Abstract

The Elo rating system, originally designed for rating chess players, has since become a popular way to estimate competitors' time-varying skills in many sports. Though the self-correcting Elo algorithm is simple and intuitive, it lacks a probabilistic justification which can make it hard to extend. In this paper, we present a simple connection between approximate Bayesian posterior mode estimation and Elo. We provide a novel justification of the approximations made by linking Elo to steady-state Kalman filtering. Our second key contribution is to observe that the derivation suggests a straightforward procedure for extending Elo. We use the procedure to derive versions of Elo incorporating margins of victory, correlated skills across different playing surfaces, and differing skills by tournament level in tennis. Combining all these extensions results in the most complete version of Elo presented for the sport yet. We evaluate the derived models on two seasons of men's professional tennis matches (2018 and 2019). The best-performing model was able to predict matches with higher accuracy than both Elo and Glicko (65.8% compared to 63.7% and 63.5%, respectively) and a higher mean log-likelihood (-0.615 compared to -0.632 and -0.633, respectively), demonstrating the proposed model's ability to improve predictions.

Keywords: Paired comparison modelling; Margin of victory; Tennis; Correlated skills

1 Introduction

The Elo rating system was originally developed by Arpad Elo for rating chess players (Elo, 1978). Elo is given by the following simple algorithm:

$$\text{new winner rating} = \text{old winner rating} + k \times (1 - p(\text{win})),$$
$$\text{new loser rating} = \text{old loser rating} - k \times (1 - p(\text{win})),$$

where the win probability $p(\text{win})$ is given by the difference in ratings, transformed by a logistic function to lie between 0 and 1, and k is a constant that has to be set; often it is set to 32.[1] Each competitor starts with a rating of 1500, after which ratings are updated recursively using the formulas given above.

The Elo rating system was first adopted by the United States Chess Federation in 1960 and is still used, with some modifications, by the International Chess Federation (FIDE).[2] Elo has also become popular in other sports: it has been used to predict games in soccer (Hvattum and Arntzen, 2010) and tennis (Kovalchik, 2016), rank Gaelic football teams (Mangan and Collins, 2016), and was the official rating system in sports such as sumo, croquet and draughts, among others, at the time of a recent review (Stefani, 2011). It has even found use in behavioural ecology to rank dominance among animal hierarchies (Neumann, Duboscq, Dubuc, Ginting, Irwan, Agil, Widdig, and Engelhardt, 2011). Elo's popularity extends beyond academia: for example, the popular website FiveThirtyEight uses Elo ratings as part of its forecasts of baseball (Boice, 2019) and American football (Silver, Boice, and Paine, 2019).

The popularity of the Elo system is likely due to a combination of factors. Firstly, it is simple and fast to compute. Secondly, it is often surprisingly

[1] The Elo rating system in chess also accounts for draws, but we do not consider this possibility here.

[2] `https://ratings.fide.com/calc.phtml?page=change`

accurate at predicting outcomes. In Hvattum and Arntzen (2010), Elo-based models outperform four other models when predicting matches from four English soccer divisions, and in Kovalchik (2016), Elo outperforms ten other tennis prediction models proposed in the literature. Finally, since ratings are updated after every contest, Elo can provide a very detailed account of changing skills over time, unlike other models which often require the specification of time windows during which skills are assumed to be constant.

Despite these advantages, the simplicity of Elo also has drawbacks. Unlike more complex systems such as Glicko (Glickman, 1999), Elo does not come with a model-based derivation. This becomes problematic when the Elo rating system does not fit a domain exactly. For example, while the categorical win / draw / loss classification in Elo may be a good fit for chess, many other sports provide additional information in the form of score differentials, or margins of victory. Many proposed systems thus adapt Elo in some way to take margins into account. Some modify k to be greater when the margin is larger, such as Hvattum and Arntzen (2010), who use the goal difference in soccer as the margin; others replace the win expectation with an expected margin, such as Carbone, Corke, and Moisiadis (2016), who use the goal difference in rugby. Without theoretical guidance, it is unclear which formulation is preferable. A recent paper (Kovalchik, 2020) compares four different margin formulations for tennis, finding that a joint additive model, combining the usual win update with an update based on the margin, performs well both when predicting match outcomes and when assessed in a simulation study. However, devising this estimator requires intuition and ingenuity, and one may wonder whether there is a more direct approach.

In addition to the margin of victory, some sports have further factors that are important for match prediction. In particular, we are motivated here by the problem of accounting for the effect of playing surfaces in tennis. The

four most important tennis tournaments, known as Grand Slams, are played on three different surfaces: clay, grass and hard courts. A player who is skilled on one surface is likely to be strong on the others, too, but players often exhibit preferences for one surface over another. For example, Rafael Nadal is particularly skilled on clay, having won the French Open, the most important clay court tournament, a record 12 times in his career. By contrast, he has been able to win Wimbledon, the most important grass court tournament, only twice. The importance of accounting for surface effects has been recognised in prior work on tennis prediction. For instance, Sipko and Knottenbelt (2015) use empirically estimated correlations between surface win percentages to weight covariates in machine learning models, and Ingram (2019) estimate player-specific surface preferences in a Bayesian hierarchical model, finding large differences between players. It is not clear how the Elo update could be modified to account for surface differences. Post-hoc corrections can be made by forming a weighted average of Elo ratings fitted to each surface separately and overall Elo ratings (proposed for example in Morris, Bialik, and Boice (2016)), but this requires the estimation of several models, and a single updating scheme may thus be preferable.

In addition to the playing surface, the format can also differ in tennis, with the four men's Grand Slam tournaments using a best-of-five set format, requiring the winner to win three sets, while other events use the best-of-three set format, requiring the winner to win only two sets. Since a larger number of points is played, best-of-five set matches should provide more information about players' relative skills, but it is not clear how this should be accounted for in Elo. One may also expect that, as with surface, some players perform better in one format rather than the other. For example, a player with great endurance may be able to outlast a player in a best-of-five set match who would otherwise be favoured in the shorter best-of-three set format.

Prior work on rating systems in sports has generally focused on developing Bayesian rating systems for paired comparisons which are able to model uncertainty in ratings over time, such as Glickman (1999), Fahrmeir and Tutz (1994), and Dangauthier, Herbrich, Minka, and Graepel (2008). In particular, Glickman (1999) recovers Elo as a special case of the Glicko rating system, giving it a Bayesian interpretation. These approaches have strong theoretical foundations but, like Elo, are not an ideal fit for all sports. Regarding the margin of victory, the approach presented by Fahrmeir and Tutz (1994) allows for ordered responses, but while these are well-suited to discrete margins like the goal difference in soccer, they are not ideal for handling continuous margins such as the difference in the percentage of points won on serve in tennis. None of the models have an obvious mechanism for incorporating player-specific surface and tournament effects.

In this paper, we tackle the general question of how to extend the Elo rating system in a principled way. We show how an updating rule almost identical to that used in Elo can be derived by considering a Taylor expansion of the log posterior distribution in a Bayesian paired comparison model. We show that this updating rule is equivalent to that used for updating the mean estimates in an extended Kalman filter, but does not update players' variance estimates. We justify this procedure by drawing a link to Kalman filtering theory, where such filters are well known and referred to as steady-state extended Kalman filters (Assimakis and Adam, 2014). They exploit the property that, assuming observations are equally spaced in time and that the number of observations at each time point are the same, variance estimates in (extended) Kalman filters converge to a constant value after a small number of updates. Crucially, the derived procedure does not rely on any particular choice of prior or likelihood, making it easy to extend to different modelling assumptions. We use it to derive three updating rules: one incorporating the margin of victory by

changing the likelihood, one allowing updates for correlated skills by changing the prior, and a third combining these two.

Our key contributions are as follows: firstly, our derivation of Elo gives it a new interpretation as a steady-state Kalman filter, providing it with theoretical justification. Secondly, we generalise the derivation to devise a comprehensive rating system for tennis, taking into account surface, margin of victory, and tournament effects. Previous work has presented an extension of Elo to incorporate the margin of victory (Kovalchik, 2020) in tennis and a weighting approach for surface-specific ratings (Morris et al., 2016), but, to the best of our knowledge, no previous work has combined the two or estimated player-specific tournament effects. Thirdly, we provide a comprehensive evaluation of the derived models, finding that they outperform both Elo and Glicko. We find that, surprisingly, Glicko does not outperform standard Elo, despite its ability to model changing uncertainty over time, suggesting that the steady-state approximation is effective for predicting tennis matches and that Glicko's assumption of varying uncertainty with calendar time may not be ideal for tennis. We note that while the steady-state approximation appears to perform well for tennis, the assumptions required for it to be reasonable may not be met in other sports.

In the rest of this paper, we start by reviewing the Elo rating system more formally, introducing some of the notation used. We follow this by a quick review of the Glicko rating system and then derive our proposed procedure, showing that it recovers the Elo update almost exactly when the likelihood considers only wins or losses. We use the procedure to derive the three extensions mentioned in the previous paragraph. We evalute the resulting models on two seasons of men's professional tennis matches, finding that the model incorporating surface, margins of victory and tournament effects performs best, outperforming Elo and Glicko in both accuracy and log-likelihood. We con-

clude by illustrating how the best model predicts a single match, the 2019 Wimbledon semi-final between Roger Federer and Rafael Nadal.

2 Methods

2.1 Review of Elo

We start with a review of the Elo rating system. Each competitor starts with a skill estimate $\hat{\theta}_i = 1500$.[3] Suppose competitor i with Elo rating $\hat{\theta}_i$ plays competitor j with Elo rating $\hat{\theta}_j$. Then the win probability for competitor i according to the Elo system is given by:

$$p(y = 1|\hat{\theta}_i, \hat{\theta}_j) = \frac{1}{1 + 10^{-(\hat{\theta}_i - \hat{\theta}_j)/400}}\,. \tag{1}$$

We note that this win probability is a rescaled inverse logit link function, and that it can be rewritten as:

$$p(y = 1|\hat{\theta}_i, \hat{\theta}_j) = \gamma(b(\hat{\theta}_i - \hat{\theta}_j))\,, \tag{2}$$

$$b = \log(10)/400\,, \tag{3}$$

where $\gamma(x) = \text{logit}^{-1}(x)$.

The second part of the Elo system is given by its update rule. If the winner's rating estimate is $\hat{\theta}_w$ and the loser's is $\hat{\theta}_\ell$, the update is given by:

$$\hat{\theta}'_w = \hat{\theta}_w + k \times (1 - p(y = 1|\hat{\theta}_w, \hat{\theta}_\ell))\,, \tag{4}$$

$$\hat{\theta}'_\ell = \hat{\theta}_\ell - k \times (1 - p(y = 1|\hat{\theta}_w, \hat{\theta}_\ell))\,, \tag{5}$$

where k is a hyperparameter that must be specified in advance. Equations 4 and 5 show the self-correcting character of the system: the winner's score

[3] We write $\hat{\theta}$ to distinguish the point estimates made by Elo from the parameter θ representing the competitor's unknown but true skill.

is corrected upwards by k times $1 - p(y = 1|\hat{\theta}_w, \hat{\theta}_\ell)$, and the loser's score is corrected downwards by the same amount. A surprising result will have a large residual, resulting in a larger change for the winner and loser. It is interesting to note that the sum $\hat{\theta}_w + \hat{\theta}_\ell$ is unchanged by the update, so the overall sum of ratings will remain constant in the system.

The constant k is the only hyperparameter in the system that has to be chosen. A value of $k = 32$ is often suggested but may not be ideal for every sport. A more data-driven estimate can be made by maximising the likelihood given by Equation 2 using numerical optimisation.

2.2 Review of Glicko

Unlike Elo, the Glicko rating system takes into account time-varying uncertainty in each player's rating (Glickman, 1999). It assumes that initial abilities at time $t = 1$ are drawn from a normal distribution with prior variance σ_0^2:

$$\theta_i^{(1)} \sim \mathcal{N}(1500, \sigma_0^2)\,. \tag{6}$$

Glicko then divides time into periods. In each period, players' abilities are assumed constant. Each player's likelihood in each period is approximated with a Gaussian distribution, leading to closed-form updates to their rating estimates.

Between periods, each player's variance estimate is increased, each rating following a random walk with variance ν^2:

$$\theta_i^{(t+1)}|\theta_i^{(t)} \sim \mathcal{N}(\theta_i^{(t)}, \nu^2)\,, \tag{7}$$

where the superscripts denote the time period. For example, $\theta_i^{(t+1)}$ is player i's rating at time $t + 1$.

Glicko requires two variables to be set: ν^2 and σ_0^2. In addition, a period length has to be chosen. A procedure to set ν^2 and σ_0^2 based on a discrepancy

measure between the observed and predicted outcomes is given in Glickman (1999). For the period length, Glickman (1999) suggests that it should be set as short as possible so that the estimates can adjust quickly to time-varying skills, while maintaining a sample size in each period that would justify the approximations used (perhaps 5-10 matches).

2.3 Connection between posterior mode estimation and Elo

We now derive Elo from a Bayesian perspective. We start by considering a single match between two competitors. Without loss of generality (at least in the sports we are considering where draws are not an option), we consider the updates from the perspective of the winner. We start by assuming a normal prior on the winner and loser's skills θ_w and θ_ℓ:

$$\theta_w \sim \mathcal{N}(\mu_w, \sigma^2); \theta_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2)\,. \tag{8}$$

Here, we have already made a first simplifying assumption in setting the prior variance for both players to be equal (we will justify this assumption shortly). We define the sum and difference of prior ratings:

$$\tau = \theta_w + \theta_\ell; \delta = \theta_w - \theta_\ell\,. \tag{9}$$

It is a standard result that the sum and difference of independent univariate normal random variables with the same variances are independent and given by:

$$\tau \sim \mathcal{N}(\mu_w + \mu_\ell, 2\sigma^2) = \mathcal{N}(\mu_\tau, \sigma_\tau^2)\,, \tag{10}$$

$$\delta \sim \mathcal{N}(\mu_w - \mu_\ell, 2\sigma^2) = \mathcal{N}(\mu_\delta, \sigma_\delta^2)\,. \tag{11}$$

So far, we have considered the prior ratings, $p(\theta_w, \theta_\ell)$. As a likelihood, we use the same likelihood used in the Elo rating system, that is:

$$p(y = 1|\theta_w, \theta_\ell) = \gamma(b(\theta_w - \theta_\ell)) = \gamma(b\delta)\,. \tag{12}$$

The reader may have already observed that Equation 12 is equivalent to a rescaled version of the likelihood used in a Bradley-Terry model (Bradley and Terry, 1952). While the models discussed in this paper share the same likelihood, they differ in that they allow ratings to vary over time. For this reason, they are sometimes referred to as dynamic Bradley-Terry models.

Note that the likelihood depends only on δ. Since δ and τ are independent, the data provide information only on δ. Once the update is made and we obtain δ', we can recover the updated θ'_w and θ'_ℓ using $\theta'_w = (\tau + \delta')/2$ and $\theta'_\ell = (\tau - \delta')/2$.

By Bayes' rule, the posterior distribution is proportional to:

$$p(\delta|y = 1) \propto p(\delta)p(y = 1|\delta) = \mathcal{N}(\delta|\mu_\delta, \sigma^2_\delta)\gamma(b\delta)\,. \tag{13}$$

This posterior is intractable since the normal prior is not conjugate to the likelihood. We thus consider approximate methods, and in particular attempt to find the posterior mode. The derivation approach closely parallels that in Glickman (1999), but is simplified somewhat by considering matches one at a time rather than in periods, and because we will only approximate the posterior mean, rather than the posterior variance. We note that other approaches have been developed for addressing this inference problem. In particular, TrueSkill attempts to match posterior moments (Dangauthier et al., 2008), iteratively refining the estimates using the expectation propagation algorithm Minka (2001), and Weng and Lin (2011) find approximate closed-form solutions to the posterior moments.

The log posterior, up to constants that do not depend on δ, is given by:

$$\log p(\delta|y=1) \propto -\frac{1}{2}\frac{(\delta-\mu_\delta)^2}{\sigma_\delta^2} + \log\gamma(b\delta) = t(\delta)\,.$$

At the mode, its derivative $t'(\delta)$ is zero:

$$t'(\delta) = -\frac{\delta-\mu_\delta}{\sigma_\delta^2} + b(1-\gamma(b\delta)) = 0\,.$$

This equation could be easily solved by numerical methods but instead we opt to find an approximate solution using a Taylor expansion around the prior mean μ_δ. To first order, this expansion is:

$$t'(x) \approx t'(\mu_\delta) + t''(\mu_\delta)(x-\mu_\delta)\,. \tag{14}$$

Setting this to zero, the approximate maximum x will be at:

$$x = \mu_\delta + \frac{t'(\mu_\delta)}{-t''(\mu_\delta)}\,. \tag{15}$$

Straightforward algebra (see Appendix A for details) yields:

$$x = \mu_\delta + 2k(1-\gamma(b\mu_\delta))\,, \tag{16}$$

where

$$\boxed{k = \frac{b/2}{\frac{1}{\sigma_\delta^2} + b^2\gamma(b\mu_\delta)(1-\gamma(b\mu_\delta))}} \tag{17}$$

We call $x = \mu'_\delta$ and use it to approximate the posterior mean. Since $\theta'_w = (\tau+\delta')/2$ and $\theta'_\ell = (\tau-\delta')/2$, by linearity of expectation this means that the approximate posterior means μ'_w and μ'_ℓ for the winner and loser will be given by:

$$\boxed{\mu'_w = \frac{1}{2}(\mu_\tau + \mu'_\delta) = \mu_w + k(1-\gamma(b\mu_\delta))} \tag{18}$$

$$\boxed{\mu'_\ell = \frac{1}{2}(\mu_\tau - \mu'_\delta) = \mu_\ell - k(1-\gamma(b\mu_\delta))} \tag{19}$$

We note that Equations 18 and 19 are almost identical to the Elo update. The only difference is that now, k is a function of μ_δ, the prior difference in

skill. It is also almost identical to the k-factor obtained in the Glicko rating system under certain simplifying assumptions (see Appendix D for more information).

The above calculations suggest a procedure for the construction of Elo-type updates:

1. Write the log posterior density as a function of the difference in skill, δ.
2. Find the first and second derivatives of that function with respect to δ.
3. Use these to calculate the update using the Taylor expansion given in Equation 15.

These updates are performed for each match. Only the means of the skill estimates are updated; the variances are kept at σ^2, so that the next update proceeds in the same way but using the new mean estimates.

From a Bayesian point of view, it may seem strange not to update players' variance estimates, which should be reduced to reflect the knowledge gained from the data. Results from Kalman filtering can provide a justification. The first step is to realise that the proposed update is equivalent to a single Newton-Raphson step towards the posterior mode. As shown in (Humpherys, Redd, and West, 2012), there is a close relationship between the Newton-Raphson method and the update in the extended Kalman filter (EKF). We show in Appendix H that the multivariate version of the proposed update (presented in Section 2.5) in fact gives the same mean estimate as a particular EKF. The EKF additionally estimates the posterior covariance using a linear approximation (see (Särkkä, 2013, p.67), for example). In a typical EKF, variance would then be added between matches, before the next update reduces it again.

Under a set of assumptions, it is well known in the theory of Kalman filtering that variance estimates quickly settle into a "steady-state", remaining approximately constant (Assimakis and Adam, 2014). These assumptions are: (1) the observations are made over equally spaced time points, and (2) the

number of observations made over each time point is the same. The proposed approach updates after each match, so the second assumption is met. The first assumption holds if changes in competitor skill either do not vary as a function of calendar time, or if they do, matches are equally spaced in time.

Matches in tennis are not equally spaced in time, but most players play frequently enough for the variation in time between matches to be relatively small, so the assumption may be reasonable. Further, it is common to start fitting Elo models many years prior to the time of interest so, at the time of interest, most players will have been observed many times, making it likely that their variance estimates will have settled into a steady state. We find later on that the steady-state approach appears to lead to good predictions in tennis. We note however that its assumptions may be less appropriate for other sports if there is reason to believe that calendar time plays a large role and thus treating matches as equally spaced in time would be problematic.

Intuitively, if the assumptions are met and the variance estimates settle into a steady state, the reduction in uncertainty from learning the match outcome is cancelled out by the uncertainty introduced between matches, resulting in the overall variance remaining constant. The variance σ^2 in the proposed update above can thus be thought of as approximating the steady-state variance in an EKF. Such filters are well-known in the literature on Kalman filtering, where they are known as "steady-state Kalman filters" or "constant-gain Kalman filters". They are used particularly when computations to update the posterior covariance would be too costly (see for example Banfield, Ingersoll, and Keppenne (1996)) and can sometimes perform almost as well as full extended Kalman filters (Wilson, 1972).

The reasoning above provides a justification of the constant-variance assumption. One may also be interested in the accuracy of the other two approximations used in the derivation above: linearisation and using the approximate

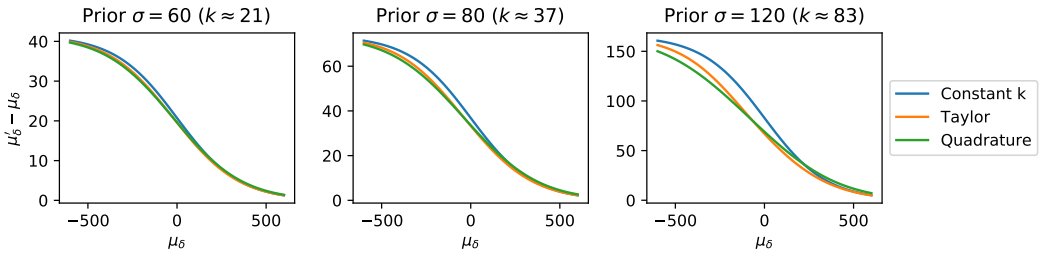

Figure 1: The figure shows the difference between the posterior mean and the prior mean, $\mu'_\delta - \mu_\delta$ as a function of the prior mean difference in ratings μ_δ, for three different methods. Two of these are approximations ("Constant k" and "Taylor") and the third should be close to exact ("Quadrature"). The three panels correspond to three different settings of the prior standard deviation σ for each player, and the three lines are coloured by the method used. Please note that while the x-axis is shared across the panels, the values of the y-axes differ. For each panel, the equivalent k is computed from Equation 17 as $k = b\sigma^2$.

mode as the mean estimate. To evaluate this, we compare three approaches to estimating the posterior mean in Figure 1. The first, which we call "Constant k", uses the constant k obtained by setting the second term in the denominator of Equation 17 to zero. This is equivalent to fitting Elo whose k-factor is set to this value. The second, dubbed "Taylor", uses the full k in Equation 17, including the second term in the denominator. Finally, to evaluate the accuracy of these approximations, we use Gauss-Hermite quadrature with 15 quadrature points to calculate the posterior mean (see Appendix F for details), which should be almost exact.

Before discussing the quality of the approximations, we first use the figure to gain intuition about the update. All three curves have a similar decaying shape, the change in the mean estimate going to zero as μ_δ grows. This is intuitively reasonable since wins with a large prior rating difference are unsurprising and thus should not result in large changes in the skill estimates. As μ_δ becomes smaller, wins become more surprising, and updates correspondingly

become larger.

Turning our attention to the quality of the three approximations, the left panel suggests that for a small prior standard deviation of $\sigma = 60$, corresponding to setting k to about $k = b\sigma^2 \approx 21$ in Elo, the three methods are almost equivalent. At $\sigma = 80$ or $k \approx 37$, the constant k approximation tends to make slightly larger updates than would be correct, and the Taylor series and quadrature solutions are still very similar. Finally, for a very large player prior standard deviation of $\sigma = 120$ or $k \approx 83$, the constant k approximation overestimates the update more strongly, and the Taylor series approach generally also slightly overestimates the posterior mean but is considerably more accurate. Overall, the figure suggests that for the k-factors around 30 typically used, both a constant k and the derived approach should be a good approximation to the posterior mean. For large prior standard deviations, the constant-k approximation starts to differ noticeably from the exact posterior mean. The derived approach still achieves good accuracy while being less computationally intensive than quadrature.

2.4 Extension to margin of victory

We now use the procedure derived in the previous section to derive an Elo-type update for tennis which takes the margin of victory into account. In tennis, the margin does not completely determine the outcome. For example, if a game spread were used as a margin, a winning score of 7-6 0-6 7-6 would have a negative margin of -4. It thus makes sense to consider the joint distribution of win, y, and margin, s. We write this joint distribution as:

$$p(y = 1, s|\delta) = p(y = 1|\delta)p(s|y = 1, \delta)\,. \tag{20}$$

The first term is the same as previously. For the second, we assume:

$$p(s|y = 1, \delta) = \mathcal{N}(s|\mu = c_1\delta + c_2, \sigma^2_{obs})\,. \tag{21}$$

The underlying idea is that the expected margin can be obtained from the skill difference δ by multiplying δ with a constant c_1 and adding an offset c_2, and that the distribution of margins follows a normal distribution. The offset's role is to take into account the fact that that the winner's margin is more likely to be positive than negative. Finally, σ^2_{obs} plays the role of noise in the observations of the margin. This model thus introduces three additional parameters: c_1, c_2, and σ^2_{obs}. We discuss how to set these later on in Section 2.7.

Equation 21 only determines the conditional distribution of the margin for the winning player, that is, the distribution of the margin given that $y = 1$. We derive the distribution given $y = 0$ from a consistency argument. Let δ_w be the difference in ratings between the winner and loser. Then $\delta_\ell = -\delta_w$ is the difference from the loser's perspective. Further define the random variables $[s_w|y = 1, \delta = \delta_w]$, the winner's margin, and $[s_\ell|y = 0, \delta = \delta_\ell]$, the loser's margin, where square brackets have been added for notational clarity. Since the sum of winner's and loser's margins equals zero, we have that $s_w = -s_\ell$. Hence

$$p(s|y = 0, \delta = -\delta_w) = \mathcal{N}(s|\mu = -c_1\delta_w - c_2, \sigma^2_{obs}), \tag{22}$$

so

$$p(s|y = 0, \delta) = \mathcal{N}(s|\mu = c_1\delta - c_2, \sigma^2_{obs}), \tag{23}$$

an intuitive result which says that the knowledge that a player won increases the expected margin by c_2, while knowledge of a loss reduces it by the same amount.

We now apply the same ideas as in the previous section to derive the update incorporating the margin of victory. Since the steps are very similar, we omit the details here (they can be found in Appendix B) and state only the final

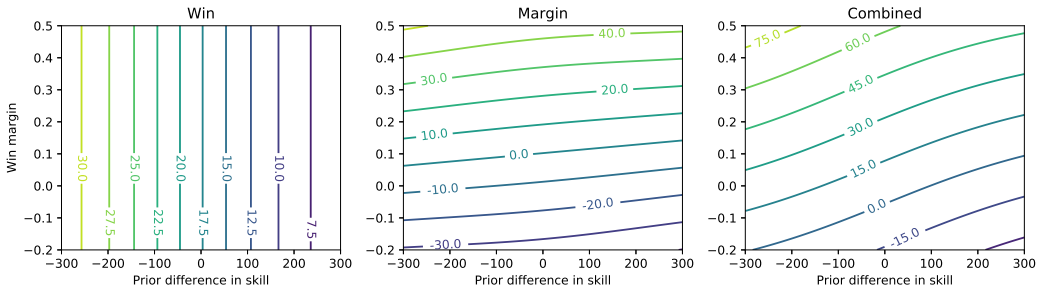

Figure 2: The figure shows the updates for the winner due to the outcome of the match (first panel from the left), the margin of victory (middle panel), and both summed (panel on the right), as a function of the prior difference in skill μ_δ and the margin. The contours show parameter settings for which the update sizes are constant.

result:

$$\mu'_w = \mu_w + k_{win}(1 - \gamma(b\mu_\delta)) + k_{margin}(s - s_{pred})\,, \tag{24}$$

$$\mu'_\ell = \mu_\ell - k_{win}(1 - \gamma(b\mu_\delta)) - k_{margin}(s - s_{pred})\,, \tag{25}$$

where

$$k_{margin} = \frac{c_1}{\sigma^2_{obs}} k_{shared};\; k_{win} = b k_{shared};\; s_{pred} = c_1\mu_\delta + c_2\,, \tag{26}$$

$$k_{shared} = \frac{1/2}{\frac{1}{\sigma^2_\delta} + b^2\gamma(b\mu_\delta)(1 - \gamma(b\mu_\delta)) + \frac{c_1^2}{\sigma^2_{obs}}}\,. \tag{27}$$

We see that the full update decomposes into two components: an update based on how well the win was predicted, given by $k_{win}(1 - \gamma(b\mu_\delta))$, and a new component, given by $k_{margin}(s - s_{pred})$, based on how well the margin was predicted. These contributions are added together, but both involve a shared k as given by Equation 27. The update sums margin and win updates just as the joint additive update proposed in Kovalchik (2020) but additionally accounts for their correlation by incorporating the winner's offset c_2.

In Figure 2, we show how the win and margin contributions combine to form an overall update for the winning player. The settings for the parameters

were approximately $c_1 = 0.00013$, $c_2 = 0.10$, $\sigma = 84$ and $\sigma_{obs} = 0.085$. For these settings, the expected margin for a win would be $c_2 = 0.10$ plus the difference in Elo points multiplied by c_1. The small magnitude of c_1 is due to this particular margin lying between -1 and 1, while Elo rating differences typically span hundreds of points.

The win update is independent of the margin, hence the contours are vertical. These updates are always positive, but decrease for high positive differences in skill, since the winning outcome is expected in that situation. The margin update depends slightly on the difference in prior skill through the shared component in the k factor, k_{shared}, but is mostly determined by the margin. Combining both contributions results in the panel on the right. Interestingly, a player with a prior difference in skill of 200 can lose points despite winning, if the margin was low enough. This seems reasonable since in this situation, a low margin would indicate underperformance compared to expectation.

Finally, we give a concrete example of the update. Consider a player with a mean rating of 1600 beating a player with a rating of 1500 with a margin of 0.2. Evaluating the k factors with the parameters listed previously gives $k_{shared} = 6185$, $k_{margin} = 111.3$, and $k_{win} = 35.6$. The predicted margin for the winner is $s_{pred} = 0.113$. The update from the win is thus $k_{win}(1 - \gamma(b\mu_\delta)) = 12.8$, and the update from the margin is $k_{margin}(0.2 - 0.113) = 9.7$, resulting in a gain of 22.5 points for the winner and a loss of the same amount for the loser.

2.5 Extension to correlated skills

As a second extension, we tackle the problem of correlated latent skills. As discussed in the introduction, this arises in tennis, where competitors' skills are expected to vary by the playing surface (such as clay, grass, or hard courts).

Players can require different skills to excel on these different surfaces, so a win on one surface may provide only limited information about their skill on other surfaces. A similar situation can be found in chess, where different formats can impose different time limits on each move. Here, too, abilities are likely to be correlated, but some players excel in shorter formats, while others do better when they are allowed more time to contemplate their moves.

To be able to model such correlated skills, we propose the following model. As before, we consider a single match update from the perspective of the winner. The prior ratings of players are given by:

$$\boldsymbol{\theta}_w \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}); \boldsymbol{\theta}_\ell \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}) \,.$$

Here, too, we make an equal (co-)variance assumption for the winner and loser. Now, however, the means $\boldsymbol{\mu}$ are (column) vectors of length n, and the covariance matrix $\boldsymbol{\Sigma}$ is of size $n \times n$. To make this discussion more concrete, the elements of $\boldsymbol{\theta}_w$ could be the ratings of tennis players for each of n surfaces, or the skills of chess players in n different formats.

We model $\boldsymbol{\theta}_w$ and $\boldsymbol{\theta}_\ell$ as independent, so the joint prior is given by concatenating them as follows:

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_w \\ \boldsymbol{\theta}_\ell \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_\ell \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix} \right) = \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \,.$$

We now introduce the vector $\boldsymbol{a}$, defined as:

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}_w \\ -\boldsymbol{a}_\ell \end{bmatrix},$$

so that

$$\delta = \boldsymbol{a}^\intercal \boldsymbol{\theta} \,.$$

In the surface example, $\boldsymbol{a}_w = \boldsymbol{a}_\ell$, a dummy vector with a 1 at the surface of interest and zero otherwise. We note that more generally, $\boldsymbol{a}_w$ need not be equal

to $\boldsymbol{a}_\ell$. This could be used to include variables which differ by competitor, such as indicators like "won last match" which could model momentum effects, or even continuous variables.

The log posterior for the case where we consider only win or loss is proportional to:

$$t(\boldsymbol{\theta}) \propto -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^\intercal \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \log \gamma(b\boldsymbol{a}^\intercal \boldsymbol{\theta})\,. \tag{28}$$

Considering a multivariate Taylor expansion leads to the following update, analogous to the 1D case in Equation 15:

$$\boxed{\boldsymbol{\mu}'_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{H}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}})\boldsymbol{j}(\boldsymbol{\mu}_{\boldsymbol{\theta}})} \tag{29}$$

where $\boldsymbol{H}$ is the Hessian and $\boldsymbol{j}$ is the Jacobian of $t(\boldsymbol{\theta})$, written as functions of $\boldsymbol{\theta}$. Omitting details of the calculation (which can be found in Appendix C.1), these are given by:

$$\boldsymbol{j}(\boldsymbol{\mu}_{\boldsymbol{\theta}}) = (1 - \gamma(b\boldsymbol{a}^\intercal \boldsymbol{\mu}_\theta))b\boldsymbol{a}\,, \tag{30}$$

$$\boldsymbol{H}(\boldsymbol{\mu}_{\boldsymbol{\theta}}) = -\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-\mathbf{1}} - \gamma(b\boldsymbol{a}^\intercal \boldsymbol{\mu}_\theta)(1 - \gamma(b\boldsymbol{a}^\intercal \boldsymbol{\mu}_\theta))b^2\boldsymbol{a}\boldsymbol{a}^\intercal\,. \tag{31}$$

Once this update is calculated, the resulting vector $\boldsymbol{\mu}'_{\boldsymbol{\theta}}$ is split in half to obtain the updated ratings for each player. Equation 29 is the new procedure which generalises the one given in Equation 15 and reduces to it in the case of one-dimensional skills.

Comparing the multivariate update to the univariate Elo update, we see that $\boldsymbol{j}(\boldsymbol{\mu}_{\boldsymbol{\theta}})$ is proportional to the residual of the win prediction, and $-\boldsymbol{H}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}})$ plays a role analogous to k. This model has $n(n+1)/2$ parameters which need to be estimated: the elements of the covariance matrix $\boldsymbol{\Sigma}$. We discuss how to set these in the next section.

We illustrate the update in Figure 3. In the example, we construct a two-dimensional covariance matrix $\boldsymbol{\Sigma}$ to have marginal variances of 100^2 and 90^2

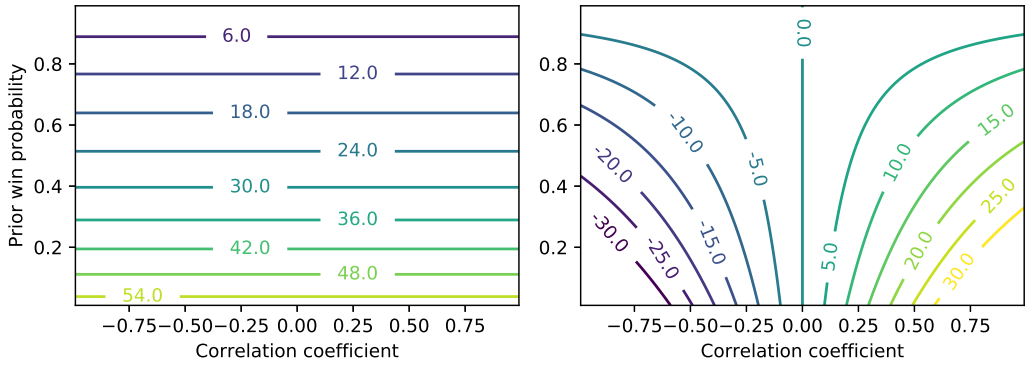

Figure 3: The figure shows the update size for the winner for a direct observation of one latent skill (left panel) and a correlated skill (right panel). Each contour shows settings for which the updates are constant.

and consider a range of correlations between the two skills. We consider an update where $\boldsymbol{a_w} = \boldsymbol{a_\ell} = [1, 0]^\intercal$, i.e. where the first skill is observed directly. Unsurprisingly, the update for the skill that is directly observed has no dependence on the correlation coefficient but only on how surprising the win was. By contrast, the update for the second skill, shown in the right panel, depends strongly on the correlation coefficient.

The derivation is easily extended to the margin of victory case. This adds a third term to Equation 28 which gives rise to one additional term in both the Hessian and Jacobian (please see Appendix C.2 for details of the derivation). The result is given by:

$$\boldsymbol{j}(\boldsymbol{\mu_\theta}) = \boldsymbol{j}_{win}(\boldsymbol{\mu_\theta}) + \frac{c_1}{\sigma^2_{obs}}(s - s_{pred})\boldsymbol{a}\,,$$

$$\boldsymbol{H}(\boldsymbol{\mu_\theta}) = \boldsymbol{H}_{win}(\boldsymbol{\mu_\theta}) - \frac{c_1^2}{\sigma^2_{obs}}\boldsymbol{a}\boldsymbol{a}^\intercal\,,$$

$$\text{where } s_{pred} = c_1\boldsymbol{a}^\intercal\boldsymbol{\mu_\theta} + c_2\,.$$

and $\boldsymbol{j}_{win}$ and $\boldsymbol{H}_{win}$ are given by Equations 30 and 31, respectively. After the Hessian and Jacobian are calculated, Equation 29 can be used to compute the

update taking the margin into account.

2.6 Accounting for differences in format

Men's tennis matches are played in either the best-of-three set format or the best-of-five set format. As mentioned in the introduction, the format can be expected to have several consequences. Firstly, it is likely to affect the win probability, since the stronger player is more likely to win the longer format (Kovalchik and Ingram, 2018). We account for this by replacing the original likelihood for the win from Equation 12 with

$$p(y = 1|\delta, I_{bo5}, m) = \gamma(b(1 + I_{bo5} \times m)\delta)\,, \tag{32}$$

where I_{bo5} is an indicator that is 1 if the match was played in best-of-five format and 0 otherwise, and m is a non-negative parameter that has to be estimated. For best-of-three set matches, this likelihood reduces to the previous likelihood, but for best-of-five set matches, the difference in skills δ is multiplied by $1 + m$, resulting in a larger win probability for the player with the higher rating.

Secondly, since, on average, a greater number of points will be played in a best-of-five set match, the standard deviation of the observation error in the margin is likely to decrease also. To take this into account, we estimate a separate observation variance σ^2_{bo5} for best-of-five set matches. We do not expect c_1 and c_2 to be very different in a best-of-five set match, hence these parameters are shared across both formats.

Finally, some players may perform better in one format than the other. There may be several reasons for this. For one, the best-of-five set format may be more physically demanding due to its length, which could favour some players over others. In addition, best-of-five set matches are now only played at the Grand Slams, which are the most prestigious of tournaments, offering

the most prize money, ranking points, and attention from fans. Some players may conserve energy at other tournaments and save their best performances for these tournaments. Tournaments outside the Grand Slams are further split into ATP Tour 250, ATP Tour 500 and ATP Tour Masters 1000 series tournaments[4], again differing in the number of ranking points and prize money offered, but sharing the best-of-three format.

To account for these factors, we add two terms to the skills vector $\boldsymbol{\theta_i}$ for each player, corresponding to additive terms for the ATP Tour Masters 1000 level and Grand Slam level tournaments relative to their performance at the less prestigious ATP Tour 250 and 500 series tournaments, which we group together. Because Grand Slam tournaments are the only ones played in the best-of-five set format, the final additive term models both each player's skill in the longer format as well as any effect due to the importance of the tournaments. A player's rating at a tournament is then determined by the sum of their skill on the surface the tournament is played on and the skill addition for this tournament.

The updates in the previous section are easily modified to account for these changes. The pre-factor b is replaced by $b(1 + m)$ in the case of a best-of-five set match, and σ^2_{obs} is replaced by σ^2_{bo5}. The skills vector $\boldsymbol{\theta_i}$ for each player is extended by the two additive tournament terms, and the dummy vectors $\boldsymbol{a_w}$ and $\boldsymbol{a_l}$ now pick out both surface and tournament skills. For example, if there are two surfaces and the match is played on the first at an ATP Masters 1000 tournament, $\boldsymbol{a_w} = [1, 0, 1, 0]^\intercal$. We model these tournament additions as independent from both each other and the surface skills to reduce the number of parameters, adding two terms to the the diagonal of the matrix $\boldsymbol{\Sigma}$ and zeros everywhere else.

[4]See, for example, `https://en.wikipedia.org/wiki/2020_ATP_Tour` for details.

2.7 Setting parameters

The models presented above have a number of parameters that have to be set. We propose to find point estimates for them by maximising their log marginal likelihood. In the most complex case presented here, the log marginal likelihood for a given match, conditional on the parameters $\boldsymbol{\alpha} = \left\{c_1, c_2, \sigma^2_{obs}, \boldsymbol{\Sigma}, m, \sigma^2_{bo5}\right\}$ and the covariates $\boldsymbol{x} = \{\boldsymbol{a_w}, \boldsymbol{a_\ell}, I_{bo5}\}$ is given by:

$$\log p(y = 1, s) = \log p(y = 1) + \log p(s|y = 1) =$$
$$\log\left[\int p(y = 1|\delta)p(\delta)d\delta\right] + \log\left[\int p(s|y = 1, \delta)p(\delta)d\delta\right],$$

where all distributions are conditioned on $\boldsymbol{\alpha}$ as well as the covariates $\boldsymbol{x}$, which we omit to keep the notation simple. A simple approximation would ignore the uncertainty in δ and instead plug in its mean μ_δ to calculate this likelihood. We show in Appendix E however that the integral is approximately equal to:

$$\log p(y = 1, s) \approx \log \gamma(b\mu_\delta/\alpha) + \log \mathcal{N}(s|c_1\mu_\delta + c_2, \sigma^2_{obs} + c_1^2\sigma^2_\delta)\,. \qquad (33)$$

which is as efficient to compute as the plug-in approximation but should be more accurate.

The full log marginal likelihood decomposes into a sum over each match. The simplest approach to maximising it is to use numerical optimisation, using finite differences to approximate its gradient with respect to the parameters. This is a valid approach but finite differences become inefficient as the dimensionality of $\boldsymbol{\alpha}$ grows. A more efficient alternative is to calculate gradients using automatic differentiation, which we implement using the software package `JAX` (Bradbury, Frostig, Hawkins, Johnson, Leary, Maclaurin, and Wanderman-Milne, 2018) in the Python programming language. `JAX` is also able to calculate the Hessians and Jacobians used in the updates automatically, further simplifying the implementation. The resulting optimisation is very fast, taking only a matter of minutes to estimate the parameters from

over 20,000 matches on a standard laptop machine running a 3.1GHz Intel Core i5 with 16GB of memory. Our code to fit the models used in this paper is available here: `https://github.com/martiningram/jax_elo`.

2.8 Dataset

We evaluate the models proposed in this paper on a dataset of professional men's tennis matches spanning the seasons from 2010 until 2019, obtained using the OnCourt dataset (`https://www.oncourt.info/`). Model parameters are estimated on the years 2010-2017, and the 2018 and 2019 seasons are used as the validation set. We drop matches played on carpet since the surface is rarely played on in recent years and discard retirements and walkovers. We note that retirements and walkovers likely do contain information, so discarding them is not ideal. However, keeping them is problematic since matches ending in retirements can result in large rating increases for the winning player, when in fact the result is due to the loser's injury rather than a strong performance by the winner. The developed approach could be extended, for example by modifying the likelihood for matches ending in retirements, but here we discard the matches for simplicity.

Table 1 shows the fractions of matches played on the different surfaces. Hard courts are the most common surface, making up about 40% of the matches in the dataset. Clay courts are the second most common surface played on, followed by indoor hard courts, and finally grass courts are rarest. The fractions of matches played on the different surfaces are similar for the training and validation sets.

We use the difference in the percentage of points won on serve as the margin of victory since it had the highest predictive accuracy and lowest log-loss among five margins compared in Kovalchik (2020) when used as the margin of victory. As shown in Table 1, this margin is 11.3% on average in the

	Train	Validation
Number of matches	20,163	5,099
Number of players	693	342
% Hard	39.9	40.9
% Clay	31.2	30.5
% Indoor hard	16.9	16.2
% Grass	12.0	12.4
Winner % serve won	69.3	69.5
Loser % serve won	58.0	59.0
Winner % margin	11.3	10.6

Table 1: The number of matches in the training and validation datasets, the fraction on each surface, as well as the average percentage of points won on serve for the winner and loser, together with the winner's margin on this measure.

training set, and 10.6% in the validation set, suggesting that the matches in the 2018 and 2019 seasons were on average slightly less lopsided than those played between 2010 and 2017.

2.9 Evaluation metrics and models

We use two metrics to evaluate the models: mean log-likelihood and accuracy. Accuracy, the fraction of times each model picked the correct winner, is an intuitive metric, but does not take into account how well calibrated the estimates were. The mean match log-likelihood is given by computing the log probability of each win and averaging this quantity across matches. This metric is equivalent to using the logarithmic scoring rule to evalute predictions, which is commonly used to evaluate probabilistic forecasts due to a number of desirable theoretical properties (see Gneiting and Raftery (2007) for a detailed discussion). Unlike accuracy, this scoring rule harshly penalises confident incorrect predictions.

We evaluate the following models:

1. *Elo* We fit standard Elo as presented in Section 2.1 with a constant k-factor estimated on the training set. We predict each match in the validation set using the pre-match ratings and Equation 2.
2. *Elo (independent surface)* To investigate how modelling correlations between surface skills compares to treating them as independent, we fit Elo to each surface separately, finding the optimal k for each surface and predicting using Equation 2.
3. *Glicko* We fit Glicko using three different period lengths: 30 days, 7 days, and 1 day. We experimented with longer periods, but found them to perform worse on the evaluation set. For each version, we minimize the discrepancy measure given in Glickman (1999) to find optimal values of σ_0^2 and v^2. As is the case for all models in this evaluation, the

parameters are estimated using the entire training set, whose statistics are given in Table 1. Our code used to fit Glicko is available online at `https://github.com/martiningram/glicko-python`.

4. *GenElo* We fit the model derived in Section 2.3, estimating the prior standard deviation σ^2 from the training set. We refer to the derived approach as "GenElo", short for "generalised Elo". We evaluate two versions of this model: one approximating the marginal likelihood by plugging in the mean μ_δ of the predicted difference, and one using the approximate marginal likelihood in Equation 33. Although the latter approach should be more accurate, we include the version using point estimates because it should be almost identical to Elo, allowing us to investigate how similar they are in practice. For all other models, we use the approximate marginal likelihood in Equation 33 to fit the models and predict outcomes.
5. *Glicko with constant variance* We fit Glicko with a constant variance by fixing players' variance estimates σ_0^2 to a constant, updating only their means. We update this version of Glicko one match at a time, like Elo, and minimise the discrepancy measure given in Glickman (1999) on the training set to find the optimal value of σ_0^2.
6. *GenElo margin* We fit the margin of victory model derived in Section 2.4, estimating σ_{obs}^2, c_1 and c_2 and prior uncertainty σ^2.
7. *GenElo surface* We fit a surface-specific model using the results from Section 2.5, estimating the prior covariance $\mathbf{\Sigma}$ across surfaces from the training set.
8. *GenElo surface + margin* We fit a surface-specific model which also accounts for the margin, again using the results from Section 2.5, estimating $\mathbf{\Sigma}$, c_1, c_2 and σ_{obs}^2.

Period length (days)	σ_0	ν
30	153.2	15.5
7	171.6	8.3
1	171.7	3.6

Table 2: Estimates made by Glicko for the prior player standard deviation σ_0 and the period-to-period standard deviation ν for three different period lengths.

9. *GenElo surface + margin + tournament* Finally, we fit the model accounting for surface and tournament effects described in Section 2.6, estimating $\mathbf{\Sigma}$, c_1, c_2, σ^2_{obs}, σ^2_{bo5}, and m.

3 Results

3.1 Parameter estimates

Elo's optimal k was found to be 32.5, which is close to the commonly-used value of 32. GenElo using point estimates for the skill distributions, rather than the approximation in Equation 33, yielded a prior standard deviation for the players of 78.3, implying a prior standard deviation of their difference of 110.8. For an evenly-matched contest where $\gamma(b\mu_\delta) = 0.5$, Equation 17 shows that this is equivalent to a k of 32.1, which is very similar to Elo, providing further evidence that these models are almost identical. The version using Equation 33 to calculate the log marginal likelihood fitted a prior player standard deviation of 84.4, or 119.4 for the difference. For an evenly-matched contest, this is equivalent to a slightly larger k of 36.7.

Glicko's estimates for the prior player standard deviation σ_0 and period-to-period standard deviation ν are shown in Table 2. The prior standard de-

viation is smallest for the model with 30-day periods at 153.2, and larger at 171.6 and 171.7 for the models with 7-day and 1-day periods, respectively. The period-to-period standard deviation decreases from 15.5 for the 30-day period model to 3.6 for the model with single-day periods. The prior standard deviations are around twice as large as the estimates for GenElo. This large difference is not unexpected as GenElo estimates the steady-state variance, which is typically smaller than the prior variance. Minimising the discrepancy measure in the constant variance version of Glicko yielded an optimal player standard deviation of 81.7, which is slightly smaller but close to the value estimated by GenElo using Equation 33.

The margin model fit a prior standard deviation of 83.4, which is similar to the model using only wins. The standard deviation of the margin, σ_{obs}, was estimated to be 0.085, and $c_1 = 0.000131$ and $c_2 = 0.102$. This suggests that a match between two evenly-matched players can be expected to end with a margin of $c_2 = 0.102 \pm 0.085$ for the winner, while a difference of 100 Elo points would produce a slightly larger expected margin of $c_1 \times 100 + c_2 = 0.013 + 0.102 = 0.115$.

The parameter estimates for the marginal standard deviations σ for the multivariate model with margin of victory and tournament effects are given in Table 3. The standard deviations vary somewhat across the surfaces, with grass having the highest, followed by clay, and finally indoor hard and hard courts. Updates on grass courts will thus be largest, followed by those on clay courts, and those on hard and indoor hard courts will be smallest. These standard deviations are generally somewhat larger than the standard deviations for the models ignoring surface. The additive effect for the ATP Tour Masters 1000 series tournaments has a standard deviation $\sigma_{masters}$ of approximately zero, indicating that players' abilities are not expected to differ when playing these tournaments compared to ATP Tour 250 and ATP Tour 500 se-

σ_{clay}	σ_{grass}	σ_{hard}	σ_{indoor_hard}	$\sigma_{masters}$	σ_{slam}
90.6	95.5	82.2	86.5	0.0	23.7

Table 3: Marginal standard deviations for the different skills in the model incorporating margin, surface and tournament effects.

ries tournaments. The additive Grand Slam effect on the other hand is large, at $\sigma_{slam} = 23.7$. This indicates that some players are estimated to perform noticeably better or worse at Grand Slams than at lower-level tournaments.

Table 4 shows the estimated correlation coefficients between the different pairs of surfaces. All correlations are positive, meaning that a positive update on any surface will lead to positive updates on all others. The strength of the correlation varies across surfaces, with the correlations between indoor hard and hard courts, as well as grass and hard courts being greatest (0.86 and 0.82, respectively) and the correlation between clay and grass courts being smallest (0.41). These correlations suggest that clay and grass courts are the most dissimilar. Common tennis knowledge suggests that this is reasonable, given that these two surfaces are usually classified as the slowest (clay) and fastest (grass), favouring different styles of play.[5]

The estimates for c_1, c_2 and σ^2_{obs} for the multivariate margin model with tournament effects are $c_1 = 0.000144$, $c_2 = 0.0998$, $\sigma_{obs} = 0.087$ and are thus similar to the margin model without tournament and surface effects. The best-of-five factor $m = 0.432$, indicating that skill differences are exaggerated by a factor of 1.432 in best-of-five set matches. Finally, as expected due to the larger number of points played, the observation standard deviation for the margin in best-of-five set matches, σ_{bo5}, is somewhat smaller than that for best-of-three set matches, at 0.071.

[5] See `https://en.wikipedia.org/wiki/Tennis_court`, for example

ρ_{cg}	ρ_{ch}	ρ_{ci}	ρ_{gh}	ρ_{gi}	ρ_{ih}
0.41	0.72	0.65	0.82	0.75	0.86

Table 4: Correlation coefficients of the surface ratings in the model including both margin and surface. We abbreviate the surfaces by their first letter; for example, ρ_{cg} is the correlation between clay and grass.

3.2 Evaluation results

Table 5 shows the summary metrics on the validation set. The mean log-likelihood ranges from -0.642 for Elo fit independently to each surface to -0.615 for the surface-specific model with margin of victory and tournament adjustments. Accuracy increases almost monotonically with the log-likelihood, from 63.4% to 65.8%.

Elo and GenElo with point estimates have essentially the same validation set performance. This is unsurprising given the close correspondence between the k-factor in Equation 17 and Elo. As expected, for the prior standard deviation performing best in tennis, the effect of the term involving the win probability in Equation 17 appears to be negligible.

Fitting and predicting using the marginal likelihood in Equation 33 appears to lead to a slight improvement for both accuracy and log-likelihood. Taking the playing surface into account leads to a larger improvement. By contrast, Elo fit independently to each surface performs considerably worse than GenElo with correlated surface effects, suggesting that modelling correlations is worthwhile. Adding the margin of victory also improves predictions, and combining surface and margin effects improves upon each separate model with an accuracy of 65.6% and a log-likelihood of -0.618 on the validation set. Finally, adding the adjustments to format and tournament level gives another slight improvement, yielding the best model with a mean log-likelihood of

-0.615 and an accuracy of 65.8%.

Surprisingly, Glicko does not outperform its constant-variance version and Elo, despite its ability to model changes in uncertainty over time. Among the three period lengths considered, the single-day period version has a slightly higher log-likelihood than the other two. Compared to constant-variance Glicko and Elo, the log-likelihood is slightly lower, at -0.633 compared to -0.631 and -0.630, respectively. It thus appears that, surprisingly, modelling changing variance with calendar time, rather than treating all matches as if they were equally spaced in time, does not improve predictions on this tennis dataset. We suggest two possible reasons for this, leaving a detailed investigation to future work. The first is that in tennis – unlike, for example, in chess – a player's skill can rapidly change from one day to the next as a result of an injury. Therefore, proximity in calendar time may not always imply similarity in skill. A second reason may be that lower-rated players can be absent from the main tour for months, playing lower-level events. When they return to the main tour, their skill is unlikely to have changed dramatically since their last match: if it were substantially lower, they would have likely failed to qualify, and if it were substantially higher, they would have likely qualified earlier. As a result, despite their long absence as measured in calendar time, their skill is unlikely to have changed dramatically.

Some of the improvements in Table 5 appear small and one may wonder whether they are due to noise. To investigate this, we first designate Elo as the reference model. For each model, we then compute the difference in log-likelihood for each match, yielding a vector d of differences. The interest lies in the mean of differences μ_d for each model compared to Elo. Approximating the likelihood with a normal distribution with mean μ_d and variance σ^2, a standard Bayesian analysis with a non-informative prior on the prior mean

Model	Mean log-likelihood	Accuracy
Elo (independent surface)	-0.642	0.634
Glicko, 30 day periods	-0.636	0.636
Glicko, 7 day periods	-0.635	0.635
Glicko, 1 day periods	-0.633	0.636
Elo	-0.632	0.637
GenElo (Point estimates)	-0.632	0.636
Glicko, constant variance	-0.631	0.636
GenElo	-0.630	0.637
GenElo Surface	-0.626	0.647
GenElo Margin	-0.623	0.650
GenElo Surface + Margin	-0.618	0.656
GenElo Surface + Margin + Tournament	-0.615	0.658

Table 5: Evaluation results on the validation set. The rows are ordered by the mean log-likelihood per match, from lowest to highest.

	2.5%	Median	97.5%
Elo (independent surface) - Elo	-0.0183	-0.0103	-0.0022
Glicko, 30 day periods - Elo	-0.0072	-0.0038	-0.0004
Glicko, 7 day periods - Elo	-0.0055	-0.0027	0.0001
Glicko, 1 day periods - Elo	-0.0038	-0.0011	0.0016
GenElo (Point estimates) - Elo	0.0001	0.0002	0.0003
Glicko, constant variance - Elo	0.0004	0.0007	0.0011
GenElo - Elo	0.0009	0.0019	0.0029
GenElo Surface - Elo	0.0028	0.0063	0.0098
GenElo Margin - Elo	0.0058	0.0095	0.0133
GenElo Surface + Margin - Elo	0.0089	0.0142	0.0195
GenElo Surface + Margin + Tournament - Elo	0.0113	0.0168	0.0223

Table 6: Credible intervals for the posterior mean μ_d of the improvement in the per-match log-likelihood compared to Elo.

and a uniform prior on $\log \sigma^2$ yields a posterior distribution for μ_d of

$$\mu_d | d \sim t_{n-1}(\bar{d}, s^2/n)\,, \tag{34}$$

where $\bar{d}$ is the sample mean and s^2 is the sample variance of the differences, and t_{n-1} denotes a Student's t-distribution with $n-1$ degrees of freedom (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, 2013, p. 66).

Table 6 shows the posterior summaries obtained using this analysis. GenElo with point estimates and Glicko with different period lengths show similar performance as Elo, and Elo fit independently to each surface performs somewhat worse. The 95% credible intervals of the other models do not include zero, indicating that they are most likely able to improve on the baseline model.

	Clay	Grass	Hard	Indoor Hard	Slam+
Rafael Nadal	2392	1991	2170	2134	51
Roger Federer	2063	2230	2214	2241	17

Table 7: Ratings prior to the match between Nadal and Federer at Wimbledon 2019 estimated by the model including surface, margin and tournament effects. Please note that the ATP Masters 1000 series addition is not shown since it is almost exactly zero for both players.

3.3 Match prediction example

We now illustrate the best-performing model by predicting the 2019 Wimbledon semi-final between Roger Federer and Rafael Nadal. Federer and Nadal are among the most successful tennis players of all time, having won 20 and 19 Grand Slam titles respectively, more than any other male player. In this particular meeting, Nadal was generally thought to be the slight favorite. For example, bet365, a bookmaker, offered decimal odds of 1.72 for Nadal and 2.10 for Federer[6], translating to a win probability of about 55% for Nadal. However, Federer defeated Nadal in four sets, 7-6 1-6 6-3 6-4. Federer won 69.8% of service points compared to Nadal's 64.6%, giving him a margin of 0.052.

Table 7 shows the ratings predicted by the best-performing model prior to the match. The ratings are strikingly different across surfaces. On clay courts, Nadal was rated over 300 points ahead of Federer; on grass however, the surface Wimbledon is played on, Federer had an edge of around 240 points, indicating that he was favoured by the model. Both Nadal and Federer have a positive addition to their ratings at Grand Slams (shown as "Slam+" in the

[6]These odds were retrieved from `https://www.flashscore.com/match/lhqi1P31/#match-summary`.

table), Nadal's being somewhat larger at 51 compared to Federer's 17. In fact, Nadal had the highest estimated Grand Slam addition among all players at the end of the 2019 season (we refer the interested reader to Appendix G for a table of players with the highest and lowest Grand Slam additions).

To predict the match outcome, we compute the mean and variance of the difference in ratings. Each player's mean rating is given by summing their grass court rating and slam addition, yielding 2247 points for Federer and 2042 for Nadal. The mean difference in ratings is thus 205 from Federer's point of view. Each player's variance is computed by summing $\sigma^2_{grass} = 95.5^2$ and $\sigma^2_{slam} = 23.7^2$ as given in Table 3. This results in a prior standard deviation of 98.4 for each player, or 139.2 for the difference in ratings.

To predict the win probability, since the match is played in the best-of-five format, this rating difference is multiplied by $1 + m = 1.432$. This results in a predicted difference of 294 points with a standard deviation of 199. Using the approximation in Appendix E, the win probability is given by $\gamma(b\mu_\delta/\alpha)$ where $\alpha = \sqrt{1 + \pi\sigma^2_\delta b^2/8}$. This results in a predicted win probability of 79.8% for Federer, which is substantially higher than the 45% predicted by the betting markets. Similarly, Glicko with single-day periods predicted a prior mean of 2109 for Federer and 2156 for Nadal with standard deviations of 64.8 and 68.2, respectively. This results in a predicted win probability of 43.6%, closely in line with the betting odds. This suggests that betting markets may have underestimated Federer and Nadal's differences in ability across surfaces.

The expected margin for Federer can be calculated by the law of total expectation as:

$$\begin{aligned}
\mathbb{E}[s|\delta] &= \mathbb{E}[\mathbb{E}[s|\delta, y]] \\
&= \mathbb{E}[\mathbb{E}[y \times [s|\delta, y = 1] + (1 - y) \times [s|\delta, y = 0]]] \\
&= \mathbb{E}[p_{win} \times [s|\delta, y = 1] + (1 - p_{win}) \times [s|\delta, y = 0]] \\
&= p_{win}(c_1\delta + c_2) + (1 - p_{win})(c_1\delta - c_2)\,,
\end{aligned}$$

where $p_{win} = 0.798$, the predicted win probability, and c_1 and c_2 were given in the previous section. Taking the expectation over δ results in a predicted margin of 0.089, which is slightly larger than the one observed, indicating that Federer was expected to win in slightly more dominant fashion.

4 Discussion and Conclusions

In the previous sections, we have derived Elo as an approximate Bayesian model and provided it with theoretical justification by linking it to steady-state extended Kalman filtering. A generalisation of the result allowed us to derive a dynamic rating system for tennis which, for the first time, models player-specific surface and tournament effects and takes into account the margin of victory. Elo has previously been shown to be one of the most accurate prediction models in tennis (Kovalchik, 2016), and since the extensions outperform Elo, the models derived in this paper are likely to be among the best forecasting models for the sport proposed in the literature.

Unlike other probabilistic rating systems such as (Glickman, 1999) or the approach presented in Fahrmeir and Tutz (1994), the derived approach does not model changes in the uncertainty around competitors' skills over time, instead keeping the variance constant after each update. This has the advantage

of simplifying the algorithm, making it easy to extend, and reducing computational complexity, but one may expect it to come at a cost of predictive accuracy. Surprisingly, this does not appear to be the case in tennis, with even standard Elo matching Glicko's predictive accuracy. A partial explanation for this may be that the steady-state approximation is a good one in tennis, since most players are observed many times at the time of prediction and would thus have an approximately constant variance in a full model. It is still somewhat surprising that accuracy is essentially equivalent and, if anything, slightly improved, compared to Glicko. We noted two possible ways in which Glicko's assumptions may be problematic in tennis, and future work could investigate these in depth.

Although our application in this paper was tennis, the derived procedure is general, and it could be applied to other sports. In soccer, for example, the likelihood for the goal difference could be modelled using a Skellam distribution, as proposed in (Karlis and Ntzoufras, 2008). Alternatively, the likelihood could take the form of an ordered probit, as used for example in Hvattum and Arntzen (2010). Another possible extension could be to estimate individual player skills in team sports, which could be done by using the extension presented in Section 2.5, forming the skill vector by concatenating individual players' ratings and modifying the vector $\boldsymbol{a}$ to compute the difference in the sum of individual team players' skills.

Overall, in the case of tennis, our results suggest that incorporating more domain-specific information such as surface, the margin of victory, and tournament effects, leads to larger improvements in predictive accuracy than relaxing the steady-state variance approximation. Although this is an intuitive result to followers of tennis who are well aware that factors like tournament level and surface are important, the dynamic rating systems presented in the literature do not allow such effects to be easily incorporated. We believe that

our work fills this gap, providing a straightforward procedure to derive rating systems that are easy to extend, accurate, and computationally efficient, and we hope that it will allow sports statisticians to devise ratings appropriate for their sport more easily in the future.

A Details of the Taylor series expansion

We already noted that the update is given by:

$$x = \mu_\delta + \frac{t'(\mu_\delta)}{-t''(\mu_\delta)}\,, \tag{35}$$

and that

$$t'(\delta) = -\frac{\delta - \mu_\delta}{\sigma_\delta^2} + b(1 - \gamma(b\delta))\,. \tag{36}$$

The second derivative $t''(\delta)$ is given by:

$$t''(\delta) = -\frac{1}{\sigma_\delta^2} - b^2\gamma(b\delta)(1 - \gamma(b\delta))\,.$$

And hence the update is:

$$\begin{aligned} x &= \mu_\delta + \frac{b(1 - \gamma(\mu_\delta b))}{\frac{1}{\sigma_\delta^2} + b^2\gamma(\mu_\delta b)(1 - \gamma(\mu_\delta b))} \\ &= \mu_\delta + 2k(1 - \gamma(b\mu_\delta))\,. \end{aligned}$$

B Derivation of margin of victory update

The likelihood and prior are given, respectively, by:

$$p(\delta) = \mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)\,, \tag{37}$$

$$p(y = 1, s|\delta) = p(y = 1|\delta)p(s|y = 1, \delta) \tag{38}$$

$$= \gamma(b\delta)\mathcal{N}(s|c_1\delta + c_2, \sigma_{obs}^2)\,. \tag{39}$$

The log posterior is thus proportional to:

$$t(\delta) \propto \log \gamma(b\delta) - \frac{1}{2}\left(\frac{\delta - \mu_\delta}{\sigma_\delta^2}\right)^2 - \frac{1}{2}\left(\frac{s - (c_1\delta + c_2)}{\sigma_{obs}^2}\right)^2 . \tag{40}$$

Hence

$$t'(\delta) = b(1 - \gamma(b\delta)) - \left(\frac{\delta - \mu_\delta}{\sigma_\delta^2}\right) + c_1\left(\frac{s - (c_1\delta + c_2)}{\sigma_{obs}^2}\right), \tag{41}$$

and

$$t''(\delta) = -b^2\gamma(b\delta)(1 - \gamma(b\delta)) - \frac{1}{\sigma_\delta^2} - \frac{c_1^2}{\sigma_{obs}^2} . \tag{42}$$

Evaluating these at the prior mean μ_δ leads to the update:

$$x = \mu_\delta + \frac{b(1 - \gamma(b\mu_\delta) + \frac{c_1}{\sigma_{obs}^2}(s - (c_1\mu_\delta + c_2))}{b^2\gamma(b\mu_\delta)(1 - \gamma(b\mu_\delta)) - \frac{1}{\sigma_\delta^2} - \frac{c_1^2}{\sigma_{obs}^2}} \tag{43}$$

$$= \mu_\delta + 2k_{shared}(b(1 - \gamma(b\mu_\delta))) + 2k_{shared}\frac{c_1}{\sigma_{obs}^2}(s - s_{pred}) , \tag{44}$$

with k_{shared} and s_{pred} defined as in the main text.

C Details of correlated skills derivation

C.1 Win/loss only

As stated in the main text, the log posterior is proportional to:

$$t_{win}(\boldsymbol{\theta}) \propto -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu_\theta})^\intercal \boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \log \gamma(b\boldsymbol{a}^\intercal\boldsymbol{\theta}) . \tag{45}$$

Hence the Jacobian and Hessian functions are given, respectively, by:

$$\boldsymbol{j_{win}}(\boldsymbol{\theta}) = -\boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu_\theta}) + (1 - \gamma(b\boldsymbol{a}^\intercal\boldsymbol{\theta}))b\boldsymbol{a} , \tag{46}$$

$$\boldsymbol{H_{win}}(\boldsymbol{\theta}) = -\boldsymbol{\Sigma}_\theta^{-1} - \gamma(b\boldsymbol{a}^\intercal\boldsymbol{\theta})(1 - \gamma(b\boldsymbol{a}^\intercal\boldsymbol{\theta}))b^2\boldsymbol{a}\boldsymbol{a}^\intercal . \tag{47}$$

Evaluating these at the prior mean $\boldsymbol{\mu_\theta}$ yields:

$$\boldsymbol{j_{win}}(\boldsymbol{\mu_\theta}) = (1 - \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu_\theta}))b\boldsymbol{a}\,, \tag{48}$$

$$\boldsymbol{H_{win}}(\boldsymbol{\mu_\theta}) = -\boldsymbol{\Sigma}_\theta^{-1} - \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu_\theta})(1 - \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu_\theta}))b^2\boldsymbol{a}\boldsymbol{a}^\top\,, \tag{49}$$

as stated in the main text.

C.2 Including margin of victory

The log posterior including the margin likelihood is proportional to:

$$t_{margin}(\boldsymbol{\theta}) \propto -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu_\theta})^\top\boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\theta) + \log\gamma(b\boldsymbol{a}^\top\boldsymbol{\theta}) - \frac{1}{2}\left(\frac{s - (c_1\boldsymbol{a}^\top\boldsymbol{\theta} + c_2)}{\sigma^2_{obs}}\right)^2 \tag{50}$$

$$= t_{win}(\boldsymbol{\theta}) - \frac{1}{2}\left(\frac{s - (c_1\boldsymbol{a}^\top\boldsymbol{\theta} + c_2)}{\sigma^2_{obs}}\right)^2\,, \tag{51}$$

with t_{win} defined as in Equation 45.

Hence the Jacobian and Hessian functions are given by:

$$\boldsymbol{j_{margin}}(\boldsymbol{\theta}) = \boldsymbol{j_{win}}(\boldsymbol{\theta}) + \frac{c_1}{\sigma^2_{obs}}(s - (c_1\boldsymbol{a}^\top\boldsymbol{\theta} + c_2))\boldsymbol{a}\,, \tag{52}$$

$$\boldsymbol{H_{margin}}(\boldsymbol{\theta}) = \boldsymbol{H_{win}}(\boldsymbol{\theta}) - \frac{c_1^2}{\sigma^2_{obs}}\boldsymbol{a}\boldsymbol{a}^\top\,. \tag{53}$$

Evaluating both at the prior mean $\boldsymbol{\mu_\theta}$ produces the equations stated in the main text.

D Comparison between derived update and Glicko

In (Glickman, 1999), Elo is derived as a special case of a more general approximate Bayesian rating system. Here, we compare the equivalent Elo update derived there with the one in this paper.

Equation 13 in (Glickman, 1999) recovers Elo as a special case of Glicko under the assumption that the opponents' skills are known exactly. Considering only a single contest in a period and matching the notation with that used in this paper, the equivalent k becomes:

$$k = \frac{b}{1/\delta^2 + 1/\sigma^2}, \tag{54}$$

$$\text{where } \delta^2 = [b^2\gamma(b\mu_\delta)(1 - \gamma(b\mu_\delta))]^{-1}. \tag{55}$$

Since $\sigma^2 = \sigma_\delta^2/2$, this k-factor is:

$$k = \frac{b/2}{1/\sigma_\delta^2 + (b^2/2)\gamma(b\mu_\delta)(1 - \gamma(b\mu_\delta))}. \tag{56}$$

This k-factor is almost exactly the same as the one given in Equation 17, with the exception of the factor of 1/2 in the second term in the denominator. This factor may be due to the assumption that opponents' skills are known exactly in the derivation.

E Derivation of approximate log marginal likelihood

The log marginal likelihood for a given match is:

$$\log p(y = 1, s) = \log p(y = 1) + \log p(s|y = 1) = \tag{57}$$

$$\log\left[\int p(y = 1|\delta)p(\delta)d\delta\right] + \log\left[\int p(s|y = 1, \delta)p(\delta)d\delta\right] = \tag{58}$$

$$\log\left[\int \gamma(b\delta)\mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)d\delta\right] +$$

$$\log\left[\int \mathcal{N}(s|c_1\delta + c_2, \sigma_{obs}^2)\mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)d\delta\right]. \tag{59}$$

The second integral is analytically tractable and reduces to the log density of s given a normal distribution with mean $c_1\mu_\delta + c_2$ and variance $\sigma_{obs}^2 + c_1^2\sigma_\delta^2$.

The first integral is not analytically tractable but can be approximated well by $\gamma(b\mu_\delta/\alpha)$, where $\alpha = \sqrt{1 + \pi\sigma_\delta^2 b^2/8}$ (Crooks, 2009). The log marginal likelihood can thus be approximated by:

$$\log p(y = 1, s) \approx \log \gamma(b\mu_\delta/\alpha) + \log \mathcal{N}(s|c_1\mu_\delta + c_2, \sigma_{obs}^2 + c_1^2\sigma_\delta^2)\,, \tag{60}$$

as stated in the main text.

F Quadrature estimate of the posterior mean

As stated in Equation 13, the posterior distribution is proportional to:

$$p(\delta|y = 1) \propto p(\delta)p(y = 1|\delta) = \mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)\gamma(b\delta)\,. \tag{61}$$

Hence,

$$p(\delta|y = 1) = c \times \mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)\gamma(b\delta)\,, \tag{62}$$

where c is an unknown normalising constant.

We seek the posterior mean of δ, that is:

$$\mathbb{E}[\delta|y = 1] = \int \delta p(\delta|y = 1)d\delta = c \int \delta \mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)\gamma(b\delta)d\delta\,. \tag{63}$$

We first compute c using Gauss-Hermite quadrature:

$$\frac{1}{c} = \int \mathcal{N}(\delta|\mu_\delta, \sigma_\delta^2)\gamma(b\delta)d\delta\,. \tag{64}$$

We then evaluate the integral in Equation 63 using Gauss-Hermite quadrature a second time and obtain the final result by multiplying its value by c.

G Estimated Grand Slam additions

Table 8 shows the players with the largest and smallest estimated Grand Slam additions, as estimated at the end of the 2019 ATP season. Rafael Nadal has

	Slam+
Rafael Nadal	51.73
Fernando Verdasco	33.48
Novak Djokovic	33.19
Stan Wawrinka	32.76
Teimuraz Gabashvili	24.74
Dominic Thiem	22.64
John Millman	20.04
Diego Schwartzman	19.10
Kevin Anderson	18.66
Andy Murray	17.88

(a) Players with the ten largest Grand Slam additions.

	Slam+
Federico Delbonis	-13.55
Potito Starace	-13.77
Florian Mayer	-14.02
Gilles Muller	-14.24
Steve Johnson	-14.56
Jack Sock	-15.88
Pablo Cuevas	-17.69
Alexander Zverev	-22.56
Carlos Berlocq	-23.05
Juan Monaco	-23.63

(b) Players with the ten smallest Grand Slam additions.

Table 8: Grand Slam additions estimated at the end of the 2019 season.

the largest estimated addition, adding around 52 rating points when playing at a Grand Slam. Stan Wawrinka also ranks highly which is consistent with his reputation as a "big match player"[7]. On the other hand, Juan Monaco, Carlos Berlocq and Alexander Zverev are all estimated to perform worse at Grand Slams than at other events.

[7]See, for example, `https://en.as.com/en/2016/09/10/other_sports/1473530143_112960.html`

H Equivalence of EKF and Newton-Raphson means

In this section, we show that the approach proposed in Section 2.5 leads to the same posterior mean estimate as an extended Kalman filter using a particular choice of measurement error covariance and observation function.

We first state the equations used to update the mean estimate in extended Kalman filtering. Writing the state estimate before the update step as $\mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$, they can be written as (Särkkä, 2013, p.70):

$$\begin{aligned}
\tilde{\boldsymbol{y}} &= \boldsymbol{z} - h(\boldsymbol{\mu}_\theta), \\
\boldsymbol{S} &= \boldsymbol{G}\boldsymbol{\Sigma}_\theta\boldsymbol{G}^\top + \boldsymbol{R}, \\
\boldsymbol{K} &= \boldsymbol{\Sigma}_\theta\boldsymbol{G}^\top\boldsymbol{S}^{-1}, \\
\boldsymbol{\mu}'_\theta &= \boldsymbol{\mu}_\theta + \boldsymbol{K}\tilde{\boldsymbol{y}},
\end{aligned}$$

where $\boldsymbol{z}$ is the vector of observations, $g(\boldsymbol{x})$ is a non-linear function from the state vector to the observation vector, $\boldsymbol{R}$ is the measurement error covariance, and $\boldsymbol{G}$ is the Jacobian matrix of g evaluated at $\boldsymbol{x} = \boldsymbol{\mu}_\theta$, the prior mean. Note that the equations above are typically written with subscripts, e.g. $\tilde{\boldsymbol{y}}_{\boldsymbol{k}}$, to denote the update at time step k, but we drop these since we consider only a single updating step.

To show the equivalence with our derived approach, we now apply these general equations to the win-only update. We set

$$g(\boldsymbol{x}) = \gamma(b\boldsymbol{a}^\top\boldsymbol{x}),$$

which implies

$$\boldsymbol{G} = \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu}_\theta)(1 - \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu}_\theta))b\boldsymbol{a}^\top = \eta b\boldsymbol{a}^\top,$$

where we have defined the scalar quantity $\eta = \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu}_\theta)(1 - \gamma(b\boldsymbol{a}^\top\boldsymbol{\mu}_\theta))$ to simplify notation in the following. Please note that here, to be consistent with

the Kalman filtering literature, the Jacobian is written as a row vector, while other Jacobians in this paper are written as column vectors.

The other equations thus become:

$$\boldsymbol{S} = \eta^2 b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} + \boldsymbol{R}$$

and

$$\boldsymbol{K} = \eta b \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} (\eta^2 b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} + \boldsymbol{R})^{-1},$$

implying the following update to the mean:

$$\boldsymbol{\mu}'_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \eta b \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} (\eta^2 b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} + \boldsymbol{R})^{-1} (1 - \gamma(b \boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})), \tag{65}$$

where z has been set to 1, reflecting the winning outcome.

The approach derived in this paper in Section 2.5 has

$$\begin{aligned} \boldsymbol{j}_{\boldsymbol{win}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}) &= (1 - \gamma(b \boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})) b \boldsymbol{a} \,, \\ \boldsymbol{H}_{\boldsymbol{win}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}) &= -\boldsymbol{\Sigma}_{\theta}^{-1} - \gamma(b \boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})(1 - \gamma(b \boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})) b^2 \boldsymbol{a} \boldsymbol{a}^\intercal \,, \end{aligned}$$

which are used in a single Newton-Raphson step:

$$\boldsymbol{\mu}'_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{H}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}) \boldsymbol{j}(\boldsymbol{\mu}_{\boldsymbol{\theta}}),$$

yielding

$$\boldsymbol{\mu}'_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \eta b^2 \boldsymbol{a} \boldsymbol{a}^\intercal)^{-1} (1 - \gamma(b \boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})) b \boldsymbol{a}. \tag{66}$$

We thus see that for the two updates in Equations 65 and 66 to be equivalent, we must have

$$(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \eta b^2 \boldsymbol{a} \boldsymbol{a}^\intercal)^{-1} \boldsymbol{a} = \eta \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} (\eta^2 b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} + \boldsymbol{R})^{-1}.$$

We rewrite the matrix inverse term on the left hand side using the Sherman-Morrison formula, yielding:

$$(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \eta b^2 \boldsymbol{a} \boldsymbol{a}^\intercal)^{-1} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\theta}} + \eta b^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}} (\boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a}) - \eta b^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\eta b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{a} + 1}.$$

Multiplying this expression by $\boldsymbol{a}$ and making use of the associativity of matrix multiplication, the second and third terms in the numerator cancel, leaving

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{a}(\eta b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{a} + 1)^{-1} = \eta \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{a}(\eta^2 b^2 \boldsymbol{a}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{a} + \eta)^{-1},$$

which is identical to the EKF update if

$$\eta = \boldsymbol{R} = \gamma(b\boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})(1 - \gamma(b\boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})),$$

which is the variance of a Bernoulli random variable with success probability $\gamma(b\boldsymbol{a}^\intercal \boldsymbol{\mu}_{\boldsymbol{\theta}})$ and thus is a reasonable choice for the measurement error variance.

References

Assimakis, N. and M. Adam (2014): "Iterative and algebraic algorithms for the computation of the steady state Kalman filter gain," *International Scholarly Research Notices*, 2014.

Banfield, D., A. P. Ingersoll, and C. L. Keppenne (1996): "A Steady-State Kalman Filter for Assimilating Data from a Single Polar Orbiting Satellite," *Journal of the Atmospheric Sciences*, 52, 737–753, URL https://doi.org/10.1175/1520-0469(1995)052<0737:ASSKFF>2.0.CO;2.

Boice, J. (2019): *How Our MLB Predictions Work*, URL https://fivethirtyeight.com/methodology/how-our-mlb-predictions-work/.

Bradbury, J., R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne (2018): "JAX: composable transformations of Python+NumPy programs," URL http://github.com/google/jax.

Bradley, R. A. and M. E. Terry (1952): "Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons," *Biometrika*, 39, 324–345, URL https://doi.org/10.1093/biomet/39.3-4.324.

Carbone, J., T. Corke, and F. Moisiadis (2016): "The Rugby League Prediction Model: Using an Elo-based approach to predict the outcome of National Rugby League (NRL) matches," *International Educational Scientific Research Journal*, 2, 26–30, URL https://iesrj.com/archive-sub?detail=THE_RUGBY_LEAGUE_PREDICTI.

Crooks, G. E. (2009): "Logistic approximation to the logistic-normal integral," *Technical Report Lawrence Berkeley National Laboratory*, URL https://threeplusone.com/pubs/on_logistic_normal.pdf.

Dangauthier, P., R. Herbrich, T. Minka, and T. Graepel (2008): "Trueskill through time: Revisiting the history of chess," in *Advances in Neural Information Processing Systems*, 337–344, URL https://papers.nips.cc/paper/3331-trueskill-through-time-revisiting-the-history-of-chess.

Elo, A. E. (1978): *The Rating of Chess Players, Past and Present*, Arco Pub.

Fahrmeir, L. and G. Tutz (1994): "Dynamic stochastic models for time-dependent ordered paired comparison systems," *Journal of the American Statistical Association*, 89, 1438–1449, URL https://www.jstor.org/stable/2291005.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013): *Bayesian Data Analysis*, CRC press, 3 edition, URL http://www.stat.columbia.edu/~gelman/book/.

Glickman, M. E. (1999): "Parameter Estimation in Large Dynamic Paired Comparison Experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48, 377–394, URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00159.

Gneiting, T. and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378, URL https://doi.org/10.1198/016214506000001437.

Humpherys, J., P. Redd, and J. West (2012): "A Fresh Look at the Kalman Filter," *SIAM Review*, 54, 801–823, URL https://doi.org/10.1137/100799666.

Hvattum, L. M. and H. Arntzen (2010): "Using ELO ratings for match result prediction in association football," *International Journal of Forecasting*, 26, 460–470, URL https://www.sciencedirect.com/science/article/abs/pii/S0169207009001708.

Ingram, M. (2019): "A point-based Bayesian hierarchical model to predict the outcome of tennis matches," *Journal of Quantitative Analysis in Sports*, 15, 313 – 325, URL https://www.degruyter.com/view/journals/jqas/15/4/article-p313.xml.

Karlis, D. and I. Ntzoufras (2008): "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference," *IMA Journal of Management Mathematics*, 20, 133–145, URL https://doi.org/10.1093/imaman/dpn026.

Kovalchik, S. (2020): "Extension of the Elo rating system to margin of victory," *International Journal of Forecasting*, URL http://www.sciencedirect.com/science/article/pii/S0169207020300157.

Kovalchik, S. A. (2016): "Searching for the GOAT of tennis win prediction," *Journal of Quantitative Analysis in Sports*, 12, 127–138, URL https://www.degruyter.com/view/journals/jqas/12/3/article-p127.xml.

Kovalchik, S. A. and M. Ingram (2018): "Estimating the duration of professional tennis matches for varying formats," *Journal of Quantitative Analysis in Sports*, 14, 13 – 23, URL https://www.degruyter.com/view/journals/jqas/14/1/article-p13.xml.

Mangan, S. and K. Collins (2016): "A rating system for Gaelic football teams: Factors that influence success," *International Journal of Computer Sci-*

ence in Sport, 15, 78–90, URL https://content.sciendo.com/view/journals/ijcss/15/2/article-p78.xml?language=en.

Minka, T. P. (2001): "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 362–369.

Morris, B., C. Bialik, and J. Boice (2016): *How We're Forecasting The 2016 U.S. Open*, URL https://fivethirtyeight.com/features/how-were-forecasting-the-2016-us-open/.

Neumann, C., J. Duboscq, C. Dubuc, A. Ginting, A. M. Irwan, M. Agil, A. Widdig, and A. Engelhardt (2011): "Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating," *Animal Behaviour*, 82, 911 – 921, URL http://www.sciencedirect.com/science/article/pii/S000334721100296X.

Silver, N., J. Boice, and N. Paine (2019): *How our NFL Predictions Work*, URL https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

Sipko, M. and W. Knottenbelt (2015): "Machine learning for the prediction of professional tennis matches," *MEng Computing Final Year Project, Imperial College London*, URL https://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf.

Stefani, R. (2011): "The Methodology of Officially Recognized International Sports Rating Systems," *Journal of Quantitative Analysis in Sports*, 7, URL https://www.degruyter.com/view/journals/jqas/7/4/article-1559-0410.1347.xml.xml.

Särkkä, S. (2013): *Bayesian Filtering and Smoothing*, Institute of Mathematical Statistics Textbooks, Cambridge University Press.

Weng, R. C. and C.-J. Lin (2011): "A Bayesian Approximation Method for

Online Ranking," *Journal of Machine Learning Research*, 12, 267–300, URL http://jmlr.org/papers/v12/weng11a.html.

Wilson, K. C. (1972): "An Optimal Control Approach to Designing Constant Gain Filters," *IEEE Transactions on Aerospace and Electronic Systems*, AES-8, 836–842.