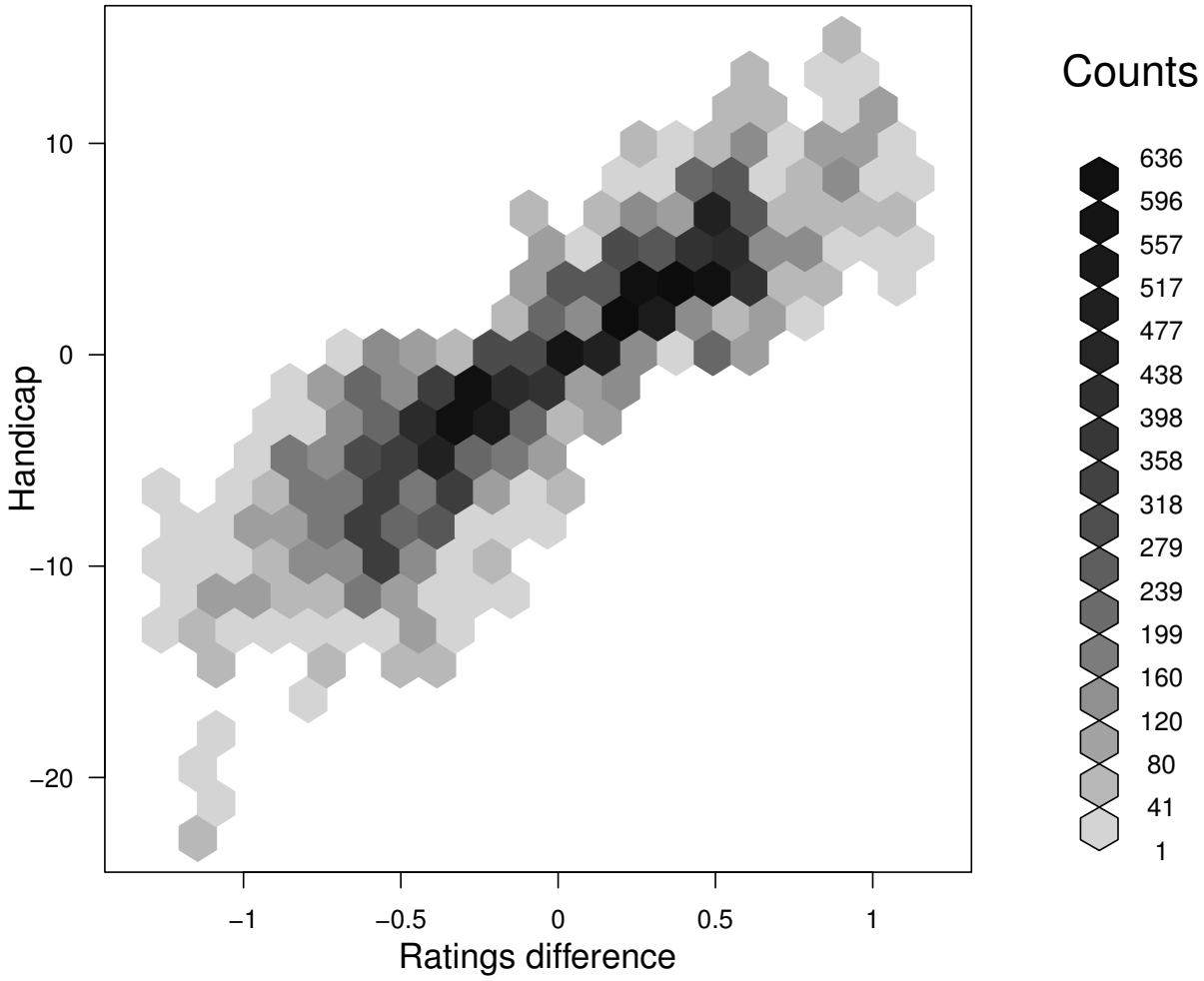
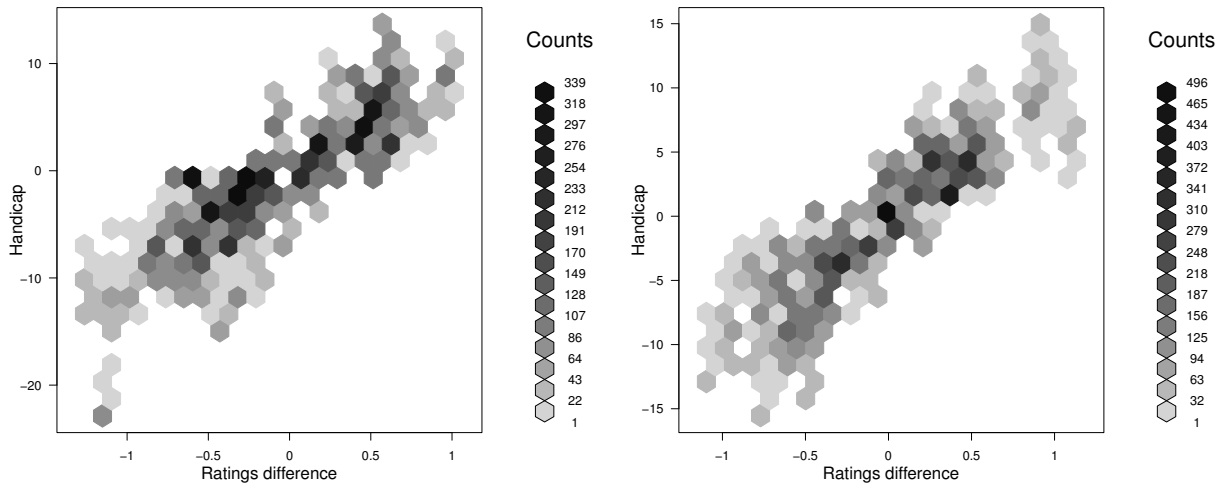


Figure 9: Scatter plot of handicap against ratings difference using hexagonal binning for the combined averaged data.

the expected value of γ is slightly lower for men (10.13) as compared to the value for women (10.67). The standard deviation parameter τ has most of its posterior mass below 3, implying a coefficient of variation of less than 30%. Results for the sex-specific data are similar.



(a) Men **(b)** Women

Figure 10: Scatter plots of handicap against ratings difference using hexagonal binning for the averaged data corresponding to men and women, respectively.

Table 4: Summary statistics of the posterior distributions of γ and τ when fitted to the averaged data for both men and women (3686 matces).

	mean	sd	2.5%	50%	97.5%
γ	10.39	0.04	10.32	10.39	10.47
τ	2.95	0.01	2.92	2.95	2.98

4 Discussion

Equation (1) has been derived in the literature previously using different methods (Carter Jr and Crews, 1974; Fischer, 1980; Liu, 2001). However, we believe our approach has several advantages. First, our approach generalizes elegantly to the case where one wishes to map p to the set-win probability or the match-win probability; all that is needed is to define the corresponding Markov chain. In Section 2.2, we map p to the match-win probability using this approach. Second, our approach is flexible enough to accommodate state-specific point-win probabilities. That is, p can be replaced by p_s, where p_s denotes the point-win probability in state s. With this approach, we can relax the iid assumption or incorporate different point-win probabilities on serve and return when modeling the outcome of a set or

Table 5: Summary statistics of the posterior distributions of γ and τ for men (subscript m) and women (subscript w) when fitted to the averaged data.

	mean	sd	2.5%	50%	97.5%
γ_m	10.13	0.06	10.02	10.13	10.25
τ_m	3.24	0.02	3.20	3.24	3.28
γ_w	10.67	0.05	10.58	10.67	10.76
τ_w	2.65	0.02	2.61	2.65	2.68

a match. Indeed, in Section 2.2, we make the point-win probability server-specific and use this approach to compute the match-win probability. Previous approaches to accommodate non-constant point-win probabilities have relied on Monte Carlo simulation (Newton and Aslam, 2009), whereas our approach is exact. Third, though we do not use this property in the current paper, it is worth mentioning that solving the absorption probability equations gives the exact match-win probability from *each* underlying state in the match. Note that there have been attempts to do so in the literature (Klaassen and Magnus, 2003; O'Malley, 2008). Klaassen and Magnus (2003) do not disclose their methodology, whereas O'Malley (2008) only compute match-win probabilities from a few states in the match. Our approach makes it simple to compute match-win probabilities from all states in the match by solving just one (sparse) system of linear equations.

Next, we discuss the reason we only use data from women's matches in validating the model presented in Section 2.2. In Section 3, we map the server-independent point-win probabilities to server-specific point-win probabilities for amateurs using the posterior samples of the Bayesian parameters (α, β, and σ) obtained from this validation. Since we do not have point-level data for amateurs, we can only use the data from professional matches for validation. In addition, we believe amateur players have lower point-win probabilities when serving than professional men's players do. As women professional players have historically won a lower fraction of points when serving compared to men (Klaassen and Magnus, 2001), we believe that data to be more representative. Moreover, the UTR of the top 100 women was lower (and hence, closer to the UTR of amateurs) than the UTR of the top 100 men.

Note that all our analysis on amateurs was done for players rated between UTR 7 and 9. We suspect a linear relationship between handicap and ratings difference will hold within a small range, but in general the handicap may depend on both the ratings difference as well as the absolute values of the ratings. Of course, the handicap between two players with a huge ratings difference is unlikely to be useful in practice because their skill difference will be too large to enjoy a competitive match anyway. We also note that in principle the handicap should be surface-specific, though at the amateur level we suspect most matches are played on hard courts. Thus, at the amateur level, we believe our handicap-ratings difference mapping is most applicable for hard courts. Finally, note that the player ratings we used correspond to the date when we downloaded the data. Ideally, we would have the ratings on the dates the matches were played, but this data was unavailable.

5 Implementation considerations

Before concluding, we briefly discuss ways in which a handicap system proposed by Chan and Singal (2016) could be implemented in practice by a tennis club or larger governing body. One way would be to track point-level data (or at least match scores) so that point-win probabilities between players could be computed. Depending on whether the point-win probabilities are server-specific, the appropriate MDP model could then be used to compute handicaps. To improve accessibility, the MDP model could be packaged as a black box, simply requiring parameters such as probabilities and length of match as input and then generating handicaps as output. However, we acknowledge that there are practical issues related to point-win probability estimation that need to be resolved before such an approach can be implemented. For example, questions such as what data history to use, what to do for players with few matches played, etc. need to be addressed. Such issues might make the MDP approach hard to implement in practice.

An alternative, possibly simpler method would be for the organization to adopt a rating system and use our regression approach presented in Section 3 to directly map ratings

difference to handicap. This approach avoids the need to track point-level data but does require adoption of a granular rating system, which could result in additional costs to the organization or players. A potential benefit of such an approach is that decisions regarding the match history used to determine the rating, for example, which influences the history of data used to calculate point-win probabilities, become endogenized within the rating system. Note that if a ratings system is used, then a question arises as to how to update ratings (if at all) in matches where handicapping is used. In principle, if the handicapping is perfect, then the ratings of players in a handicapped match should be updated based on the outcome as if the match was between two equally competitive players.

A valid critique of implementing the proposed approach in practice is that it would yield a handicap with a high amount of uncertainty for individual matches (as seen in Section 3.2). The underlying reason for this undesirably high level of uncertainty is that the ratings difference can only explain a limited variation in the *realized* performance of the tennis players on a given day. Consequently, one should view the parameter γ as the *expected* increase in handicap per unit difference in ratings. On a given day, due to the variance in the performance of the players (and the inability of the ratings difference to explain such variance), the *realized* handicap might deviate from the expected handicap. However, on average, the expected handicap is correct. Such a critique raises the following research question: what factors does one need to account for to explain the variance in tennis match outcomes to a good extent? If one can find such a set of factors, then one can estimate the handicap as a function of these factors (using the methodology developed in this paper). However, the more factors one adds to explain the variance, the more challenging it becomes to keep the handicap equation "easy-to-remember".

Adding more to the above point, we expect our one-factor model (the one factor being the ratings difference) to be a rough guidance for individual players, who can adjust the exact amounts given over time by actually playing with handicap and observing what happens. Perhaps a player might need to give less / more to her opponent than our model suggests

because of other factors we have not accounted for, e.g., handedness (right or left), surface of play (clay or grass or hard), etc.

6 Conclusion

In this paper, we developed a Bayesian regression-based approach to map tennis ratings differences to a handicap determined from a Markov Decision Process. We used data from the Universal Tennis Rating system as the basis for the analysis. For a three-set match, we found that a one point difference in UTR between two players translated to 10-12 handicap points. Using a rating system instead of the point-win probabilities required by the MDP model for computing handicaps allows amateur players to calculate handicaps easily. Our overall framework, while rooted in tennis, may be applicable in other sports where handicap calculations are based on complex models but handicaps themselves may be directly estimated from a sport-specific ratings system.

References

Bertsekas, D. and J. Tsitsiklis (2008). Introduction to Probability (Athena Scientific). *Nashua, NH*.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2016). Stan: A probabilistic programming language. *Journal of Statistical Software 20*, 1–37.

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

Chan, T. C. and R. Singal (2016). A Markov Decision Process-based handicap system for tennis. *Journal of Quantitative Analysis in Sports 12*(4), 179–188.

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching statistics 25*(3), 76–80.

Fischer, G. (1980). Exercise in probability and statistics, or the probability of winning at tennis. *American Journal of Physics 48*(1), 14–19.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. CRC press Boca Raton, FL.

Kemeny, J. G. and J. L. Snell (1960). *Finite Markov Chains*, Volume 356. van Nostrand Princeton, NJ.

Klaassen, F. J. and J. R. Magnus (2001). 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.

Klaassen, F. J. and J. R. Magnus (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research 148*(2), 257–267.

Liu, Y. (2001). Random walks in tennis. *Missouri Journal of Mathematical Sciences 13*(3).

Newton, P. K. and K. Aslam (2009). Monte carlo tennis: a stochastic markov chain model. *Journal of Quantitative Analysis in Sports 5*(3).

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

UTR (2015). Universal Tennis, Universal Tennis Rating System. http://universaltennis.com. [Accessed: 23-November-2015].

A Additional figures/tables

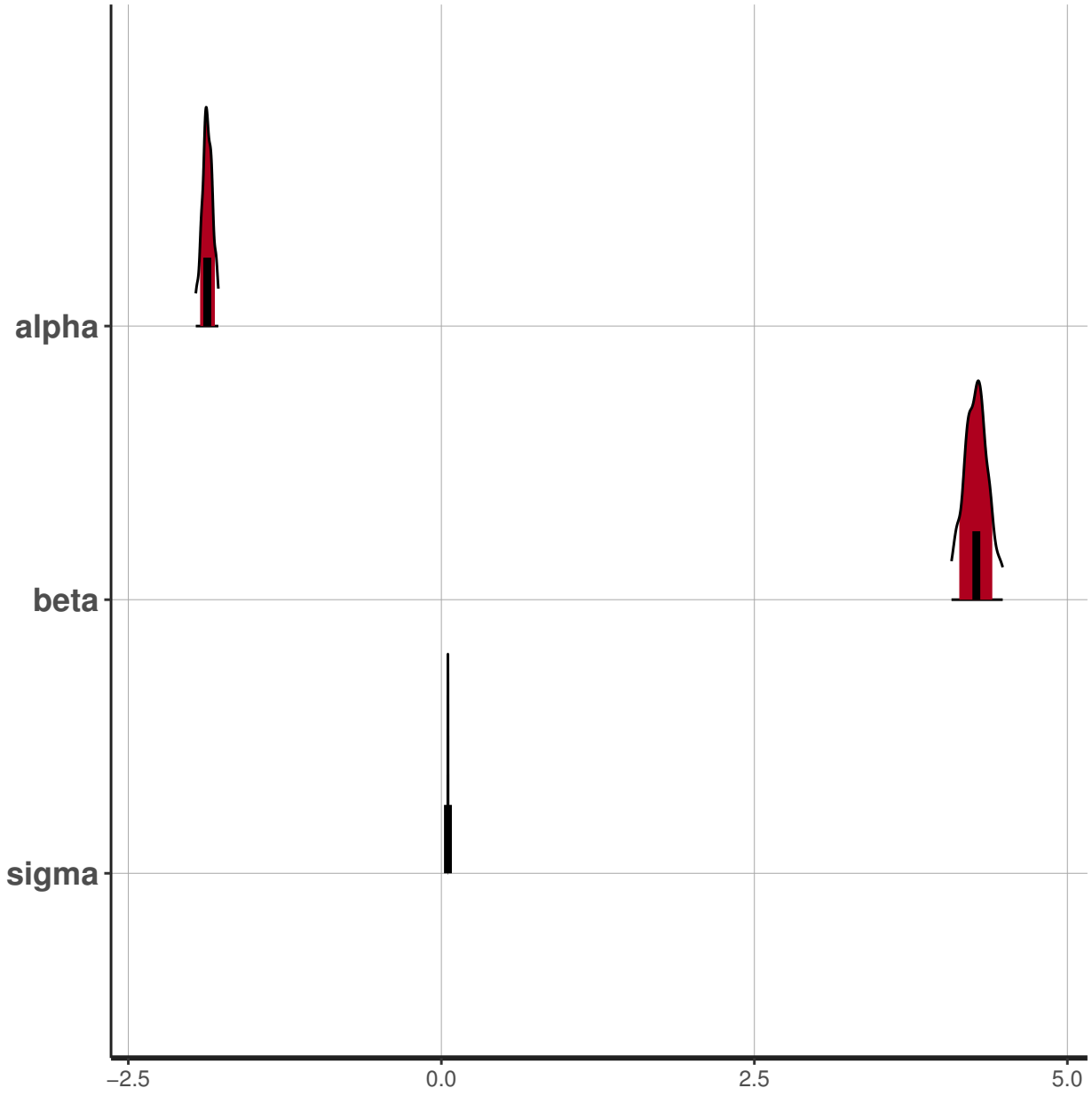


Figure 11: Posterior distributions of the Bayesian parameters α, β, and σ corresponding to the Bayesian model (Equation (2)). Red area denotes the 80% posterior interval and black curve denotes the 95% posterior interval. The distributions are unimodal and symmetric with low standard deviations.

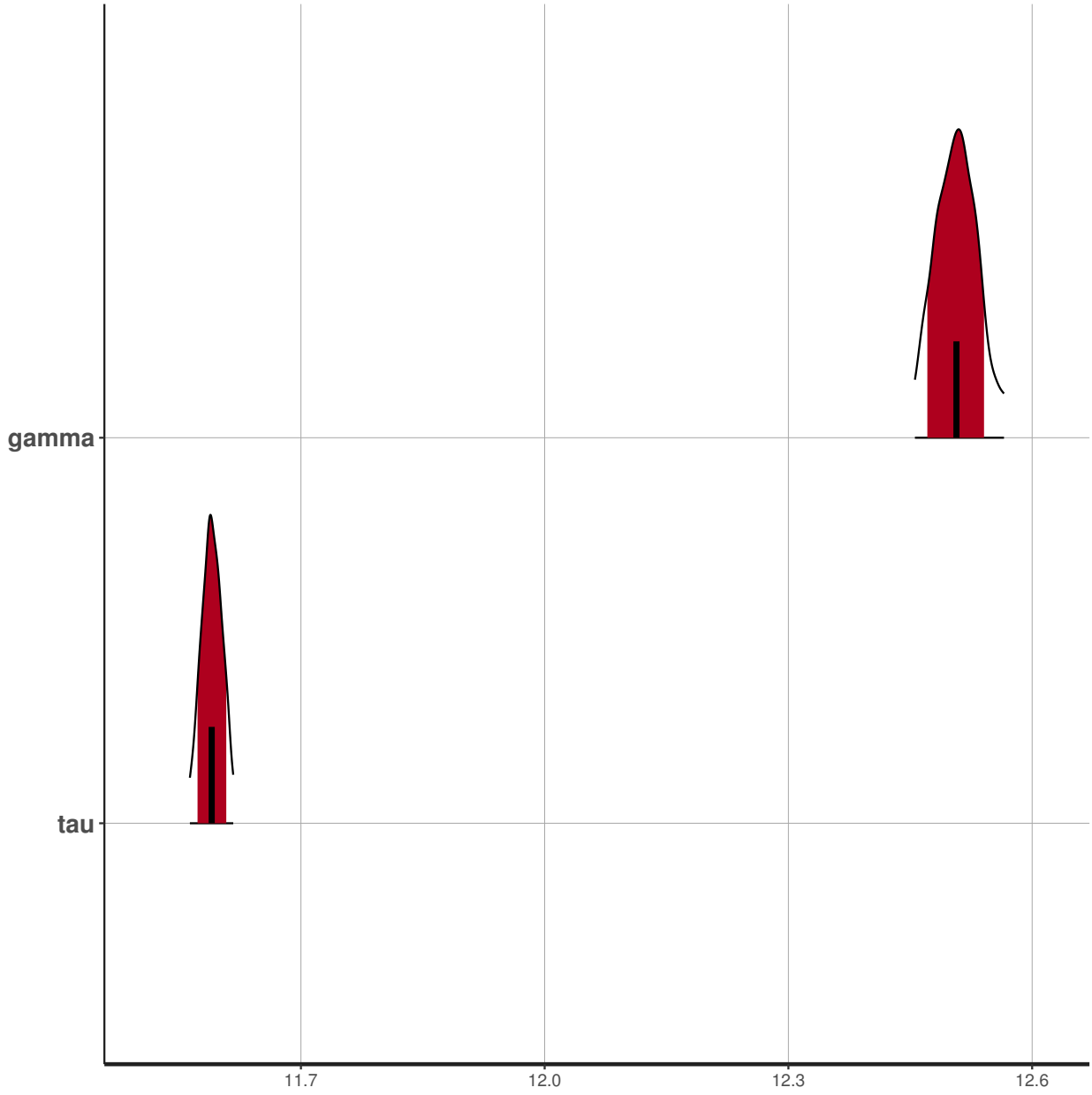


Figure 12: Posterior distributions of the Bayesian parameters γ and τ corresponding to the Bayesian linear regression model (Equation (3)) when fitted to the combined data (3686 matces). Red area denotes the 80% posterior interval and black curve denotes the 95% posterior interval. The distributions are unimodal and symmetric with low standard deviations.