Statistical Science
2017, Vol. 32, No. 4, 616–629
DOI: 10.1214/17-STS628

Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability?

David Aldous

Abstract. In a simple model for sports, the probability A beats B is a specified function of their difference in strength. One might think this would be a staple topic in Applied Probability textbooks (like the Galton–Watson branching process model, for instance) but it is curiously absent. Our first purpose is to point out that the model suggests a wide range of questions, suitable for "undergraduate research" via simulation but also challenging as professional research. Our second, more specific, purpose concerns Elo-type rating algorithms for tracking changing strengths. There has been little foundational research on their accuracy, despite a much-copied "30 matches suffice" claim, which our simulation study casts doubt upon.

Key words and phrases: Elo rating, Bradley–Terry model, dynamic ratings, sports forecasting.

1. INTRODUCTION

This article provides an overview of several topics from an applied probability viewpoint: more details of the mathematics can be found in Aldous (2017). A very useful recent account of the technical statistical side of the topic and further references to the statistical literature can be found in Király and Qian (2017).

1.1 The Basic Probability Model

Each team A has some "strength" x_A, a real number. When teams A and B play

$$\mathbb{P}(\text{A beats B}) = W(x_A - x_B)$$

for a specified "win-probability function" W satisfying the following conditions (which we regard as the minimal natural conditions):

$$\begin{aligned} &W : \mathbb{R} \to (0,1) \text{ is continuous, strictly increasing,} \\ &W(-x) + W(x) = 1; \quad \lim_{x\to\infty} W(x) = 1. \end{aligned} \tag{1}$$

Implicit in this setup:

- each game has a definite winner (no ties);
- no home field advantage, though this is easily incorporated by making the win probability be of the form $W(x_A - x_B \pm \Delta)$;
- strengths do not change with time.

A common choice for W is the *logistic* function

$$L(u) = e^u/(1+e^u), \quad -\infty < u < \infty.$$

We will take $W = L$ as a default, though we investigate the likely errors from using the wrong W in Section 2.6. With $W = L$, the model is (under the reparametrization $v = e^u$) equivalent to the *Bradley–Terry model* (Bradley and Terry, 1952), which has attracted a large literature in statistics by virtue of its "consensus ranking" interpretation. The basic statistics theory (MLEs, confidence intervals, hypothesis tests, goodness-of-fit tests) of that model is treated in Chapter 4 of David (1988). See Cattelan (2012) for a recent survey.

To use the model, we need to specify how matches are scheduled. The following three formats are representative of real-world sports and games:

- League format as in the English Premier League.
- Single-elimination tournament as in Wimbledon.
- No centralized scheduling, as in online games.

The first almost always involves *teams*, the third individual *players*, the second either: we use the words

David Aldous is Professor, Department of Statistics, 367 Evans Hall # 3860, University of California, Berkeley, Berkeley, California 94720, USA (e-mail: aldous@stat.berkeley.edu).

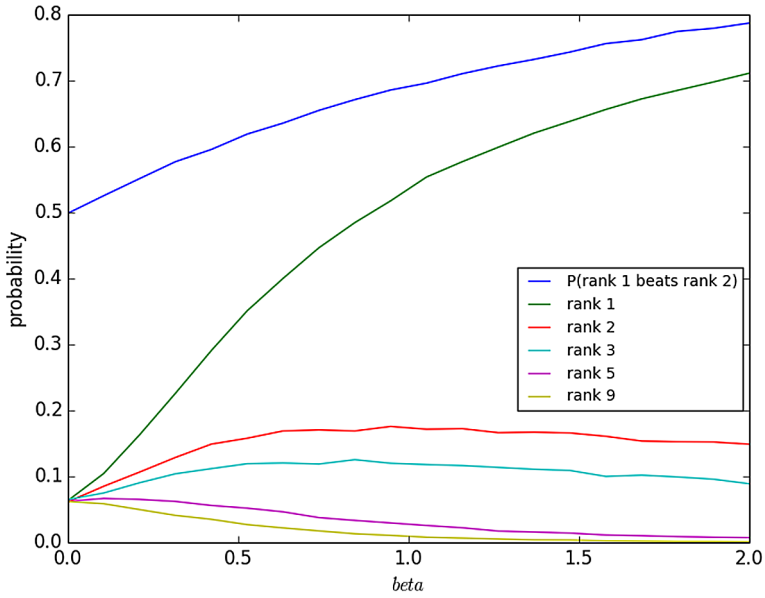

FIG. 1. *Probabilities of different-ranked players winning the tournament, compared with probability that rank-*1 *player beats rank-*2 *player* (*top curve*).

teams and *players* interchangeably in describing the scheduling models in this paper, which are usually simplified analogs of real-world scheduling.

Using $W = L$ involves a "standard unit" of strength difference, which can be interpreted as follows. Note $L'(0) = 1/4$. So the effect of a small increase δ in strength is that the probability of beating an originally equal strength team increases from 0.5 to about $0.5 + \delta/4$. See Section 2.2 for the transformation to "unit of strength" implied by conventional Elo ratings.

1.2 Existing Literature

We emphasize that our focus is not upon statistical analysis of data from a particular sport, which has a huge literature using more realistic modeling. Instead, analogous to use of the Galton–Watson model as a rough guide to the behavior of more realistic branching-style models, one can hope that the basic model provides a rough guide to the consequences of the stochastic nature of sports results. But there has been surprisingly little "applied probability" style mathematical treatment of the basic model. Indeed, the only textbook mention we know in that field concerns an algorithm (Lange, 2010, Example 3.4.1) for finding MLEs in the Bradley–Terry model. As recent examples of papers, Adler et al. (2016) gives upper and lower bounds for each player to win a randomly-matched tournament, in terms of the strengths x_i. And Chetrite, Diel and Lerasle (2017) considers the $n \to \infty$ limit probabilities that the best player wins a n-player league, under different models of random strengths. These papers cite scattered previous work, but the fact that these are *recent* papers on very basic questions underscores the lack of extensive literature.[1]

1.3 A Selection of Questions

The first objective of this article is to point out that there is a much broader variety of questions one could study within the basic model. We describe a few such questions here.

Robustness of second seed winning probability. A mathematically natural model for relative strengths of top players is $(\beta\xi_i, i \geq 1)$ where $0 < \beta < \infty$ is a scale parameter and

$$\xi_1 > \xi_2 > \xi_3 > \cdots$$

form the inhomogeneous Poisson point process on $\mathbb{R}$ of intensity e^{-x} arising in extreme value theory (Resnick, 1987). So we can simulate a tournament, in the conventional deterministic-over-seeds pattern schedule (Wikipedia: Tournament), to see which seed wins. Figure 1 gives data from simulations of such a tournament with 16 players, assuming the seeding (or *rank*) order coincides with the strength order. The interpretation of the parameter β is not so intuitive, but that parameter determines the probability that the top seed would beat the second seed in a single match, and

[1] Literature involving Elo ratings is discussed later.

TABLE 1
Seed of winner, men's and women's singles, Grand Slam tennis tournaments, 1968–2016

	Seed of winner					
	1	**2**	**3**	**4**	**5+**	**Total**
Frequency	148	94	42	29	77	390
Percentage	38%	24%	11%	7%	20%	100%
Model, $\beta = 0.65$	41%	17%	11%	7%	24%	100%

this probability is shown (as a function of β) in the top curve in Figure 1. The other curves show the probabilities that the 16-player tournament is won by the players seeded as 1, 2, 3, 5 or 9.

The results here are broadly in accord with intuition. For instance, it seems intuitively obvious that the probability that the top seed is the winner is monotone in β. What is perhaps surprising and noteworthy is that the probability that the second seed player is the winner is quite insensitive to parameter values, away from the extremes, at around 17%.

This kind of testable prediction is an appealing basis for an undergraduate research project. Is this 17% prediction in fact accurate? How robust is it to alternate models? As a start, data from tennis tournaments[2] in Table 1 shows a moderately good fit.

Annual changes in strength. In professional league-based team sports, changes in a team's winning percentages from year to year represent a combination of chance and actual changes in strength. For a specific team, it will usually be difficult to separate these effects, but for the league as a whole one can try to measure an average "change in strength" between successive years. It is not obvious how to formalize this notion in a way that can be estimated from the data of observed wins and losses, but our model suggests a way. For n teams with each pair playing twice per year, write N_i and N_i' for the number of wins of team i in successive years. A natural statistic to consider is

$$S = \frac{1}{n}\sum_{i=1}^{n}(N_i - N_i')^2.$$

If the strengths $(x_i, 1 \le i \le n)$ do not change, then

$$\mathbb{E}S = \frac{2}{n}\sum_{i=1}^{n} \operatorname{var} N_i$$

and because $\operatorname{var} N_i \le 2(n-1) \times \frac{1}{4}$ we obtain $\mathbb{E}S \le n - 1$. So observing a value of S significantly larger than $n - 1$ would confirm our common sense expectation that strengths overall do indeed change from year to year. But a more challenging question arises if we ask: how large would the change of strengths need to be, to make the observed changes in outcomes from last year to this year be statistically significant, for realistic size leagues? Figure 2 compares via simulation the distribution of S for unchanged strengths with the distribution S_1 where team strengths change, with typical (RMS) change[3] 0.4. Here, we took $n = 20$ teams and 2 games per year between each pair, and Normal(0, 1) distribution of team strengths. The considerable overlap in Figure 2 implies that in this scenario, the observed outcome changes in a given successive pair of years might well not be statistically significant. What would more realistic (incorporating draws) analysis of Premier League data show?

Promotion and relegation. In *promotion and relegation* schemes (Wikipedia: Promotion and relegation), each year a specified small number of bottom teams from a top division are exchanged with the top teams from a second division. This is of course intended to allow changes in strengths to be reflected in division placement, and such changes could be modeled as later in this article. But even with unchanging strengths there is a question about how well the resulting division placement reflects strength. For instance, for what k is the true strength of the kth best-performing team in the second division approximately equal to the true strength of the kth worst-performing team in the top division?

Large tournaments. How might one arrange a tournament to choose a winner out of (say) 200 players, given a constraint on the total number of matches to be played? Optimal schemes involve weaker players being eliminated progressively to allow more matches between the better players. One simple implementation (for the case of no prior information about strengths) would be to split into 10 leagues of 20 teams, do league play within each, then make a final league of 20 comprised of the top 2 teams from each original league. How does this compare with other schemes?

[2]Wimbledon, and the U.S., French and Australian Opens form the prestigious "Grand Slam." tournaments.

[3]Implying a win-probability previously 50% changes by around ±10%.

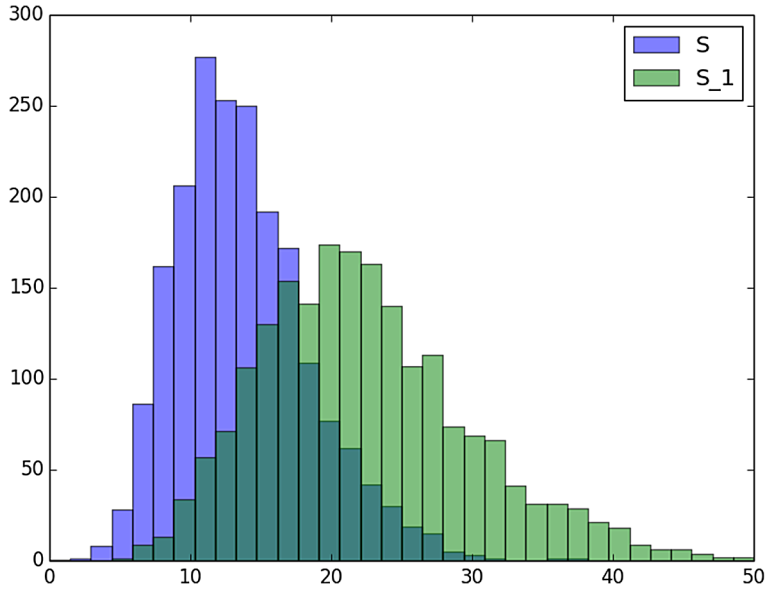

FIG. 2. *Distributions of the statistic S for unchanged and changed strengths.*

Elementary ranking schemes. The Langville and Meyer (2012) textbook discusses simple rating and ranking methods based on undergraduate linear algebra. How accurate are these, under the basic probability model?

2. ELO-TYPE RATING SYSTEMS

The particular type of rating systems we study are known loosely as Elo-type systems[4] and were first used systematically in chess. The Wikipedia page *Elo rating system* is quite informative about the history and practical implementation. What we describe here is an abstracted "mathematically basic" form of such systems.

Each player i is given some initial rating, a real number y_i. When player i plays player j, the ratings of both players are updated using a function Υ (`Upsilon`):

$$\begin{aligned}&\text{if } i \text{ beats } j, \text{ then } y_i \to y_i + \Upsilon(y_i - y_j) \text{ and}\\ &\quad y_j \to y_j - \Upsilon(y_i - y_j),\\ &\text{if } i \text{ loses to } j, \text{ then } y_i \to y_i - \Upsilon(y_j - y_i) \text{ and}\\ &\quad y_j \to y_j + \Upsilon(y_j - y_i).\end{aligned} \tag{2}$$

Note that the sum of all ratings remains constant. We require the function $\Upsilon(u)$, $-\infty < u < \infty$ to satisfy the qualitative conditions

$$\begin{aligned}&\Upsilon : \mathbb{R} \to (0, \infty) \text{ is continuous,}\\ &\quad \text{strictly decreasing,}\\ &\text{and} \quad \lim_{u\to\infty} \Upsilon(u) = 0.\end{aligned} \tag{3}$$

We will also impose a quantitative condition

$$\kappa_\Upsilon := \sup_u |\Upsilon'(u)| < 1. \tag{4}$$

To motivate the latter condition, we want the functions

$$x \to x + \Upsilon(x - y) \quad \text{and} \quad x \to x - \Upsilon(y - x),$$

that is the rating updates when a player with (variable) strength x plays a player of fixed strength y, to be an *increasing* function of the starting strength x.

Note that if Υ satisfies (3) then so does $c\Upsilon$ for any scaling factor $c > 0$. So given any Υ satisfying (3) with $\kappa_\Upsilon < \infty$ we can scale to make a function where (4) is satisfied. The complementary logistic function

$$L(-x) = \frac{1}{1+e^x} = 1 - \frac{e^x}{1+e^x}, \quad -\infty < x < \infty$$

is a common choice for the "update function shape" in Elo-type rating systems. That is, one commonly uses $\Upsilon(x) = cL(-x)$ for some scaling parameter c.

2.1 What Is the Connection Between Ratings and the Probability Model?

Elo-type rating algorithms have nothing to do with probability, *a priori*. But there is a simple *heuristic*

[4]Named after Arpad Elo; sometimes mistakenly written ELO as if an acronym.

connection between the probability model and the rating algorithm. This connection is "well known" in the sense of being implicit in much discussion of Elo ratings, but we have never seen a careful mathematical discussion, so we will attempt one here.

Consider n teams with unchanging strengths $x_1, \ldots, x_n$, with match results according to the basic probability model with some win probability function W, and ratings (y_i) given by the update rule with some update function Υ. When team i plays team j, the expectation of the rating change for i equals

$$\Upsilon(y_i - y_j)W(x_i - x_j) - \Upsilon(y_j - y_i)W(x_j - x_i). \tag{5}$$

So consider the case where the functions Υ and W are related by

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty. \tag{6}$$

In this case,

> (*)If it happens that the difference $y_i - y_j$ in ratings of two players playing a match equals the difference $x_i - x_j$ in strengths, then the expectation of the change in rating difference equals zero

whereas if unequal then (because Υ is decreasing) the expectation of $(y_i - y_j) - (x_i - x_j)$ is closer to zero after the match than before.

Call (6) the *balance relation*. These observations suggest that, under this balance relation, there will be a tendency for player i's rating y_i to move towards[5] its strength x_i though there will always be random fluctuations from individual matches. So if we believe the basic probability model for some given W, then in a rating system we should use an Υ that satisfies the balance relation.

Now we have a mathematical question: given W, what is the solution of the balance equation (6) for unknown Υ? Curiously, this does not have a good answer. The first observation is that $\Upsilon(u) = W(-u)$ is a solution, so we can use

$$\Upsilon(u) = \delta W(-u) \tag{7}$$

with any scaling factor $0 < \delta < 1$ we like. But there are more general solutions. For symmetric ϕ [i.e., $\phi(u) \equiv \phi(-u)$] with $\phi(0) = 1$,

$$\Upsilon(u) = \delta W(-u)\phi\big(|u|\big) \tag{8}$$

is a solution, provided the qualitative conditions (3, 4) remain satisfied. Other than simplicity, we know no explicit reason why choice (7) should be preferable to some other choice of form (8). One approach to this issue is to note that equation (6) arose from setting the expectation of the rating change to equal zero when ratings equal true strengths. The extra freedom of (8) allows us to impose a second requirement, that the *variance* of the rating change should be constant, and this leads to a specific "variance-stabilizing" form for the update function

$$\Upsilon(u) = \delta\sqrt{W(-u)/W(u)}. \tag{9}$$

Unfortunately, for logistic W the resulting Υ does not satisfy (4). But for the Cauchy distribution function the resulting Υ does indeed satisfy (4). This suggests a project: look at real-world Elo ratings based on the logistic, and see whether the alternative update function (9) arising from the Cauchy gives better predictions.

However, the mathematics are clearer if we consider the balance relation the other way round. Given an update function Υ there is a *unique* win-probability function W satisfying (6):

$$W_\Upsilon(u) := \Upsilon(-u)/\big(\Upsilon(u) + \Upsilon(-u)\big). \tag{10}$$

So when Elo ratings are derived from Υ we can regard this W_Υ as an associated implicit win-probability function. In particular, the conventional use of a scaled logistic $\Upsilon(u) = \delta L(-u)$ is implicitly using the logistic L as win-probability function.

2.2 Relating Our Mathematical Set-up to Published Ratings

As mentioned before, what we discuss in this article is the "mathematically basic" form of Elo ratings. In practice, the algorithm is adapted in different ways to different sports (see, e.g., Curiel, 2017 for international football) so numerical values in this article would need to be adjusted before attempting serious data analysis.

In published real-world data, ratings are integers, mostly in range 1000–2000. For instance, at time of writing the ratings for the England and Australian football teams (Curiel, 2017) are 1909 and 1701. The conventional implementation is that 1 standard unit (for logistic) in our model corresponds to 174 rating points.[6] So the implicit probabilities for an upcoming match would be

$$\mathbb{P}(\text{Australia beats England}) = L\left(\frac{1701 - 1909}{174}\right) = 0.23.$$

[5] More precisely, we center both strengths and ratings.

[6] 174 arises as $400/\log(10)$.

By convention, a new player is given a 1500 rating. If players never departed, the average rating would stay at 1500. However, players leaving (and no re-centering) will make the average tend to drift. One can define "expert" in a given sport by a threshold rating, but the drift makes it problematic to compare "expert" in different sports, or in the same sport over long time periods.

In published data, the update scaling factor δ (there called the *K-factor*) varies; for international football the factor depends on the significance of the match but $\delta = 0.2$ is typical. For tennis, United (2017) uses $\delta = 0.12$.

2.3 A Convergence Theorem

What can we do, in the setting above, via standard mathematical probability theory? Assume the basic probability model with nonchanging strengths, and use Elo-type ratings—what happens? We need to specify how the matches are scheduled, so let us use the mathematically simplest "random matching" scheme in which there are n players and for each match a pair of players is chosen uniformly at random. This gives a continuous-state Markov chain

$$\mathbf{Y}(t) = \big(Y_i(t), 1 \le i \le n\big), \quad t = 0, 1, 2, \ldots,$$

where $Y_i(t)$ is the rating of player i after a total of t matches have been played. Call this the *update process*. Note that this process is parametrized by the functions W and Υ, and by the vector $\mathbf{x} = (x_i, 1 \le i \le n)$ of player strengths. We center player strengths and rankings: $\sum_i x_i = 0$ and $\sum_i Y_i(0) = 0$.

The following convergence theorem is perhaps intuitively obvious; the main point is that no further technical assumptions are needed for W, Υ.

THEOREM 1. *Under our standing assumptions* (1), (3), (4) *on* W *and* Υ, *for each* $\mathbf{x}$ *the update process has a unique stationary distribution* $\mathbf{Y}(\infty)$, *and for any initial ratings* $\mathbf{y}(0)$ *we have* $\mathbf{Y}(t) \to \mathbf{Y}(\infty)$ *in distribution as* $t \to \infty$.

This can be proved (Aldous, 2017) by standard methods—coupling and Lyapounov functions. Note here we are not assuming the balance relation (6) between W and Υ. Note also that given nonrandom initial rankings $\mathbf{y}(0)$ the distribution of $\mathbf{Y}(t)$ has finite support for each t, so we cannot have convergence in variation distance, which is the most familiar setting for Markov chains on $\mathbb{R}^d$ (Meyn and Tweedie, 2009).

Alas these techniques do not give useful quantitative information about the stationary distribution. Theorem 1 suggests a wide range of quantitative questions—how close is $\mathbf{Y}(\infty)$ to $\mathbf{x}$?—which we cannot answer via theory, but for which we can hope to gain insight via simulation.

A parallel analysis can be carried out in the "continuum limit" framework of $n \to \infty$ limits under $\Upsilon_n(u) = \delta_n W(-u)$. In this continuum framework, there is a density function of ratings at time t, which evolves according to a dynamical system, and it is shown in Jabin and Junca (2015) that analogous to Theorem 1, as $t \to \infty$ the density function of ratings does converge to the density of the strengths.

2.4 How to Measure Error?

Within the basic probability model, we could measure ratings *error* directly as the difference between rating and strength. But for real-world sports we cannot measure true strengths, and (as explained below) it seems more useful to consider errors in predicted *win probabilities*.

Within our model, to any update function Υ we associate the implied win-probability function $W_\Upsilon(u)$ at (10). So when the Elo algorithm with Υ gives ratings $(Y_i(t))$ at the current time t, our implicitly predicted probability of team i beating team j is

$$W_\Upsilon\big(Y_i(t) - Y_j(t)\big).$$

In the model, the true probability is $W(x_i - x_j)$. So there is a "prediction error"

$$W_\Upsilon\big(Y_i(t) - Y_j(t)\big) - W(x_i - x_j)$$

and a natural way to measure the size of prediction error is via the "RMS prediction error" statistic

$$\left(\mathbb{E}\frac{1}{n(n-1)}\sum_i\sum_{j\ne i}\big(W_\Upsilon\big(Y_i(t) - Y_j(t)\big) - W(x_i - x_j)\big)^2\right)^{1/2}. \tag{11}$$

We use this notion of *error* throughout. Note it is **root**-mean-square, and so is directly interpretable as typical additive error in estimating a probability.

2.5 The Prediction Tournament Paradox

Can one assess a person's accuracy at estimating probabilities of future real-world events, when the true probabilities are unknown? To do so seems paradoxical, like saying one can grade exam answers to questions whose correct answers are unknown. But the key point is that one can assess *relative* accuracy. If C and D give estimates $p_C(i)$, $p_D(i)$ for events $A(i)$ with unknown true probabilities $(q(i), 1 \le i \le m)$, then

we can quantify their prediction MSE as $\text{error}(C) = m^{-1}\sum_i (p_C(i) - q(i))^2$ and similarly $\text{error}(D)$. After the events are determined, we can award a score to C as $\text{score}(C) = m^{-1}\sum_i (1_{A(i)} - p_C(i))^2$, and similarly award $\text{score}(D)$ to D. Decomposition of variance gives

$$\mathbb{E}\big(\text{score}(C) - \text{score}(D)\big) = \text{error}(C) - \text{error}(D). \tag{12}$$

So for large m the difference in accuracy, that is $\text{error}(C) - \text{error}(D)$, is well estimated by the observed score difference, without assuming anything about the unknown true probabilities.

The fascinating book (Tetlock, 2006) describes this basic mathematics in the course of detailed study of the accuracy of "expert judgement" in geopolitics. In our sports context, this means that we can compare relative accuracy of different win-probability forecasts for real-world matches, whether from Elo-type ratings or from more elaborate modeling or from gambling odds. See Section 5.1 for further comments. Potential undergraduate projects mentioned in this article would use this methodology.

2.6 Mismatch Error

In the real world, we do not know the win-probability function, so we cannot use an update function satisfying the balance relation. Is there much harm in using the logistic regardless? Persi Diaconis (personal communication) notes that, in the analysis of binary data, statisticians use the logistic as a default and regard it as quite robust; see, for example, Cox and Snell (1989), Section 2.7. Is this true in our context?

We can study this question, in the nonchanging strengths setting, by considering the "slow update" limit in which we fix Υ but use a scaled update function $\delta\Upsilon$. In the $\delta \to 0$ limit, the (random) Elo rating process, appropriately time-scaled, converges to a deterministic dynamical system. The conclusion (Aldous, 2017) is that, given (W, Υ) and strengths $\mathbf{x} = (x_1, \ldots, x_n)$, the large-$t$ Elo ratings $\mathbf{Y}(t)$ will approximate, for small δ, the solution $\mathbf{y}(\mathbf{x})$ of a certain fixed-point equation $\mathbf{y} = \Gamma_{\mathbf{x}}\mathbf{y}$. This solution is different from $\mathbf{x}$ when the balance relation (6) does not hold—call this *mismatch error*. As at (11), we quantify that as RMS mismatch error

$$\nu_{\mathbf{x}} = \Bigg(\frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\big(W_\Upsilon\big(y_i(\mathbf{x}) - y_j(\mathbf{x})\big) - W(x_i - x_j)\big)^2\Bigg)^{1/2}.$$

TABLE 2
RMS mismatch error

W / Υ	Logistic Linear	Logistic Cauchy	Cauchy Logistic	Cauchy Linear	Linear Logistic	Linear Cauchy
$\sigma = 0.5$	0.9%	1.1%	1.2%	2.4%	3.2%	6.4%
$\sigma = 1.0$	2.9%	2.9%	2.5%	5.4%	3.2%	6.0%

Even in this deterministic limit, giving theoretical bounds on mismatch error seems a hard problem. Table 2 shows some simulation results, in which we averaged the RMS mismatch error $\nu_{\mathbf{x}}$ over i.i.d. Normal$(0, \sigma^2)$ realizations of strengths $\mathbf{x} = (x_i, 1 \le i \le n)$. We use $\sigma = 0.5$ and 1 because the resulting spreads of win-proportions in a season bracket those of most real-world professional league sports. In the table, *Cauchy* means W is the distribution function of the standard Cauchy distribution, and *linear* means W is the distribution function of the uniform $[-1, 1]$ distribution, with corresponding names for update functions $\Upsilon(u) = W(-u)$.

Note that errors are unchanged if Υ is "stretched", that is replaced by $\Upsilon^{(c)}(u) = \Upsilon(u/c)$. Instead, what matters is the spread of strengths (here represented by σ) relative to the spread of the distribution W.

Compared with the other sources of error described later, the mismatch errors in Table 2 are surprisingly small, except perhaps for the linear/Cauchy combinations, which intuitively seem rather extreme possibilities.

3. THE CENTRAL QUESTION: HOW WELL DOES ELO TRACK CHANGING STRENGTHS?

The Elo rating algorithm is implicitly intended for games where many individuals are to be rated but where there is no systematic scheduling of matches. Chess and tennis are longstanding examples, but nowadays these ratings are widely used in online games. Ratings have several uses: enabling individuals to know their relative strength, eligibility for advanced tournaments, arranging matches to be between equally strong players. A key point is that we expect strengths to change with time. Indeed this is a vital feature of both amateur and professional sport—players in the former, and spectators in the latter, hope performance will improve. Sports would be very dull otherwise.

At this point, we diverge from the analogy with Bradley–Terry consensus rankings from a fixed data set. In that context, given real world data, one can simply calculate MLEs (or Bayes analogs) of the strength

parameters. In the sport context, to do this precisely one would need to update *all* strength estimates after each single match. More importantly, the Bradley–Terry set-up does not naturally allow strengths to vary with time. Of course, one can make particular models of time-varying strengths and study optimal estimation procedures for the particular model (see, e.g., Cattelan, Varin and Firth, 2013, Glickman, 2001, Knorr-Held, 2000) but it is hard to believe that any particular model would be plausible across the range of sports where Elo ratings are used.

To summarize, the plausible conceptual advantages of Elo-type ratings over classical statistical parameter estimation methods are:

- They are dynamic: a rating is updated only after a match by that player, and in a very simple human-interpretable way.
- They implicitly give more weight to recent matches, and avoid needing to explicitly choose how much past data to use.
- They provide a general purpose method of tracking strength changes without assuming any specific model for changing strengths.

But can we go beyond rhetoric: how accurate are these ratings, either in absolute terms or in comparison with other methods?

The web site (Curiel, 2017) maintaining Elo ratings for international football teams asserts:

> ratings tend to converge on a team's true strength relative to its competitors after about 30 matches.

A search engine finds this sentence verbatim in other online venues, evidently copied from some original source, but we have been unable to find its origin, or any theoretical or empirical foundation. By analogy, a search on "seven shuffles suffice" (to mix a deck of cards) finds not only nontechnical discussion but also an actual underlying theorem (Bayer and Diaconis, 1992). So, is there any foundation for the "30 matches suffice" Elo assertion?[7]

Within our models, one can obtain various asymptotic mathematical results (Section 4.6) but such asymptotic regimes are scarcely relevant to real world sports. More informative results can be found via simulation, and such simulation study is the main focus of this article.

[7]Several people have suggested this might arise from traditional Statistics textbooks asserting that sample size 30 suffices for Normal approximation.

Conceptually, there are three sources of error in ratings:

- **Mismatch error** caused by using an update function not adapted to the win-probability function.
- **Lag error** caused by our data coming from past results affected by past, rather than current, strength.
- **Noise** caused by the randomness of recent match results.

In simulations, we use logistic W and Υ, so there is no mismatch error. As we shall see, in choosing the update scaling δ there is an obvious tradeoff between lag error (which increases as δ decreases) and noise error (which increases as δ increases). The tables later show the RMS prediction errors for the optimal value of δ, which is also shown. Note that these choices will tend to make the simulation predictions better than real-world predictions. We will use three different models for time-varying strengths, each being a stationary random process with the same long-run average strength for each team (but Section 4.5 removes this rather unrealistic assumption) and each having a "relaxation time" parameter τ indicating the number of matches required for a team's strength to change substantially. So in each model the errors depend on σ and τ, where as earlier σ^2 is the variance of strength over teams.

To jump to the bottom line, let us summarize the conclusion of our simulations as follows:

> Under plausible models of time-varying strengths, the typical error of predicted win-probabilities will not be substantially less than 10%, regardless of number of matches played.

Readers may judge for themselves whether our models and parameter values are indeed plausible. To us, this contradicts the "30 matches suffice" assertion quoted earlier, but again this is a matter of judgment.

4. SIMULATION RESULTS

The subsections below describe simulation results for each of three models of time-varying strengths. In these models, at each time step we randomly assign all n teams into $n/2$ matches, so here time $t = 0, 1, 2, \ldots$ indicates the number of games that each team has played. In the simulations, we use $n = 20$, but the results are insensitive to the exact value of n.

Our models give a stationary process $(\mathbf{X}(t), t = 0, 1, 2, \ldots)$ of team strengths $\mathbf{X}(t) = (X_i(t), 1 \le i \le n)$, The Elo rating algorithm gives ratings $\mathbf{Y}(t) =$

$(Y_i(t), 1 \le i \le n)$, and our simulation results refer to the jointly stationary process $((\mathbf{X}(t), \mathbf{Y}(t)), t = 0, 1, 2, \ldots)$ for which initial ratings do not matter.

As before, our statistic for measuring the quality of the Elo ratings is the root-mean-square error of the win-probabilities predicted from the Elo ratings, in this context the number μ defined by

$$\mu^2 = \frac{1}{n(n-1)} \mathbb{E} \sum_i \sum_{j \ne i} \big(L(X_i(t) - X_j(t)) - L(Y_i(t) - Y_j(t))\big)^2$$

for the jointly stationary process. Call this statistic *RMSE-p*, where the p is a reminder that we are estimating *probabilities*, not outcomes. Our strength models are exchangeable over teams, so

$$\mu^2 = \mathbb{E}\big(L(X_1(t) - X_2(t)) - L(Y_1(t) - Y_2(t))\big)^2.$$

In each model, we show "realization" figures, which show the strengths and ratings of two teams over a 128-game window, so one can see how well the ratings track the changing strengths (labeled as "ability" here).

4.1 The Cycle Model

This first model is obviously unrealistic, but provides a test of mathematical intuition. Strengths follow a deterministic cycle, with a random shift for each team:

$$X_i(t) = 2^{1/2}\sigma \sin\Big(U_i + \frac{\pi}{2\tau} t\Big),$$

where U_i is uniform on $[0, 2\pi]$. So $\operatorname{var} X_i(t) = \sigma^2$, and the "relaxation time" parameter τ indicates the number of games required for strength to decrease from the maximum to the average value.

The realizations in Figure 3 agree with intuition. When the update scaling is comparatively small (left,

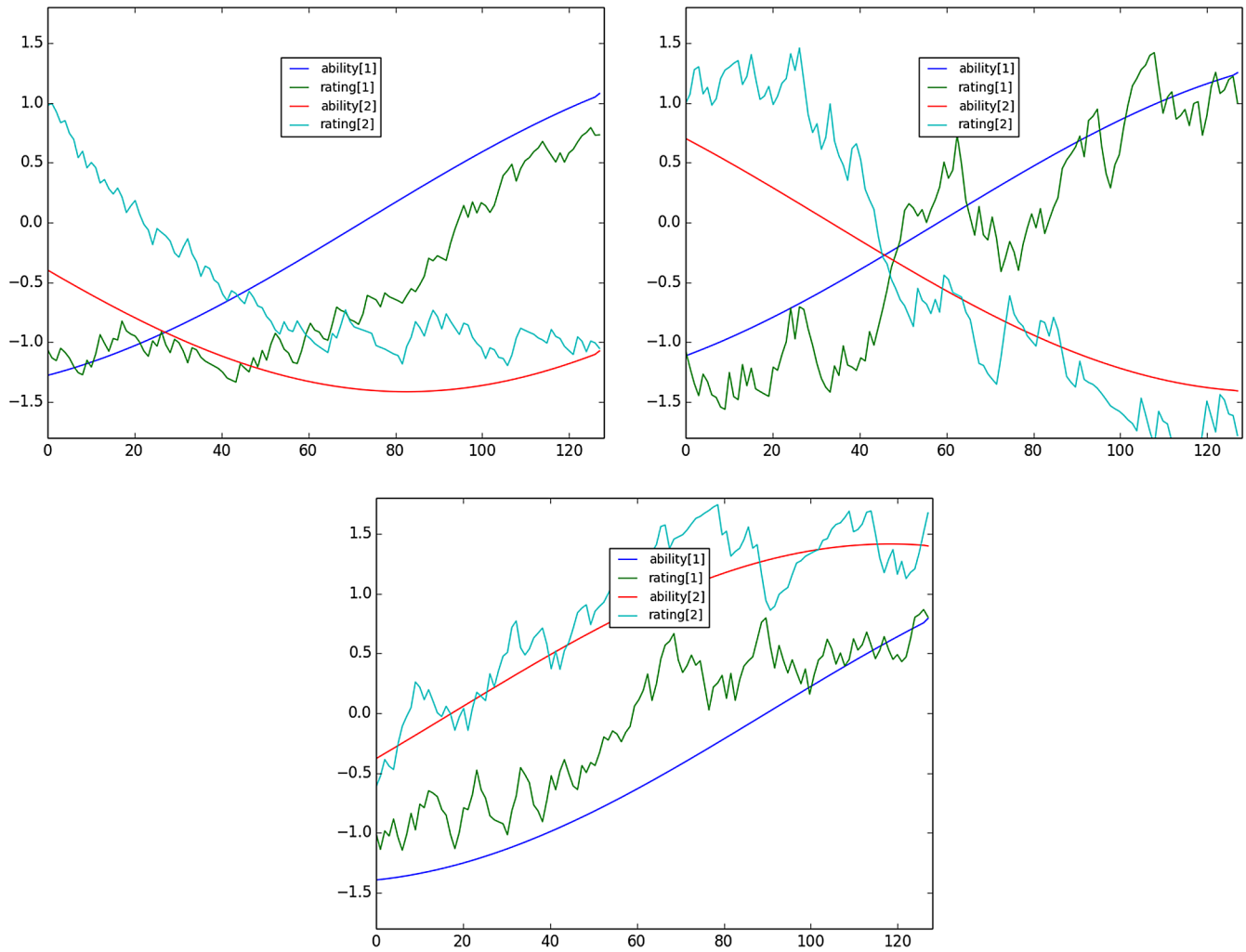

FIG. 3. *Realizations of the cycle model* ($\sigma = 1$, $\tau = 100$, *logistic* W *and* Υ) *for* $\delta = 0.17$ (*top left*) $\delta = 0.35$ (*top right*) *and* $\delta = 0.26$ (*optimal*, *bottom*).

TABLE 3
Cycle model: *RMSE-p and* (*optimal* δ)

	τ			
σ	**50**	**100**	**200**	**400**
0.5	13.4% (0.20)	10.8% (0.14)	8.6% (0.09)	6.8% (0.06)
1.0	15.7% (0.40)	12.5% (0.27)	9.9% (0.17)	8.0% (0.12)

TABLE 4
Ornstein–Uhlenbeck model: *RMSE-p and* (*optimal* δ)

	τ			
σ	**50**	**100**	**200**	**400**
0.5	12.9% (0.11)	11.1% (0.09)	9.5% (0.08)	8.2% (0.07)
1.0	17.0% (0.28)	14.6% (0.24)	12.4% (0.16)	10.4% (0.14)

$\delta = 0.17$) the ratings follow a cycling curve fairly closely (the "noise" from random wins/losses is small) but lag in time behind the true strength curve. When the update scaling is comparatively large (right, $\delta = 0.35$) the ratings show larger noise but their averages track strengths better. This bias-variance tradeoff is optimized at some intermediate value, at $\delta = 0.26$.

Table 3 shows numerical values of RMSE-p for the optimal values of δ in this cycle model, We will discuss these numbers from the three models in Section 4.4.

4.2 The Ornstein–Uhlenbeck Strength Model

Here, we model strengths as a discretized Ornstein–Uhlenbeck process (ARMA process). In the *standard* process, for each team i the process $(X_i^{(\tau)}(t), t = 0, 1, 2, \ldots)$ of strengths evolves as

$$X_i^{(\tau)}(t+1) = (1-\tau^{-1})X_i^{(\tau)}(t) + \sqrt{1-(1-\tau^{-1})^2}Z_i(t),$$

where $(Z_i(t), t \geq 1)$ are IID Normal$(0, 1)$. This gives a stationary process with Normal$(0, 1)$ marginal. The processes are independent for different teams, so for a league of n teams we have a combined process $\mathbf{X}^{(\tau)}(t) = (X_i^{(\tau)}(t), 0 \leq i \leq n)$. The parameter $\tau \gg 1$ is the relaxation time. Finally, we can scale by the factor σ to define a 2-parameter "Ornstein–Uhlenbeck strength model"

$$\mathbf{X}(t) = \mathbf{X}^{(\tau,\sigma)}(t) = \sigma\mathbf{X}^{(\tau)}(t).$$

Figure 4 shows realizations, comparing $\tau = 400$ (left) with $\tau = 50$ (right), using the optimal values of δ (0.16 and 0.28). Table 4 gives numerical values of RMSE-p for the optimal values of δ.

4.3 The Jump Model

In this model, for each team i the strength process remains constant for a Geometric$(1/\tau)$ time, then jumps to an independent Normal$(0, \sigma^2)$ value. Figure 5 shows realizations, comparing $\tau = 400$ (left) with $\tau = 50$ (right), using the optimal values of δ (0.13 and 0.30). Table 5 gives numerical values of RMSE-p for the optimal values of δ.

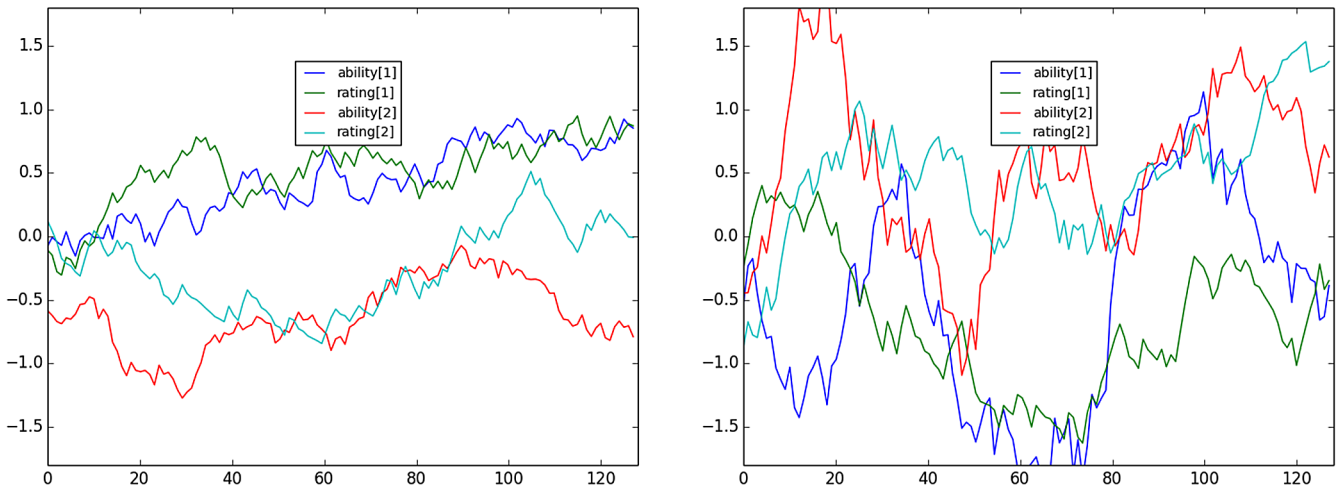

FIG. 4. *Realizations of the Ornstein–Uhlenbeck model*: $\sigma = 1$.

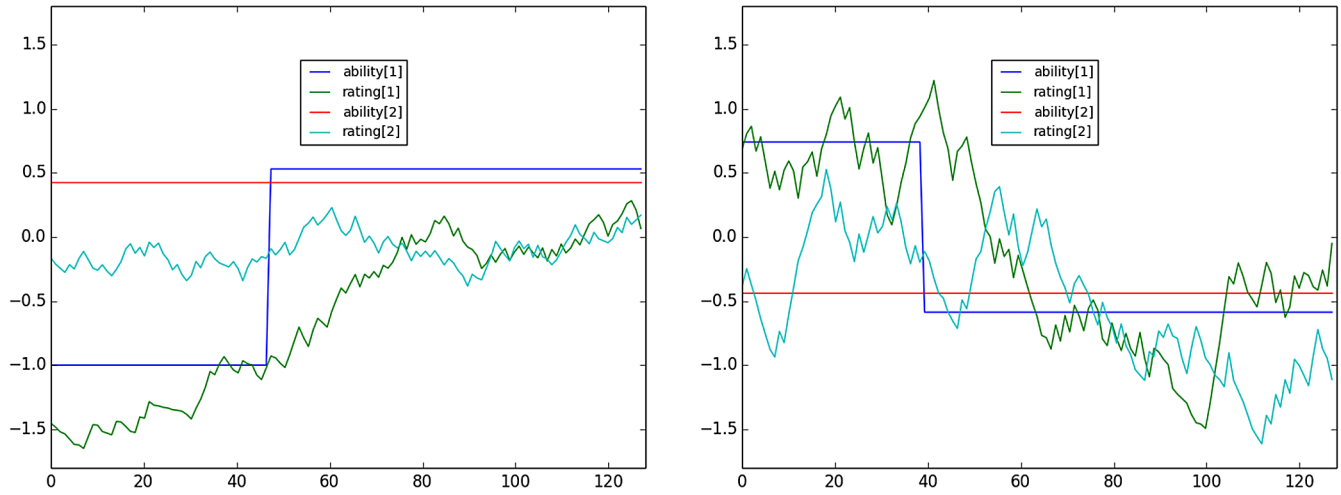

FIG. 5. *Realizations of the jump model*: $\sigma = 1$.

4.4 Discussion of Simulation Results in These 3 Models

Of course, none of these three models is very plausible from a real-world viewpoint, but they serve to represent different "extreme" possibilities.

In relating the graphics to the tabled values, recall that the maximum slope of the logistic function is $L'(0) = 1/4$, so that a 10% error in estimating win probability corresponds to a somewhat more than 0.4 error in strength difference.

As mentioned before, the optimal choice of δ involves a trade-off between *lag*—the fact that our data arises from past strength, not current strength—and the *noise* from the random outcomes of recent matches. This type of bias-variance trade-off is of course very familiar to statisticians.

As intuition suggests, increasing τ make the Elo ratings more accurate and decreases the optimal δ. Increasing σ, in the range 0.5–1.0 we consider, makes them less accurate, which is less immediate intuitively. One might think that increasing the variability of strengths would make it easier to assess a team's strength. As regards assessing a given team's strength as a percentile amongst the league of teams, this intuition is likely correct. However, the fact the typical strength-differences are larger makes the effect of errors on assessing strengths larger; also our parametrization makes the absolute rate of change of strength per game proportional to σ for fixed τ, so that changes in strength are harder to track.

TABLE 5
Jump model: *RMSE-p and* (*optimal* δ)

	τ			
σ	**50**	**100**	**200**	**400**
0.5	12.8% (0.12)	11.2% (0.09)	9.8% (0.06)	8.4% (0.05)
1.0	16.9% 0.30	14.5% 0.24	12.4% 0.19	10.5% 0.14

The numerical values in Tables 4 and 5 are remarkably close, which suggests that the prediction accuracy of Elo ratings may be somewhat insensitive to the details of the model for time-varying strengths, and instead determined mainly by the parameters σ and τ (in our "no long term difference in average strength" setting). On the other hand, one can easily give convincing heuristics for $\tau \to \infty$ asymptotics in these models (see Section 4.6 for the cycle and Ornstein–Uhlenbeck models) and these show different asymptotic behavior in the three models. So we cannot explain the coincidence of numerical values via asymptotics.

Because RMSE-p is a smooth function of the update scaling parameter δ, its values must be almost constant in some neighborhood of the optimal δ, and this neighborhood can be quite large. For instance, for the jump model with $\tau = 50$, $\sigma = 1.0$ (Table 5) the value of RMSE-p hardly varies over the range $0.24 \le \delta \le 0.36$. This suggests that accuracy of Elo ratings are not very sensitive to choice of update scaling parameter.

For the cycle model, and the Ornstein–Uhlenbeck model with larger τ, we see numerically that the optimal value of δ for $\sigma = 1$ is around twice the optimal value for $\sigma = 0.5$. We do not have a good explanation,

but it suggests how one might take into account the diversity of strengths in a real-world implementation.

Regarding the absolute values of RMS errors, if we always predicted 50% win probabilities, the RMS error would be 15.9% ($\sigma = 0.5$) or 26.1% ($\sigma = 1$). The tables show, very roughly, that in order for the Elo ratings to cut this error in half, we need a relaxation time τ of order 150–200 games. This is larger than intuition might suggest. At first sight, it might be a consequence of the unrealistic "no long term difference in average strength" setting, but in fact this has less effect than one might think (Section 4.5).

Given any specific model for time-varying strengths, there is in theory some optimal way to predict win-probabilities, via some model-dependent rule involving the entire past of the win/lose process of all teams. Where it is computationally feasible to calculate the optimal prediction? This suggests another simulation project:

> How close to optimal—as regards RMSE-p or some other criterion—is the Elo rating, in a particular model?

4.5 Differing Long-Term Average Strengths

In the previous models, the long-run average strength was the same for each team. Manchester United and the New York Yankees remind us that for real-world professional team sports, long-term average performance has not been the same for all teams in a league. Let us look at one model. Take $0 < \alpha < 1$ and suppose:

- Long-term average strengths of different teams are independent Normal$(0, \alpha\sigma^2)$.
- Fluctuations of a team's strength around its long-term average follow the discrete Ornstein–Uhelnbeck process as in Section 4.2, with relaxation time τ and stationary variance $(1-\alpha)\sigma^2$.

So the team strengths have variance σ^2, with the relative contributions to variance being α from the diversity of long-term averages and $1-\alpha$ from the shorter-term fluctuations.

Intuition suggests that Elo ratings should be more accurate as α increases (note that $\alpha = 0$ case is the "equal long-term average strengths" model), and this is borne out by simulations. Table 6 gives numerical values of RMSE-p for the optimal values of δ. But even in the rather extreme case where 75% of variance is due to different long-term strengths, the win-probability prediction errors are not greatly reduced.

TABLE 6
RMSE-p and (optimal δ) for the "differing long-term average strengths" model: $\sigma = 1.0$

	τ			
α	**50**	**100**	**200**	**400**
0.0	17.0% (0.28)	14.6% (0.24)	12.4% (0.16)	10.4% (0.14)
0.5	13.9% (0.20)	12.0% (0.17)	10.4% (0.13)	8.8% (0.10)
0.75	11.2% (0.14)	9.8% (0.11)	8.5% (0.09)	7.3% (0.07)

4.6 Scaling Laws for Slowly Varying Strengths

Consider a model of smoothly time-varying strengths, and introduce the relaxation time τ as a "stretch" parameter, as in our cycle model. If τ is large, then the update factor δ will be small. Here, we are in a classical setting of Ornstein–Uhlenbeck approximations to stable dynamical systems with small noise (Gardiner, 1983, Chapter 6). As outlined in Aldous (2017) (informally, but could be readily be rephrased as rigorous asymptotics), we find the following small noise asymptotics:

> In a model of smoothly time-varying strengths, as $\tau \to \infty$ the optimal update factor δ scales as $\tau^{-2/3}$, and the resulting RMSE-p error statistic scales as $\tau^{-1/3}$.

And the data from Table 3 (the cycle model) is consistent with this conclusion. Similarly, one can analyze the Ornstein–Uhlenbeck strength model from Section 4.2 in the $\tau \to \infty$ limit. Here, the informal calculations say:

> the optimal update factor δ scales as $\tau^{-1/2}$, and the resulting RMSE-p error statistic scales as $\tau^{-1/4}$.

Looking at the Table 4 data, this prediction works quite well for the RMSE-p error but less well for the optimal δ, perhaps because of the next order terms.

5. FINAL REMARKS

5.1 Do Simulations Relate to Real Data?

It is important to remember that the simulations above are within a model in which we know the true win-probability for each match and then study the accuracy of the win-probability implied by Elo ratings.

For real world sports data, we do not know true win-probabilities, so what can such simulations tell us?

There is some literature comparing Elo-type ratings with other methods of predicting *outcomes* of matches. See, for instance, the recent work of Kovalchik (2016) for tennis, Hvattum and Arntzen (2010) for English league football and Lasek, Szlávik and Bhulai (2013) for international football.[8] The latter article shows different methods having outcome MSEs in the range 12%–15%, though because of the frequency of draws in football these numbers are not directly comparable to our win-or-lose setting.[9] In the terminology of Section 2.5, these outcome-MSEs are the *scores* in a prediction tournament. Identity (12) shows that the *difference* in scores of two methods is a good estimate of the *difference* in their accuracy in predicting probabilities, defined as the *difference* in MSE-p, that is the square of our RMSE-p.

To relate this to simulations, suppose we see outcome MSE scores of 11.75% and 14% for two prediction schemes S_1 and S_2. This tells us that the difference in MSE-p is 2.25%, but is consistent with a spectrum of possible *absolute* errors in predicting win-probabilities:

- S_1 has 0 RMS-p error and S_2 has 15% RMS-p error;
- S_1 has 13% RMS-p error and S_2 has 20% RMS-p error;
- or S_1 has 20% RMS-p error and S_2 has 25% RMS-p error.

We cannot distinguish between these possibilities because we do not know the true win-probabilities.

As a criterion for accuracy in predicting probabilities in our theoretical study, we invoked an arbitrary benchmark of 10% RMSE-p, that is 1% MSE-p, as a goal. If that were achieved by Elo rankings on real data, that would say[10] that we could expect no other algorithm based only on past results to beat Elo by more than 1% in outcome MSE, however accurate the stochastic model of changing strengths.

Another project would be to treat the probabilities derived from gambling odds as true probabilities, and examine the accuracy of the probabilities derived from Elo ratings.

[8]Such work usually finds that Elo-type rankings work better than official rankings.

[9]MSEs would be larger in our setting.

[10]After modifying for actual Elo implementations.

5.2 Statistical Analysis

This article has not attempted any substantial discussion of statistical analysis of data, for several reasons including lack of expertise by the author. Fortunately, Király and Qian (2017) provides a very recent lengthy account of statistical methods related to the topics here. As observed by a reviewer, a natural framework for analysis of the basic time-static Bradley–Terry model is provided by generalized linear models. To quote Király and Qian (2017):

> Generalized linear models generalize both linear and log-linear models (such as the Bradley–Terry model) through so-called link functions, or more generally (and less classically) link distributions, combined with flexible structural assumptions on the target variable. The generalization aims at extending prediction with linear functionals through the choice of link which is most suitable for the target (for an overview, see McCullagh and Nelder, 1989). Particularly relevant for us are generalized linear models for ordinal outcomes which includes the ternary (win/draw/lose) case, as well as link distributions for scores. Some existing extensions of this type, such as the ternary outcome model and the score model, may be interpreted as specific choices of suitable linking distributions. How these ideas may be used as a component of structured log-odds models will be discussed later.

Readers may pursue these topics in Király and Qian (2017).

5.3 "Probability in the Real World"

This article is an extended write-up of a lecture in the author's "Probability in the Real World" course at U.C. Berkeley. The course consists of around 20 lectures on different topics, each (ideally) "anchored" by contemporary data (here the Elo football ratings), and each (ideally) offering scope for student projects based on contemporary data. Two other extended write-ups are available, one on martingales and prediction markets (Aldous, 2013) and one on a game-theoretic analysis of an online game which one can observe being played ("by your grandmothers") in real time (Aldous and Han, 2017).

ACKNOWLEDGMENT

Research supported by NSF Grant DMS-1504802.

REFERENCES

ADLER, I., CAO, Y., KARP, R., PEKOZ, E. and ROSS, S. M. (2016). Random knockout tournaments. *Oper. Res.* To appear. Available at arXiv:1612.04448.

ALDOUS, D. J. (2013). Using prediction market data to illustrate undergraduate probability. *Amer. Math. Monthly* **120** 583–593. MR3096465

ALDOUS, D. (2017). Mathematical probability foundations of dynamic sports ratings: Overview and open problems. To appear.

ALDOUS, D. and HAN, W. (2017). Introducing Nash equilibria via an online casual game that people actually play. *Amer. Math. Monthly* **124** 506–517. MR3654829

BAYER, D. and DIACONIS, P. (1992). Trailing the dovetail shuffle to its lair. *Ann. Appl. Probab.* **2** 294–313. MR1161056

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925

CATTELAN, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statist. Sci.* **27** 412–433. MR3012434

CATTELAN, M., VARIN, C. and FIRTH, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 135–150. MR3042315

CHETRITE, R., DIEL, R. and LERASLE, M. (2017). The number of potential winners in Bradley–Terry model in random environment. *Ann. Appl. Probab.* **27** 1372–1394. MR3678473

COX, D. R. and SNELL, E. J. (1989). *Analysis of Binary Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **32**. Chapman & Hall, London. MR1014891

CURIEL, R. S. DA S. (2017). World Football Elo Ratings. Available at http://www.eloratings.net.

DAVID, H. A. (1988). *The Method of Paired Comparisons*, 2nd ed. *Griffin's Statistical Monographs & Courses* **41**. Charles Griffin & Co., Ltd., London. MR0947340

GARDINER, C. W. (1983). *Handbook of Stochastic Methods*: *For Physics*, *Chemistry and the Natural Sciences*. *Springer Series in Synergetics* **13**. Springer, Berlin. MR0714148

GLICKMAN, M. E. (2001). Dynamic paired comparison models with stochastic variances. *J. Appl. Stat.* **28** 673–689. MR1862491

HVATTUM, L. M. and ARNTZEN, H. (2010). Using Elo ratings for match result prediction in association football. *Int. J. Forecast.* **26** 460–470.

JABIN, P.-E. and JUNCA, S. (2015). A continuous model for ratings. *SIAM J. Appl. Math.* **75** 420–442. MR3323555

KIRÁLY, F. J. and QIAN, Z. (2017). Modelling competitive sports: Bradley–Terry–Elo models for supervised and online learning of paired competition outcomes. Available at arXiv:1701.08055.

KNORR-HELD, L. (2000). Dynamic rating of sports teams. *Statistician* **49** 261–276.

KOVALCHIK, S. (2016). Searching for the GOAT of tennis win prediction. *J. Quant. Anal. Sports* **12** 127–138.

LANGE, K. (2010). *Applied Probability*, 2nd ed. Springer, New York. MR2680838

LANGVILLE, A. N. and MEYER, C. D. (2012). *Who's #1? The Science of Rating and Ranking*. Princeton Univ. Press, Princeton, NJ. MR2894870

LASEK, J., SZLÁVIK, Z. and BHULAI, S. (2013). The predictive power of ranking systems in association football. *Int. J. Appl. Pattern Recogn.* **1** 27–46.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. MR3223057

MEYN, S. and TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2509253

RESNICK, S. I. (1987). *Extreme Values*, *Regular Variation*, *and Point Processes*. *Applied Probability. A Series of the Applied Probability Trust* **4**. Springer, New York. MR0900810

TETLOCK, P. E. (2006). *Expert Political Judgment*: *How Good Is It? How Can We Know?* Princeton Univ. Press, Princeton, NJ.

UNITED, O. (2017). Weekly Tennis ELO Rankings. Available at https://tenniseloranking.blogspot.com.

WIKIPEDIA (2014a). Elo rating system. Wikipedia, the free encyclopedia. [Online; accessed, 31-October-2014].

WIKIPEDIA (2014b). Tournament. Wikipedia, the free encyclopedia. [Online; accessed. 4-December-2014].

WIKIPEDIA (2017). Promotion and relegation. Wikipedia, the free encyclopedia. [Online; accessed 19-February-2017].