SWI - Tennis rating for KNLTB

Anatoliy Babic[1], Jeremy Budd[2], Fokko van de Bult[1], Stijn Cambie[3], Thomas Gomez[4], David Kok[5], Roel Lambers[6], Yukihiro Murakami[1], Len Spek[7], Esmee te Winkel[8], Bas van der Wulp[9]

Abstract

The KNLTB is looking for a new or improved rating system for doubles matches in tennis. We have analysed how the current rating system functions as a predictor of match outcome and how to assign a team rating based on the current individual ratings of the players. We propose two new rating systems based on the Elo and Glicko ratings used in other sports. In both systems each player gets a new rating which is the old rating plus or minus a small amount based on the match results. On top of that the Glicko system also tracks how much uncertainty there is in each players rating.

[1]STOLBV, The Netherlands
[2]Delft University of Technology, The Netherlands
[3]Radboud University, The Netherlands
[4]Utrecht University, The Netherlands
[5]Leiden University, The Netherlands
[6]Eindhoven University of Technology, The Netherlands
[7]University of Twente, The Netherlands
[8]University of Warwick, England
[9]Fontys University of Applied Sciences, The Netherlands

KEYWORDS: Rating system, implementation, mathematical model, data analysis

1.1 Introduction

The KNLTB, Koninklijke Nederlandse Lawn Tennis Bond, is the national governing body for tennis in the Netherlands. They promote tennis as a sport and represent over half a million people who play tennis. They organise tournaments from the amateur to the professional level tennis and padel, an other racket sport. To unsure a fair competition, they rate each player in their skills in singles tennis, doubles tennis and padel.

The KNLTB has approached SWI, Studiegroup Wiskunde en Industrie, with a problem they have with their rating system. The problem mainly focused on how to design a good rating system for doubles tennis. During the week at SWI we first brainstormed on possible alternative rating sytems. In the end we chose to look at two systems: the Elo and the Glicko rating system. We then split into three groups with different objectives.

The first group did a statistical analysis of the current rating system. We have looked how it functions as a predictor of the outcome of matches in singles tennis. We also investigated whether, based on the current ratings, the stronger or the weaker player has a larger impact on the team rating. The second group did some theoretical work to extend the Elo and Glicko rating systems to handle doubles matches. The last group made a prototype implementation of these systems for a complete season of tennis. The results of this implementation are not discussed in this report.

This report is structured as follows: First we clearly define the problem as we have received it from the KNLTB and look at some other rating systems used in sports. Next we show our data analysis of the current rating system. Then introduce our proposed rating systems with a theoretical analysis. Finally we summarise our results and give recommendations

1.1.1 Statement of the problem

The main problem proposed by the KNLTB at SWI 2020 was to give a rating model that gives a good indication for the strength of pairs in double matches of tennis. Nevertheless, also suggestions for adaptations of the single rating are welcome, as at the end of the season multiple data corrections in match results are needed with the current system. An other question is how to deal with new players and re-entrants without rating or prior matches.

1.1.2 Criteria for a good rating system

In this subsection, we summarize the properties that the KNLTB wants the rating system to have. The old system of the KNLTB has some of these, but not all. The important and essential properties of the rating system are listed below.

- There should be always a rating change, in the sense that every match is taken into account. Also the rating change should not depend on the moment in the season the match has been played. In particular, matches with unbalanced partnerships should not be disregarded, they should still have an effect. The current system of the KNLTB has not this property.
- It should be impossible to win and get a worse rating. This is a logical assumption and the current KNLTB system has this property.
- The system should be explainable to all players. Players have to be able to know how winning or loosing a game will (approximately) influence their rating. This may be possible by implementing a FAQ or a calculator at the website, such that people can compute the possible outcomes for their rating in advance.
- The match result only affects the ratings of the players who were playing during the match.
- Rating should be a good predictor of performance. The reason for this is that tournaments are usually organised using ratings to predict performance to have a fair tournament.

Some other properties which can be taken into account, but are less important, are listed next.

- Incentivise playing by small positive effect on rating. Here we note that this should be a small effect as otherwise it can be manipulated by people to find strategies to make their rating better than it should be.
- People should not refuse to play games because they fear it hurts their rating. Exactly as with the previous property, the KNLTB wants people to enjoy and play a lot of tennis games.
- The double system should be a generalisation of the single system. If two partnerships compete and both players of each partnership have equal strength, the rating result should agree with the rating result for singles of these strengths.
- Frequent partners should converge to the same rating. An unbalanced partnership should converge to the same rating for both players if they play always together.
- We should disregard the amount of sets or games won. The outcome should only depend on winning or loosing the match.

1.2 Rating systems

In this section we explain and review some rating systems used in tennis and other sports, most notably chess.

1.2.1 Dynamisch Speelsterkte Systeem (DSS) of the KNTLB

In the Dynamisch Speelsterkte Systeem (DSS), KNLTB (2017), the rating system used by the KNTLB, each player has a singles and doubles rating between 1 and 9, where 1 is the best rating and 9 is the worst. At the start of the year each player is given a starting rating based on the previous year. After a match a player is given a rating result, based on their own or the opponents rating and whether they lost or

win. However when the players skill varies too much, no rating result is noted. The interpretation of this rating result is they played the match like a player with that rating. During the year their current rating is the average of the rating results achieved during that year. If there are less than 6 rating results, the starting rating is added the average up to 6 results.

For a match of singles tennis, the rating result is determined as follow: The result of the winner is the opponents current rating minus 1 and the result of the loser is the opponents current rating plus 1. Suppose player 1 has a rating of 4.3 and player 2 has a rating of 4.8. If player 1 wins, he gets a result of 3.8 and player 2 a result of 5.3. However if the difference in the rating is larger than 1.5 and the stronger player wins, no player gets a rating result. This means that players cannot improve their rating whenever they play against much lower rated players, but can worsen their rating (drastically). For this reason, it is very unattractive for competitive players that focus on getting a low rating to play against much weaker players.

For a doubles match this becomes more complicated as we have 4 players each with their own rating. The result is based on the outcome of the match, the own rating, the average rating of the opponents, the sum of the winners ratings minus the sum of the losers ratings and the largest within-team difference between ratings. The KNLTB uses a flowchart to guide the players through the possibilities, which we have summarised in Figure 1.1.

1.2.2 UTR

The Universal Tennis Rating is a global rating system. Every tennis player in the world can get a rating between 1.0 and 16.5, with 16.5 being the best rating. Each match which is entered in the system receives a match rating based on the difference of the amount of sets won and the rating of the opponent. The final player rating is then a weighted rolling average of the last 30 matches in the previous 12 months. The match weight is higher for longer of the games, more recent games, games played against a similarly rated player and games played against a player with a reliable rating. One of the consequences is the rating of the player which had just won, can get deteriorate when

Sum of winner ratings minus sum of loser ratings: D			Largest within-team difference: W			
			W < 1	1 ≤ W < 1.5	1.5 ≤ W ≤ 2.5	W > 2.5
D > 1.5		Winners	Average of opponents -1	Average of opponents -1	Own rating - 1	No result
		Losers	Average of opponents +1	Average of opponents +1	Own rating +1	
-1.5 ≤ D ≤ 1.5		Winners	Average of opponents -1	Own rating - ½	Own rating - ½	No result
		Losers	Average of opponents +1	Own rating + ½	Own rating + ½	
D < -1.5		Both pairs	No result	No result	No result	No result

Figure 1.1: The different rating results for a doubles match according to the DSS of the KNLTB

they won less sets than expected or when the match result is worse than the one which it supplants in the rolling average.

1.2.3 Elo rating system

The Elo-rating is a widely used rating system, initially developed by Arpad Elo (see Elo (1978)) to determine the relative strengths of chess players. The skill of players is relative to their rating and some uncertainty. In contrary with the rating systems above, the Elo-rating is also a predictive rating. Given the rating difference between two players, it is possible to calculate the probability that a player wins.

After every match the rating is updated by taking the old rating and adding or subtracting some points based on who won and how likely they were to win. So if you win against a stronger player, your rating improves much more than winning against a weaker player.

1.2.4 Glicko rating system

The Glicko-rating system Glickman (1999) was developed by Mark Glickman as an extension to the Elo-rating. It takes the Elo rating

as a basis, but instead of using a fixed amount of uncertainty for each player, it tracks how certain we are that the rating is a good reflection of the skill of the player. So if a player plays a lot of matches and plays very consistently, the uncertainty should be low. But for a new players or a very inactive players, we shouldn't be certain at all in their ratings.

This also factors into the win probability we can calculate, as the chance of an upset win or loss should be higher if there are players with uncertain ratings. This a similar effect on how the ratings are updated. So for example if we are uncertain about a players rating, they can gain or lose rating faster. However the opponent gains or loses rating slower, as winning or losing against an opponent with an uncertain rating does provide less information about their own skill.

1.2.5 Microsoft True Skil

The TrueSkill rating system Herbrich, Minka, and Graepel (2007) has been developed by Microsoft to be used in a broad spectrum of games offered on their Xbox game console. The main goal of this proprietary rating system is to match players with equivalent skills, in essence maximizing draw probability of matches. This way competitive matches between users on the platform ideally are between players of equal strength. TrueSkill is also suitable for multi-player games, but it assumes that player qualities are additive.

1.3 Analysis of the current system

Before we look at our proposed rating systems, we first investigate how the current system, DSS, of the KNLTB performs. The KNLTB has provided us with a data-set of all tennis matches of the last 3 years. Using this data-set we look if the difference in rating can predict the outcome of a match. In a good rating system, the probability of winning should be a logistic function of the difference in rating.

For the doubles rating, we check if the stronger or the weaker player has a larger impact on the team rating. We model the team rating by the weighted average of the individual ratings of the players. Using statistical analysis, we try to determine the optimal weight of each

player's rating. We repeat this analysis for a mixed doubles where we team rating is a weighed average of the male and female player.

1.3.1 Singles rating

Figure 1.2 shows frequency histograms for rating and age across singles matches for different groups of players. The peaks at integer ratings correspond to starting/returning players, because not enough information is available to assign a precise rating, so the KNLTB assigns them an integer rating. Compared to adult male players, adult female and young players are more concentrated around the lower ratings. The age distributions are similar for male and female players. The histograms clearly show that more matches are played by younger players (< 20 y.o.).

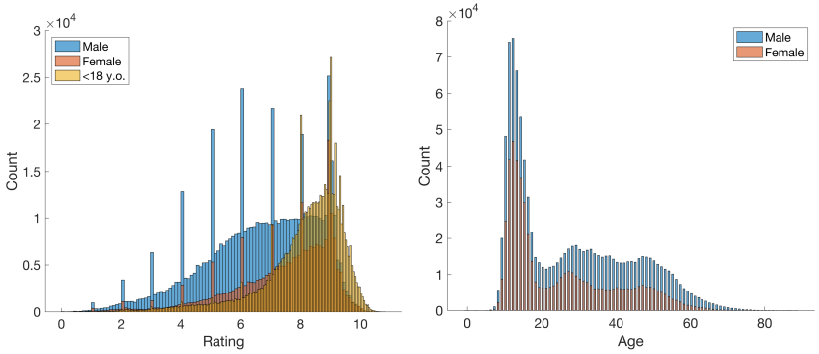

Figure 1.2: *Left:* Rating histogram across all singles matches for adult male, adult female, and underage players. *Right:* Age histogram across all singles matches of male and female players.

We set out to estimate the performance of the current rating system, as well as determine the effects of age and playing a home match on the odds of winning. We assume that when player A plays a match

against player B, the log-odds of player A wining are given by

$$\text{logit}(p_A) = \log \frac{p_A}{1-p_A} = \beta_0 + \beta_1 \text{Home}_A + \beta_2 (R_A - R_B) + \beta_3 (\text{Age}_A - \text{Age}_B), \tag{1.1}$$

where p_A denotes the probability of player A winning the match, R_X indicates the rating of player X, Age_X indicates the age of player X, and Home_A is a dummy variable that equals 1 if player A plays a home match and 0 otherwise.

We estimate the coefficients β_i by logistic regression, where we treat every singles match as one observation. Matches usually have multiple entries in the dataset (one for each player), but we remove all duplicate match entries, so that for each match, one entry remains for one of the two players who participated in the match. Player A in (1.1) is the reference player for which the match is entered in the dataset.

The estimates are presented in Table 1.1. The first column focuses on the role played by rating differences by excluding all control variables. This model based purely on rating differences correctly predicts 70% of all singles match outcomes. The negative coefficient indicates that when a player's rating increases compared to that of her opponent (note that a worse player has a larger rating), her odds of winning go down. Specifically, compared to the baseline of having equal ratings, improving one's rating by 1 increases the probability of winning from 50% to 85%. A rating difference of 2 implies a win probability of 97% for the stronger player.

The second column includes controls for playing a home game and age differences. We find evidence for a significant home advantage: Given the rating and age differences between the two players, playing a home game increases the odds of winning by 18.3%. In a match that is equal in terms of age and rating, the home player has a win probability of 52.8%. The negative coefficient on the age difference shows that given the rating difference between two players and the home advantage, the older player has lower odds of winning the match. The model predicts that someone playing a home match against a 20-years-younger player with the same rating has a probability of winning of only 42.8%.

The third and fourth column of Table 1.1 show the estimates for matches between two females and between two males, respectively. Comparing logistic regression coefficients across groups can be prob-

lematic, but the results suggest that the effects are similar to those discussed above for all singles matches. Regressions where we also control for the province where each player is based do not give much new insight: Results are mixed across specifications for female and male players, and very few provinces have significantly different odds. The other coefficients are practically unchanged by taking into account regional differences.

Table 1.1: Singles estimates

Logit(p_A)	All		Female	Male
Intercept	–	−0.0568	−0.0560	−0.0671
	–	(0.0039)	(0.0071)	(0.0050)
Home_A	–	0.168	0.156	0.184
	–	(0.0055)	(0.010)	(0.0070)
$R_A - R_B$	−1.74	−1.75	−1.80	−1.71
	(0.0045)	(0.0045)	(0.0082)	(0.0056)
$Age_A - Age_B$	–	−0.0201	−0.0169	−0.0212
	–	(0.00025)	(0.00045)	(0.00030)
Correct (%)	70.0	71.9	72.5	72.2
N (matches)	753,224	749,344	234,543	468,458
Controls	N	Y	Y	Y

Note. Standard errors are given in parentheses. Correct (%) indicates the percentage of matches for which the outcome is correctly predicted by the model. Player A is the reference player for which the match is entered in the dataset. Where indicated, controls are included for playing a home match, and for the age difference between the reference player and her opponent. The Female and Male columns refer to matches between two females and between two males, respectively.

1.3.2 Doubles rating

We now consider the case where two teams of two players play a match. We want to estimate a model similar to (1.1), but with single player

ratings and ages replaced by effective team ratings and ages. We explain this for the rating, as for age the analysis is similar.

We assume that the effective rating of a team, consisting of a stronger player A and a weaker player B ($R_A \leq R_B$), is given by a weighted average of the ratings of those two players:

$$R_{AB} = \theta R_A + (1-\theta) R_B, \tag{1.2}$$

where $0 \leq \theta \leq 1$ measures the degree to which the stronger player carries the team. We denote by $\langle R \rangle_{AB}$ the average rating of team AB:

$$\langle R \rangle_{AB} = \frac{R_A + R_B}{2}. \tag{1.3}$$

Now consider the case where team AB plays team CD, consisting of stronger player C and weaker player D. The difference between the effective ratings is given by

$$\begin{aligned} R_{AB} - R_{CD} &= \theta R_A + (1-\theta) R_B - [\theta R_C + (1-\theta) R_D] \\ &= \left(\theta - \frac{1}{2}\right) \left(\Delta^R_{AB} - \Delta^R_{CD}\right) + \langle R \rangle_{AB} - \langle R \rangle_{CD}, \end{aligned}$$

where we have defined

$$\Delta^R_{AB} = R_A - R_B, \quad R_A \leq R_B, \tag{1.4}$$

and used that

$$R_A + R_B = R_C + R_D + 2(\langle R \rangle_{AB} - \langle R \rangle_{CD}).$$

Similarly, we model the effective age of the team to be a weighted average of the age of its two players:

$$\mathrm{Age}_{XY} = \Phi \mathrm{Age}_X + (1-\Phi) \mathrm{Age}_Y,$$

where X is the older player, so that $0 \leq \Phi \leq 1$ measures the weight on the older player.

We can now define the doubles equivalent of (1.1) for a match between team AB and team CD:

$$\begin{aligned}\text{logit}(p_{AB}) &= \beta_0 + \beta_1 \text{Home}_{AB} + \beta_2 \left(R_{AB} - R_{CD}\right) + \beta_3 \left(\text{Age}_{AB} - \text{Age}_{CD}\right) \\ &= \beta_0 + \beta_1 \text{Home}_{AB} + \beta_2 \left(\theta - \frac{1}{2}\right) \left(\Delta^R_{AB} - \Delta^R_{CD}\right) + \\ &\quad \beta_2 \left(\langle R \rangle_{AB} - \langle R \rangle_{CD}\right) + \beta_3 \left(\Phi - \frac{1}{2}\right) \left(\Delta^{\text{Age}}_{AB} - \Delta^{\text{Age}}_{CD}\right) \\ &\quad + \beta_3 \left(\langle \text{Age} \rangle_{AB} - \langle \text{Age} \rangle_{CD}\right).\end{aligned}$$

Table 1.2 presents some results on the estimates of θ and Φ.

Table 1.2: Doubles estimates: All, female, male

Logit(p_{AB})	All	Female	Male
θ	0.524	0.519	0.535
	(0.0012)	(0.0019)	(0.0019)
Φ	0.517	0.470	0.512
	(0.0076)	(0.015)	(0.011)
Correct (%)	73.8	74.5	73.9
N (matches)	1,526,808	616,209	585,242

Note. Controls are included for playing a home match, for the difference between the average age of the two teams, and for the difference in age differences within the two teams. The Female and Male columns refer to matches between four females and between four males, respectively.

This suggests that the stronger player carries the team to a higher level, as the estimate for θ is larger than 0.5. Note that although θ is statistically significantly different from 0.5, it is not very large. In designing a rating system it might make sense to keep it at 0.5 for simplicity. The effect of a team's unbalancedness on rating outcomes could be dampened by increasing its variance, for example. Similar to singles matches, teams that are older on average have lower odds of winning, given their effective rating. Asymmetries with respect to age do not matter as much however, with Φ barely larger than 0.5. The female and male estimates do suggest that rating asymmetry is

more important in male doubles matches, as can be seen by the larger value for θ. The difference might be negligible in terms of practical implementation in a new rating system, however.

1.4 Proposal for a new system

In this section we propose two rating methods. A relatively simple method based on the Elo rating and a more advanced method based on the Glicko rating. We first show how this rating system would work for singles and then we discuss the extension to doubles.

1.4.1 Elo method for singles

In this rating system, each player has a rating which signifies the players strength. Suppose we have a match of a player with rating R_1 against a player rating R_2, then using the Elo method the probability that player 1 wins can be computed by

$$prob = \frac{1}{1 + e^{-q(R_1 - R_2)}}$$

We can tune the constant q such that these probabilities match what we found for the original ratings in the previous sections. When player 1 wins, his rating should improve and when he loses his rating should deteriorate. We can compute the new rating of player 1 by

$$R_{1,new} = R_{1,old} + K(prob - result)$$

Here *result* is 0, when player 1 loses and 1 when he wins. This new rating can then be used to compare other games. Note that for an expected win, $prob - result$ is small and thus the change in ratings is small, however for an upset win the change in ratings is much larger. The K-factor is also used to scale the size of the ratings change. For young players this should be higher as their skills improves faster and for professional players this factor should be smaller.

One of the main benefits of this system is that it treats all the games equally. Every game gives a change in rating and the rating always improves when the player wins. It would also be a good idea

to make a table of the formula for the win probability, as is also done in chess. This makes it easier to understand for the players what they can expect and how their rating is changed.

1.4.2 Glicko method for singles

An alternative to using the Elo system for singles ratings is to use the Glicko system. The Glicko system can be seen as an extension to the Elo system. Thus in comparison to the Elo system it is more difficult, but also more accurate. The main improvement is that it can better handle big differences in the number of matches played by different people; and as a consequence new players can be added without a special procedure.

In this section we first discuss how the implementation works and then consider the positives and negatives of this system.

In the case of tennis singles it is actually possible to implement standard Glicko, as written for chess. The only change we make in this section is to allow for a change to the 1-9 scale of tennis.

The strength of every player R has two variables, their rating μ, and the variance σ. The rating μ is a measure of how good a player is on average. The variance measures how well we know the rating; it will be higher for new players and less-active players and lower for players who play lots of matches and who play very consistently.

Update for a match

Suppose players 1 and 2 compete and player 1 wins. For the purposes of this subsection we assume the original strength of the two players are R_1 and R_2 with ratings μ_1 and μ_2 and the associated variances are σ_1^2 and σ_2^2. These should be up to date (see the next subsection), and for new players they should be generated first (see the subsequent subsection).

For the new ratings of the players we get

$$g(\sigma^2) = \frac{1}{\sqrt{1 + 3q^2\sigma^2/\pi^2}}$$

$$E_1 = \frac{1}{1 + e^{-qg(\sigma_2^2)(\mu_1 - \mu_2)}} \qquad E_2 = \frac{1}{1 + e^{-qg(\sigma_1^2)(\mu_2 - \mu_1)}}$$

$$\sigma_{1,new}^2 = \frac{\sigma_1^2}{1 + q^2\sigma_1^2 g(\sigma_2^2)^2 E_1(1 - E_1)} \qquad \sigma_{2,new}^2 = \frac{\sigma_2^2}{1 + q^2\sigma_2^2 g(\sigma_1^2)^2 E_2(1 - E_2)}$$

$$r_{1,new} = \mu_1 + q\sigma_{1,new}^2 g(\sigma_2^2)(1 - E_1) \qquad r_{2,new} = \mu_2 - q\sigma_{2,new}^2 g(\sigma_1^2) E_2,$$

where q is a scaling parameter which for chess is taken (standard) to be $q_{\text{chess}} = \ln(10)/400$. To adjust to the tennis scale a different value can be used, the choice should be inferred from the data.

While the formulas look somewhat imposing, anyone with a calculator can easily calculate the results.

As with Elo, the new rating is equal to the old rating plus a factor times obtained score (1 for winning or 0 for losing) minus the expected score of the match (the probability of winning). This prefactor is varies according to the variance of the player: Uncertain ratings mean high variance thus big adjustments to the rating, and if the confidence in a rating is high, the adjustment to the rating based on a single match is low. As a secondary factor, when the rating of the opposing player is uncertain the change in rating is reduced.

The formula for the new variance means that it is slightly less than the old variance. If the match was unbalanced, so the outcome was almost predetermined (a very good vs. a very bad player), the variance does not decrease by much, as this match does not give a lot of information, likewise if the rating of the opponent is very uncertain.

Introducing new players

For new players you initially do not know anything. Therefore the initial rating is placed at an average value and the initial variance is made to be very large.

$$\mu = \mu_{\text{init}} \qquad \sigma^2 = \sigma_{\text{init}}^2$$

Here μ_{init} and σ^2_{init} are parameters that can be chosen. The only effect of r_{init} is to move the ratings a fixed amount up or down, so it can be adjusted to fit the current rating distribution if desired. The choice for σ^2_{init} does not have a big impact on the ratings. A reasonable value is

$$\sigma^2_{\text{init}} = \frac{4}{q^2}.$$

Update for time

Before each match the variances of the players should be made up to date. This means that the variance σ^2 for a player is increased by the passage of time. The longer someone is inactive, the less certain we are of their rating.

For the time update a new parameter ν^2 of the system is introduced. The time update follows the formula

$$\sigma^2(t + \Delta t) = \min(\sigma^2(t) + \nu^2 \Delta t, \sigma^2_{\text{init}})$$

Here Δt is the change in time passed between the previous update of the ratings and the new one. The minimum is taken to ensure the new variance is never higher than the variance for a new player. Basically this only happens if a player has retired, or there is a long time between the first match of a new player and the second match.

Implementationwise it suffices to measure time in days (or even weeks or months), so Δt can be the number of days since the last match. Updating the variances every day seems cumbersome, so it might be wise to give each player a "time of last update"-variable and only update the variance of a player when you need it: Either when the player participates in a match, or when a ratings list is published.

A good choice of ν^2 should be inferred from the data. As a guideline you can use that with $\Delta t = \sigma^2_{\text{init}}/\nu^2$ any variance will become σ^2_{init}, so any rating will be completely uncertain after this period of time.

Presenting the ratings

As this system is a bit more complicated some thought must be put in how to present the resulting rating list for less mathematically inclined people. Even if people cannot follow the way the ratings are

calculated, they should understand the outcomes of the rating system. The rating list consists of two numbers for every player. One number, the rating gives the strength of that player, and big differences in that number correspond to big differences in strengths, which should be understandable for everyone. In particular this is the variable upon which you want to sort the ratings.

The variance for a player is more complicated to interpret. It is suggested to suppress the numerical value from any rating list (though it should be accessible for the mathematically inclined, perhaps after a click or two). Instead the ratings can be coded to give the level of confidence in the ratings. The coding can be done by a color or stars or something. In particular uncertain ratings should be designated as such (otherwise new players who win their first two matches might end up very high on the list). It might also be wise to exclude people from any ranking lists if the variance is too large. The exact cutoff points for these codes are of course up to the rating administrators, and should be decided upon inspection of the actual data.

Another option to show the variance is to give a confidence interval $(\mu - 2\sigma, \mu + 2\sigma)$, which gives 95% confidence that the interval contains the actual strength. With this one can only really say that they are stronger than someone else if their confidence intervals don't overlap.

1.4.3 Extension to doubles

Our rating system for doubles relies on computing a team rating. Then we pit the teams against each other as if it was a singles match. As in section 3, we use a weighted average of the players ratings to compute a team rating. Suppose we have a team with the stronger player with rating R_1 and with the weaker player with rating R_2, then the team rating R_{12} is given by

$$R_{12} = \theta R_1 + (1 - \theta) R_2$$

If we take $\theta = \frac{1}{2}$, take we take a simple average and each team member contributes equally. However, if $\theta = 1$ then it is basically a match between the two strongest players and the weaker players have no impact on the game. Currently the KNLTB uses the rating of the strongest

player to determine tournament eligibility, so it basically assumes $\theta = 1$ for this case.

For Glicko, we can do a similar thing for the team rating μ_{12} and variance σ_{12}. Now we can use the formulas above for either the Elo or Glicko methods to compute a new team rating. Then we can divide the change in rating between the individual players, such that this matches the new team rating. Here it used that the player who contributed the most to the team rating (based on θ) will also get the largest share of the rating change.

There is also a different way to think about this problem. We can view a doubles match as a 4 player game, where 2 players win. This alternate view gives the same changes in rating for the Elo method, but there is a slight difference in the Glicko method.

1.5 Theoretical analysis of the proposal

In this section we dive deeper into the mathematical aspects of the rating systems. We will go in detail how the win probabilities and ratings changes are derived using Bayesian statistics. We will also give two ways to compute the Glicko update for doubles, via comparison of team ratings (section 5.4) and as a 4 player game (section 5.5).

1.5.1 Relation between rating difference and win probability

A one-sided game is to be expected in the event of a large disparity between player ratings. To reflect for such relative ease/difficulty of a match, the win probability serves as an important component in determining the new ratings of both players, by compensating fairly to those who has seemingly over/under-performed in a given match. Since tennis matches have a binary outcome (either a win or a loss), the win probability is equivalent to the expected score (which is generally the win probability plus half the probability of a draw), and we use these two terms interchangeably throughout the text.

In both Elo and Glicko systems, the win probability is calculated using the difference between current player ratings and a scaling factor q

(taken to be $-\ln(10)/400$ for chess). We examine exactly how these probabilities are calculated, starting with the Elo system.

The Elo rating system assumes that the true rating of a player i follows a logistic distribution with mean μ_i and a scale q, for $i = 1, 2$. It follows that the expected score for player 1 against player 2, or the win probability for player 1 against player 2, is

$$E_1 = \frac{1}{1 + e^{-q(\mu_1 - \mu_2)}}.$$

The Glicko rating system assumes that the true rating R_i of a player i is normally distributed with mean μ_i and variance σ_i^2 for $i = 1, 2$. The win probability of player 1 against player 2, is, as in the case of the Elo rating system, given by the distribution function of the logistic distribution with scale q.

$$\mathbb{P}(1 \text{ wins against } 2 | R_1, R_2) = \frac{1}{1 + e^{-q(R_1 - R_2)}}. \tag{1.5}$$

Let $\mathcal{N}(r; \mu, \sigma^2)$ denote the probability density function of a normal distribution with mean μ and variance σ^2, evaluated at the point r. Then we may estimate the expectation of player 1 winning against player 2 as follows.

$$\begin{aligned}
\mathbb{E}[1 \text{ wins against } 2 | R_1] &= \int_{-\infty}^{\infty} \mathbb{P}(1 \text{ wins against } 2 | R_1, R_2 = r_2) \mathcal{N}(r_2; \mu_2, \sigma_2^2) dr_2 \\
&= \int_{-\infty}^{\infty} \frac{1}{1 + e^{-q(R_1 - r_2)}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(r_2 - \mu_2)^2}{2\sigma_2^2}} dr_2 \\
&\approx \frac{1}{1 + e^{-qg(\sigma_2^2)(R_1 - \mu_2)}},
\end{aligned}$$

where the approximation from Glickman (1999) is used in the final line, and where

$$g(\sigma^2) = \frac{1}{\sqrt{1 + 3q^2\sigma^2/\pi^2}}.$$

Finally, we substitute $R_1 = \mu_1$ as this is a reasonable approximation of player 1's current rating.

$$E_1 = \frac{1}{1 + e^{-qg(\sigma_2^2)(\mu_1 - \mu_2)}}.$$

1.5.2 Derivation of the Elo update for singles and doubles

For singles matches, the Elo update has a fairly straightforward derivation. Given players 1 and 2, with Elo ratings μ_1 and μ_2, we model the *performance* of each player as the random variables

$$P_1 \sim \mathcal{L}(\mu_1, s/\log 10), \qquad P_2 \sim \mathcal{L}(\mu_2, s/\log 10),$$

where $\mathcal{L}(\mu, s)$ is the logistic distribution with mean μ and scale s. Then the *expected score* of A vs B (for a single game) is given by

$$E_A = \Pr(P_1 > \mu_2) = \frac{1}{1 + 10^{(\mu_2 - \mu_1)/s}} = \frac{Q_A}{Q_A + Q_B} \tag{1.6}$$

where $Q_A := 10^{\mu_1/s}$ and likewise for Q_B. Then given the actual score S_A of A vs B, we update the rating of A linearly:

$$\mu_{1,new} := \mu_1 + K(S_A - E_A) \tag{1.7}$$

and likewise for B.

For doubles matches, we apply the method that we will go into more detail on in the Glicko case in section 1.5.4. That is, given players A and B forming a doubles team AB, we combine the Elo ratings μ_1, μ_2 into a team Elo rating μ_{12}. Then when AB play CD we update the team ratings via the above formulae to find $\mu_{12,new}$. Then we separate this rating to find $\mu_{1,new}$ and $\mu_{2,new}$.

By the same argument as in Theorem 1.5.1, the only reasonable and scale-invariant combination method is (choosing labels so that $\mu_1 \leq \mu_2$):

$$\mu_{12} = \theta\mu_1 + (1-\theta)\mu_2 \tag{1.8}$$

where $\theta \in [0, 1]$ is a parameter we infer from data (see section 1.3.2).

Then, given an updated team Elo rating $\mu_{12,new}$, we update the individual ratings by finding the new ratings nearest to the old ones which combine to form the updated team rating, i.e. we solve

$$\min(\mu_{1,new} - \mu_1)^2 + (\mu_{2,new} - \mu_2)^2 \text{ s.t. } \theta\mu_{1,new} + (1-\theta)\mu_{2,new} = \mu_{12,new}$$

which gives updated ratings

$$\mu_{1,new} = \mu_1 + \frac{\theta}{\theta^2 + (1-\theta)^2} \left(\mu_{12,new} - \mu_{12}\right), \tag{1.9}$$

$$\mu_{2,new} = \mu_2 + \frac{(1-\theta)}{\theta^2 + (1-\theta)^2} \left(\mu_{12,new} - \mu_{12}\right). \tag{1.10}$$

1.5.3 Derivation of the Glicko update for singles

The Glicko system has two types of updates, one updates the ratings after a match, one updates the ratings after a passage of time. We consider both in the following, but first we detail the general setting. The derivations here were obtained by Glickman in Glickman (1999).

We assume all players have at all times an intrinsic strength R, which we do not know exactly. We do, however have an estimate for this strength determined by the rating and variance, which means that given rating and variance at a certain time we have $R \sim \mathcal{N}(\mu, \sigma^2)$. We assume that all these ratings are independent of each other.

Match update

For simplicity we consider a single match between players 1 and 2. Both players have their own strengths R_i, and our estimates of those strengths, the ratings μ_i and the variances σ_i^2. The probability that a player 1 wins from player 2, given the hypothesis $R_i = r_i$ is given by the distribution function of the logistic distribution:

$$P(1 \text{ wins from } 2) = \frac{e^{q(r_1 - r_2)}}{1 + e^{q(r_1 - r_2)}}$$

Likewise we have

$$P(1 \text{ loses from } 2) = 1 - \frac{e^{q(r_1 - r_2)}}{1 + e^{q(r_1 - r_2)}} = \frac{1}{1 + e^{q(r_1 - r_2)}} = \frac{e^{q(r_2 - r_1)}}{1 + e^{q(r_2 - r_1)}}$$

We can combine these two formulas to

$$P(\text{player 1 wins from opponent } 2) = \frac{e^{sq(r_1 - r_2)}}{1 + e^{q(r_1 - r_2)}},$$

where $s = 1$ if player 1 wins, and $s = 0$ if player 1 loses.

Now we can use a Bayesian analysis. The prior is given by

$$prior = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(r_2-\mu_2)^2}{2\sigma_2^2}} = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2} - \frac{(r_2-\mu_2)^2}{2\sigma_2^2}}$$

The likelihood of the result is given by the probability that the outcome occurs, so

$$llh = \frac{e^{sq(r_1-r_2)}}{1+e^{q(r_1-r_2)}} = \frac{1}{1+e^{(1-2s)q(r_1-r_2)}}, \qquad s = 0, 1.$$

Using Bayes' rule we find that the posterior distribution is proportional to

$$post(r_1, r_2) \propto prior \cdot llh = e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2} - \frac{(r_2-\mu_2)^2}{2\sigma_2^2}} \frac{1}{1+e^{(1-2s)q(r_1-r_2)}},$$

where we removed R_i-independent factors as we only get a proportionality relation anyway. You will find that in the new posterior distribution there is a positive correlation between the ratings of the two players, so they are clearly not independent. The posterior distribution is also not a normal distribution. However we want to approximate this posterior distribution with a distribution where R_1 and R_2 are independent normally distributed random variables to bring it back to our framework. Therefore we first take the marginal distribution for player 1, that is

$$margpost(r_1) \propto \int_{-\infty}^{\infty} post(r_1, r_2)\, dr_2$$

This integral cannot be expressed in terms of simple functions, so we have to approximate this integral. We do this by first approximating the distribution function of the logarithmic distribution by the distribution function of a normal distribution with identical mean and variance, after which the integral can be evaluated and expressed in terms of a normal distribution function (with different mean and variance). We can subsequently substitute a logistic distribution function with identical mean and variance for the distribution function of the new normal distribution.

Thus writing

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt$$

for the distribution function of the standard normal distribution we use

$$\frac{1}{1+e^{-x}} \approx \Phi(\frac{\sqrt{3}}{\pi}x)$$

and we get

$$\begin{aligned}
marg post(r_1) &\propto \int_{-\infty}^{\infty} e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2} - \frac{(r_2-\mu_2)^2}{2\sigma_2^2}} \frac{1}{1+e^{(1-2s)q(r_1-r_2)}}\, dr_2 \\
&\approx \int_{-\infty}^{\infty} e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2} - \frac{(r_2-\mu_2)^2}{2\sigma_2^2}} \Phi\left(\frac{\sqrt{3}}{\pi}(1-2s)q(r_1-r_2)\right) dr_2 \\
&= \sigma_2 e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}} \Phi\left(\frac{\sqrt{3}q(1-2s)(r_1-\mu_2)}{\sqrt{\pi^2+3q^2(1-2s)^2\sigma_2^2}}\right) \\
&\approx \sigma_2 e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{1+e^{-\frac{\pi q}{\sqrt{\pi^2+3q^2(1-2s)^2\sigma_2^2}}(1-2s)(r_1-\mu_2)}}
\end{aligned}$$

We can simplify this expression slightly by observing that $(1-2s)^2 = 1$ for both $s = 0$ and $s = 1$. Also, recall the definition of the function g:

$$g(\sigma^2) = \frac{1}{\sqrt{1+\frac{3q^2}{\pi^2}\sigma^2}}.$$

Thus we get

$$marg post(r_1) \propto e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{1+e^{-g(\sigma_2^2)q(1-2s)(r_1-\mu_2)}}$$

Next we want to approximate this marginal posterior distribution with a normal distribution. As an approximation of the mean of the normal distribution we take the mode of this marginal posterior distribution. For the variance we take $\sigma^2_{new} = -\left(\frac{d^2 \ln(marg post(r_1))}{dr_1^2}\right)^{-1}$, which would be constant if the distribution truly was a normal distribution. To find

the mode we need to solve that the derivative equals 0, and for easy of calculations we actually consider the log-derivative. This gives

$$\frac{d\ln(margpost(r_1))}{dr_1} = -\frac{r_1-\mu_1}{\sigma_1^2} + \frac{g(\sigma_2^2)q(1-2s)e^{-g(\sigma_2^2)q(1-2s)(r_1-\mu_2)}}{1+e^{-g(\sigma_2^2)q(1-2s)(r_1-\mu_2)}}$$

Solving this equation equals 0 is not possible in terms of elementary functions, so we approximate the zero using a single Newton-Raphson iteration, starting at $r_1 = \mu_1$ (this is a sensible starting point as you would expect the new rating to be close to the old one). This gives

$$\mu_{1,new} = \mu_1 - \frac{\frac{d\ln(margpost)}{dr_1}(\mu_1)}{\frac{d^2\ln(margpost)}{dr_1^2}(\mu_1)}$$

Now observe that we have the second derivative of the log-marginal posterior appearing both in the expression for the new rating, as the expression for the new variance. In the latter expression it was still unclear at which point to evaluate this function (though it should not matter too much). For consistency it now makes sense to evaluate the new variance at the same point, $r_1 = \mu_1$; this also makes sense as μ_1 is a reasonable approximation of the new rating. This last choice fixes the approximations. Thus we get

$$E_1 = \frac{1}{1+e^{-qg(\sigma_2^2)(\mu_1-\mu_2)}},$$

$$\sigma_{1,new}^2 = -\frac{1}{\frac{d^2\ln(margpost)}{dr_1^2}(\mu_1)} = -\frac{1}{-\frac{1}{\sigma_1^2} - g(\sigma_2^2)^2q^2E_1(1-E_1)},$$

$$\mu_{1,new} = \mu_1 + \sigma_{1,new}^2\frac{d\ln(margpost)}{dr_1}(\mu_1) = \mu_1 + \sigma_{1,new}^2qg(\sigma_2^2)(s-E_1),$$

(where we again simplify the formulas by using $s = 0, 1$). You should now be able to recognize these formulas as identical to what we had before.

Time update

In this part we only consider the ratings of a single player. Thus we can use the subscripts to refer to the time at which we take the parameter, instead of using subscripts to denote whose rating we are talking about.

We model the change in strength over time as following a Brownian motion. That means that the strength $r_{t+\Delta t}$ at time $t+\Delta t$ given the value of the strength r_t at time t is chosen from a normal distribution around r_t with a variance which is a multiple of the change Δt in time. Thus $r_{t+\Delta t} - r_t \sim \mathcal{N}(0, \nu^2 \Delta t)$ for some parameter ν^2. We assume that this change in rating is independent of the ratings of all players (and also of the rating of the relevant player at time t).

Since we have assumed $R_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ to begin with we thus obtain that $R_{t+\Delta t} = (R_{t+\Delta t} - R_t) + R_t \sim \mathcal{N}(0, \nu^2 \Delta t) + \mathcal{N}(\mu_t, \sigma_t^2) = \mathcal{N}(\mu_t, \sigma_t^2 + \nu^2 \Delta t)$. Here we use that the sum of two independent normally distributed random variables is again a normally distributed random variable. We thus observe that for the time update we have

$$\mu_{t+\Delta t} = \mu_t, \qquad \sigma_{t+\Delta t}^2 = \sigma_t^2 + \nu^2 \Delta t.$$

1.5.4 Derivation of the Glicko update for doubles by information projection

The basic idea of this method is to apply the following three steps to reduce the 2v2 match to a single 1v1 Glicko update:

1. **Combination** Given players $\mathbf{x} = \{x_i\}_{i=1}^2$ with ratings (μ_i, σ_i), construct a Glicko rating $(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}})$ for the team $\mathbf{x}$.

2. **Match update** When $\mathbf{x}$ play $\mathbf{y}$, given $(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}})$, $(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$ and the result of the match, use Glicko to update the combined rating to $(\mu'_{\mathbf{x}}, \sigma'_{\mathbf{x}})$.

3. **Separation** Given $(\mu'_{\mathbf{x}}, \sigma'_{\mathbf{x}})$ and the original ratings (μ_i, σ_i), construct the updated ratings (μ'_i, σ'_i).

Combination

We have players $\mathbf{x} = \{x_i\}_{i=1}^2$ with 'true ratings'

$$R_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

and we wish to combine these into

$$R_{\mathbf{x}} \sim \mathcal{N}(\mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2).$$

We note some properties that this combination should obey:

I. If $\forall i\ R_i \mapsto R_i + c$, then $R_{\mathbf{x}} \mapsto R_{\mathbf{x}} + c$ as Glicko ratings are shift-invariant.

II. $R_{\mathbf{x}}$ is independent of the ordering of the x_i.

III. $\mu_{\mathbf{x}}$ is monotonic increasing in each of the μ_i.

IV. If $\mu_i \equiv \mu$, then $\mu_{\mathbf{x}} = \mu$.

Theorem 1.5.1. *If $R_{\mathbf{x}}$ obeys* (I)-(IV)*, and furthermore we impose that*

$$\mu_{\mathbf{x}} = f(\mu_1, \mu_2) \tag{1.11}$$

i.e., that the mean of the combined rating depends only on the means of the players, and

$$\text{if } \forall i\ R_i \mapsto kR_i \text{ for } k > 0, \text{ then } R_{\mathbf{x}} \mapsto kR_{\mathbf{x}} \tag{1.12}$$

then f must have the form:

$$f(\mu_1, \mu_2) = \begin{cases} a\mu_1 + b\mu_2, & \text{if } \mu_1 \geq \mu_2, \\ b\mu_1 + a\mu_2, & \text{if } \mu_2 > \mu_1, \end{cases} \tag{1.13}$$

for $a, b \geq 0$ with $a + b = 1$.

Proof. By (I) and (1.12), for any $k > 0$ we have that

$$f(\mu_1, \mu_2) = \mu_1 + kf(0, k^{-1}(\mu_2 - \mu_1)).$$

By (IV), it suffices to consider the case $\mu_1 \neq \mu_2$. Then, setting $k = |\mu_2 - \mu_1|$,

$$f(\mu_1, \mu_2) = \begin{cases} \mu_1 + (\mu_1 - \mu_2)f(0, -1), & \text{if } \mu_1 > \mu_2, \\ \mu_1 + (\mu_2 - \mu_1)f(0, 1), & \text{if } \mu_2 > \mu_1. \end{cases}$$

By (II) $f(\mu_1, \mu_2) = f(\mu_2, \mu_1)$, so

$$\left\{ \begin{array}{c} \mu_1 + (\mu_1 - \mu_2)f(0, -1) = \mu_2 + (\mu_1 - \mu_2)f(0, 1) \\ \text{or} \\ \mu_1 + (\mu_2 - \mu_1)f(0, 1) = \mu_2 + (\mu_2 - \mu_1)f(0, -1) \end{array} \right\},$$

which implies that $f(0,1) = 1 + f(0,-1)$. Finally, recalling (III), we define $a := f(0,1) \geq f(0,0)$ and $b := -f(0,-1) \geq -f(0,0)$. By (IV), $f(0,0) = 0$, so $a, b \geq 0$ and $a + b = 1$, and the result follows. □

By (III), without loss of generality we can suppose that $\mu_1 \leq \mu_2$. Then Theorem 1.5.1 suggests that we should combine the ratings as a weighted average with weights θ corresponding to a player's skill ranking within the team. This allows for the possibility that strong players/weak players may have a larger impact on the outcome than others, e.g. a strong player carrying the team. Mathematically

$$R_{\mathbf{x}} = \sum_i \theta_i R_i \tag{1.14}$$

where $\theta_i \geq 0$, $\sum_i \theta_i = 1$ and the values of θ are inferred from data (see section 1.3.2).

Then supposing that the R_i are independent, the standard formulae for linear sums of independent Gaussian variables give:

$$\mu_{\mathbf{x}} = \sum_i \theta_i \mu_i, \tag{1.15}$$

$$\sigma_{\mathbf{x}}^2 = \sum_i \theta_i^2 \sigma_i^2. \tag{1.16}$$

Separation

After the Glicko update, we now have $\mu'_{\mathbf{x}}$ and $\sigma'_{\mathbf{x}}$ describing some

$$R'_{\mathbf{x}} \sim \mathcal{N}(\mu'_{\mathbf{x}}, \sigma'^2_{\mathbf{x}}).$$

We suppose that this random variable can be modelled as

$$R'_{\mathbf{x}} = \sum_i \theta_i R'_i. \tag{1.17}$$

This then gives

$$\mu'_{\mathbf{x}} = \sum_i \theta_i \mu'_i \tag{1.18}$$

$$\sigma'^2_{\mathbf{x}} = \sum_i \theta_i^2 \sigma'^2_i \tag{1.19}$$

and we seek to recover the new ratings $R'_i \sim \mathcal{N}(\mu'_i, \sigma'_i)$. One problem which makes this hard is that formally speaking the new ratings are correlated with the ratings of the other players. However, for our method we require that the strengths of all players are independent. We solve this by finding the 'smallest' change from the old ratings $R_i \sim \mathcal{N}(\mu_i, \sigma_i)$ to new independent ratings $R'_i \sim \mathcal{N}(\mu'_i, \sigma'_i)$ which give the correct new team rating R'_x. This change should be 'as small as possible' as you don't want to change anything besides what you learned from the match.

We formulate this as the problem of minimising the Kullback–Leibler divergence Kullback and Leibler (1951) of the posterior ratings s.t. the combination condition:

$$\min \sum_i D_{KL}(R'_i||R_i) \text{ s.t. } R'_{\mathbf{x}} = \sum_i \theta_i R'_i. \tag{1.20}$$

In words, out of all the updated ratings for the players $\mathbf{x}$ that would give rise to $R'_{\mathbf{x}}$, we choose the ones that minimise the amount of information gained in moving from the old ratings to the new ratings. As R'_i and R_i are assumed to be all independent Gaussian variables we can explicitly write the Kullback–Leibler divergences as:

$$D_{KL}(R'_i||R_i) = \log(\sigma_i) - \log(\sigma'_i) + \frac{\sigma'^2_i + (\mu'_i - \mu_i)^2}{2\sigma_i^2} - \frac{1}{2}. \tag{1.21}$$

So the problem becomes

$$\min \sum_i -\frac{1}{2}\log(\sigma'^2_i) + \frac{\sigma'^2_i + (\mu'_i - \mu_i)^2}{2\sigma_i^2} \text{ s.t. } (1.18), (1.19). \tag{1.22}$$

This decouples for the μ' and σ'.

Solving for μ' The problem for μ' becomes

$$\min \sum_i \frac{(\mu'_i - \mu_i)^2}{2\sigma_i^2} \text{ s.t. } \mu'_{\mathbf{x}} = \sum_i \theta_i \mu'_i.$$

Which has solution

$$\mu'_i = \mu_i + \frac{\theta_i \sigma_i^2}{\sum_j \theta_j^2 \sigma_j^2} (\mu'_{\mathbf{x}} - \mu_{\mathbf{x}}). \tag{1.23}$$

Solving for σ' Next, the problem for the σ' becomes

$$\min \sum_i -\frac{1}{2}\log(\sigma_i'^2) + \frac{\sigma_i'^2}{2\sigma_i^2} \text{ s.t. } \sum_i \theta_i^2 \sigma_i'^2 = \sigma_{\mathbf{x}}'^2.$$

Let $v_i := \sigma_i'^2 \sigma_i^{-2}$, $\beta_i := \theta_i^2 \sigma_i^2$. Then we seek to solve

$$\min f(v) := \sum_i v_i - \log v_i \text{ s.t. } \sum_i \beta_i v_i = \sigma_{\mathbf{x}}'^2.$$

Note that f is strictly convex and bounded below, and that the constraint is linear, so this problem has a unique minimiser v^*. For some dual variable $\nu \in \mathbb{R}$, it is straightforward to show that for all i,

$$v_i^* = (1 + \nu\beta_i)^{-1}$$

which gives update rule

$$\sigma_i'^2 = \frac{\sigma_i^2}{1 + \nu\theta_i^2\sigma_i^2}. \tag{1.24}$$

Note that we can write this in Glicko-like terms as

$$\sigma_i'^{-2} = \sigma_i^{-2} + \nu\theta_i^2. \tag{1.25}$$

Plugging into the constraint we get that ν is the unique solution to

$$\sigma_{\mathbf{x}}'^2 = \sum_i (1 + \nu\beta_i)^{-1}\beta_i = \sum_i \frac{\theta_i^2\sigma_i^2}{1 + \nu\theta_i^2\sigma_i^2}. \tag{1.26}$$

Note that since $\sigma_{\mathbf{x}}'^2 \leq \sigma_{\mathbf{x}}^2$ (by the formula for a Glicko update) we must have $\nu \geq 0$. Multiplying (1.26) out, it becomes

$$\nu^2 + (\theta_1^{-2}\sigma_1^{-2} + \theta_2^{-2}\sigma_2^{-2} - 2\sigma_{\mathbf{x}}'^{-2})\nu + (\theta_1^{-2}\sigma_1^{-2}\theta_2^{-2}\sigma_2^{-2} - (\theta_1^{-2}\sigma_1^{-2} + \theta_2^{-2}\sigma_2^{-2})\sigma_{\mathbf{x}}'^{-2}) = 0$$

which we can solve and take the unique positive root to compute ν, i.e.

$$b := \frac{1}{2}(\theta_1^{-2}\sigma_1^{-2} + \theta_2^{-2}\sigma_2^{-2} - 2\sigma_{\mathbf{x}}'^{-2}),$$

$$c := -(\theta_1^{-2}\sigma_1^{-2}\theta_2^{-2}\sigma_2^{-2} - (\theta_1^{-2}\sigma_1^{-2} + \theta_2^{-2}\sigma_2^{-2})\sigma_{\mathbf{x}}'^{-2}) = \frac{\sigma_{\mathbf{x}}^2 - \sigma_{\mathbf{x}}'^2}{\sigma_{\mathbf{x}}'^2\theta_1^2\sigma_1^2\theta_2^2\sigma_2^2} \geq 0,$$

$$\nu = -b + \sqrt{b^2 + c}.$$

1.5.5 Derivation of the Glicko update for doubles by direct estimation

In this part we use a different approach to derive updating formulas for the doubles system. It follows closely both the method behind the Glicko system as the approximations that were made there.

As above we assume that every participant has a strength parameter r_i, and that we have prior estimates of these strength parameters as being normally distributed around the ratings μ_i, according to $R_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Here we label the participants such that 1 and 2, correspond to the winning team, and 3 and 4 to the losing team. Thus the prior becomes

$$prior \propto e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}-\frac{(r_2-\mu_2)^2}{2\sigma_2^2}-\frac{(r_3-\sigma_3)^2}{2\sigma_3^2}-\frac{(r_4-\sigma_4)^2}{2\sigma_4^2}}$$

We assume the probability of winning is given by the same formula as for singles, but with teams performing as though their rating was the average of the ratings of the members. This is the $\theta = \frac{1}{2}$ case of the previous subsection. We use $\theta = \frac{1}{2}$ mostly for ease of exposition, the method also works for other values of θ. However for $\theta \neq \frac{1}{2}$ the approximations need a slightly new justification.

Thus we find for the likelihood of team 1-2 winning from team 3-4 that

$$llh = \frac{1}{1+e^{-\frac{1}{2}q(r_1+r_2-r_3-r_4)}}$$

Therefore the posterior is equal to

$$post \propto e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}-\frac{(r_2-\mu_2)^2}{2\sigma_2^2}-\frac{(r_3-\sigma_3)^2}{2\sigma_3^2}-\frac{(r_4-\sigma_4)^2}{2\sigma_4^2}} \frac{1}{1+e^{-\frac{1}{2}q(r_1+r_2-r_3-r_4)}}$$

Now we want the marginal posterior for the first participant, giving

$$margpost(r_1) \propto \iiint \frac{e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}-\frac{(r_2-\mu_2)^2}{2\sigma_2^2}-\frac{(r_3-\sigma_3)^2}{2\sigma_3^2}-\frac{(r_4-\sigma_4)^2}{2\sigma_4^2}}}{1+e^{-\frac{1}{2}q(r_1+r_2-r_3-r_4)}} dr_2 dr_3 dr_4.$$

In the derivation of the original Glicko system we derived what amounts

to the approximation

$$\int_{-\infty}^{\infty} \frac{1}{1+e^{-q(x-y)}}\,dy \approx \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{1+e^{-qg(\sigma^2,q)(\mu-y)}},\quad g(\sigma^2,q)=\frac{1}{\sqrt{1+\frac{3q^2}{\pi^2}\sigma^2}}$$

Here we made g explicitly depend on q, as we will need to use the approximation for other values of q. Indeed using this approximation three times, and subsequently simplifying the product of g-functions that arise we find

$$\begin{aligned} marg post(r_1) &\propto e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{1+e^{-\frac{1}{2}qg_2g_3g_4(r_1+\mu_2-\mu_3-\mu_4)}} \\ &= e^{-\frac{(r_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{1+e^{-\frac{1}{2}qg(\sigma_2^2+\sigma_3^2+\sigma_4^2,\frac{1}{2}q)(r_1+\mu_2-\mu_3-\mu_4)}} \\ g_2 = g\left(\sigma_2^2,\frac{1}{2}q\right),&\quad g_3 = g\left(\sigma_3^2,\frac{1}{2}qg_2\right),\quad g_4 = g\left(\sigma_4^2,\frac{1}{2}qg_3\right) \end{aligned}$$

This expression looks very much like the marginal distribution in the singles case. We can once again apply the same approximations (in this case literally the same), to arrive at formulas for the new ratings for player 1.

$$\begin{aligned} E_1 &= \frac{1}{1+e^{-\frac{1}{2}qg(\sigma_2^2+\sigma_3^2+\sigma_4^2,\frac{1}{2}q)(\mu_1+\mu_2-\mu_3-\mu_4)}} \\ \frac{1}{\sigma_{1,new}^2} &= \frac{1}{\sigma_1^2}+\frac{1}{4}q^2g(\sigma_2^2+\sigma_3^2+\sigma_4^2,\frac{1}{2}q)E_1(1-E_1) \\ \mu_{1,new} &= \mu_1+\frac{1}{2}\sigma_{1,new}^2qg(\sigma_2^2+\sigma_3^2+\sigma_4^2,\frac{1}{2}q)(1-E_1) \end{aligned}$$

The formulas for the update of the rating of participant is of course similar, for participants 3 and 4 the big difference is that the change in rating is multiplied by $-E_3$ (respectively $-E_4$) instead of $(1-E_1)$.

The updates of the ratings for a change in time and the introduction of new players is identical to the system for singles.

$\theta \neq \frac{1}{2}$

If we do take $\theta \neq \frac{1}{2}$ the obvious likelihood would be

$$llh = \frac{1}{1 + e^{-q(\theta \max(r_1,r_2)+(1-\theta)\min(r_1,r_2)-\theta\max(r_3,r_4)-(1-\theta)\min(r_3,r_4)}}$$

However this likelihood is non-differentiable (when $r_1 = r_2$ or $r_3 = r_4$), and our approximations become trickier. If we assume $\mu_1 > \mu_2$ and $\mu_3 > \mu_4$ we can instead use the likelihood

$$\frac{1}{1 + e^{-q(\theta r_1+(1-\theta)r_2-\theta r_3-(1-\theta)r_4}},$$

which has the benefit of the derivation of the approximations to proceed as above, but the detriment of breaking the symmetry between r_1 and r_2. Now if μ_1 is much bigger then μ_2, this would not change much, because in practice r_1 will be bigger than r_2. If the difference $\mu_1 - \mu_2$ is small, this does seem a bit stranger. Anyway, the resulting updating formulas are

$$\begin{aligned}
\theta_1 = \theta = \theta_3, \qquad \theta_2 = 1 - \theta = \theta_4 \\
g = \frac{1}{\sqrt{1 + \frac{3q^2\theta_2^2\sigma_2^2}{\pi^2} + \frac{3q^2\theta_3^2\sigma_3^2}{\pi^2} + \frac{3q^2\theta_4^2\sigma_4^2}{\pi^2}}} \\
E_1 = \frac{1}{1 + e^{-\frac{1}{2}qg(\theta_1\mu_1+\theta_2\mu_2-\theta_3\mu_3-\theta_4\mu_4)}} \\
\frac{1}{\sigma_{1,new}^2} = \frac{1}{\sigma_1^2} + \theta_1^2 q^2 g E_1(1 - E_1) \\
\mu_{1,new} = \mu_1 + \theta_1\sigma_{1,new}^2 qg(1 - E_1)
\end{aligned}$$

1.6 Conclusion

We have done a statistical analysis of DSS, the current rating system used by the KNLTB. We found that a difference in rating of 1 increases the odd of winning to 85% and a difference of 2 even to 97% in singles tennis. For doubles tennis we found that the stronger player carries the team more than you should expect if both player contributed equally to

the strength of a team. Although this result was statistically significant, the size of the effect was very small.

We have proposed two alternative rating systems based on the Elo and Glicko methods. One of the main benefits of this system is that every game gives a change in rating and the rating always improves when the player wins. The Glicko model is more advanced by tracking also how certain we are that the assigned rating is correct. This makes the players rating change faster when it is uncertain, but the ratings change is smaller when playing against someone with an uncertain rating. The trade-off is that the Glicko model is more complex and harder to understand intuitively. However for a good implementation, one needs to determine the optimal values for certain parameters in the rating systems.

References

Elo, A.E. (1978). *The rating of chessplayers, past and present.* Arco Pub.

Glickman, Mark E (1999). "Parameter estimation in large dynamic paired comparison experiments". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48.3, pp. 377–394.

Herbrich, Ralf, Tom Minka, and Thore Graepel (2007). "TrueSkill (TM): A Bayesian Skill Rating System". In: *Microsoft Research*.

KNLTB (2017). *"Dynamic Playing Strength" system (DSS) used by the KNLTB (Dutch Tennis Federation) to rank all their players according to their playing level.* `http://www.knltb.nl/tennissers/speelsterkte/`.

Kullback, S. and R. A. Leibler (1951). "On Information and Sufficiency". In: *Ann. Math. Statist.* 22.1, pp. 79–86. DOI: `10.1214/aoms/1177729694`.