Understanding and Pushing the Limits of the Elo Rating Algorithm

Leszek Szczecinski and Aymen Djebbi

Abstract

This work is concerned with the rating of players/teams in face-to-face games with three possible outcomes: loss, win, and draw. This is one of the fundamental problems in sport analytics, where the very simple and popular, non-trivial algorithm was proposed by Arpad Elo in late fifties to rate chess players. In this work we explain the mathematical model underlying the Elo algorithm and, in particular, we explain what is the implicit but not yet spelled out, assumption about the model of draws. We further extend the model to provide flexibility and remove the unrealistic implicit assumptions of the Elo algorithm. This yields the new rating algorithm, we call κ-Elo, which is equally simple as the Elo algorithm but provides a possibility to adjust to the frequency of draws. The discussion of the importance of the appropriate choice of the parameters is carried out and illustrated using results from English Premier League football seasons.

I. Introduction

Rating of players/teams is, arguably one of the most important issues in sport/ competition analytics. In this work we are concerned with rating of the players/teams in the sports with one-on-one games yielding ternary results of win, loss and draw; such a situation appear in almost all team sports and many individual sports/competitions.

Rating in sports consists in assigning a numerical value to a player/team using the results of the past games. While most of the sports' ratings use the points which are attributed to the game's winner, the rating algorithm which was developed in late fifties by Arpad Elo in the context of chess competition (Elo, 2008), and adopted later by Fédération Internationale des Échecs (FIDE), challenged this view.

Namely, the Elo algorithm changes the players' rating using not only the game outcome but also the ratings of the players before the game. The Elo algorithm is arguably one of the most popular, non-trivial rating algorithm and was used to analyze different sports, although mostly informally (Langville and Meyer, 2012, Chap. 5)(Wikipedia contributors, 2019); it is also used for rating in eSports (Herbrich and Graepel, 2006). Moreover, in 2018, the Elo algorithm was adopted, under the name "SUM", by Fédération Internationale de Football Association (FIFA) for the rating of the national football teams (FIFA, 2019). The Elo algorithm thus deserves a particular attention particularly because it is often presented without mathematical details behind its derivation, which may be quite confusing.

In this work we adopt the probabilistic modelling point of view, where the game outcomes are related to the rating by conditional probabilities. The advantage is that, to find the rating, we can use the conventional estimation strategies, such as maximum likelihood (ML); moreover, with the well defined model, once the ratings are found, they can be used for the purpose of prediction, which we understand as defining the distribution over the results of the game to come. This well-known mathematical formalism of rating in sport has been developed in psychometrics for rating the preferences in pairwise-

L. Szczecinski and A. Djebbi are with INRS, Montreal, Canada. [e-mail: {leszek, aymen.djebbi}@emt.inrs.ca]. The work was supported by NSERC, Canada.

comparison setup (Thurston, 1927)(Bradley and Terry, 1952) and the problem was deeply studied and extended in different directions e.g., (Cattelan, 2012)(Caron and Doucet, 2012).

In this work we are particularly concerned with the mathematical modelling of draws (or ties). This issue has been addressed in psychometrics via two distinct approaches: in (Rao and Kupper, 1967), using thresholding of the unobserved (latent) variables and in (Davidson, 1970)–via an axiomatic approach. These two approaches have also been applied in sport rating, e.g., (Herbrich and Graepel, 2006)(Joe, 1990); the former, however, is used more often than the latter.

We note that the draws are not modelled in the Elo algorithm (Elo, 2008). In fact, and more generally, the outcomes are not explicitly modelled at all; rather, to derive the algorithm, the probabilistic model is combined with the strong intuition of the author; no formal optimality criteria is defined. Nevertheless, it was later observed that the Elo algorithm actually finds the approximate ML ratings estimates in the binary-outcome (win-loss) games (Király and Qian, 2017).

As for the draws, the Elo algorithm considers them by using the concept of a fractional score (of the game). However, since the underlying model is not specified, in our view there is a logical void: on the one hand, the Elo algorithm includes draws, on the other hand, there is no model allowing us to calculate the draw probability. The objective of this work is to fill this gap.

The paper is organized as follows. We define the mathematical model of the problem in Sec. II. In Sec. III we show how the principle of ML combined with the stochastic gradient (SG) yield the Elo algorithm in the binary-outcome games. We treat the issue of draws in Sec. IV; this is where the main contributions of the paper are found. Namely, we show and discuss the implicit model underlying the Elo algorithm; we also extend the model to increase its flexibility; finally we show how to define its parameters to take into account the known frequency of the draws. In Sec. V we illustrate the analysis with numerical results and the final conclusions are drawn in Sec. VI.

II. Rating: Problem definition

We consider the problem of M players (or teams), indexed by $m = 1, \ldots, M$, challenging each other in face-to-face games. At a time n we observe the result/outcome y_n of the game between the players defined by the pair $\boldsymbol{i}_n = \{i_{\mathrm{H},n}, i_{\mathrm{A},n}\}$. The index $i_{\mathrm{H},n}$ refers to the "home" player, while $i_{\mathrm{A},n}$ indicates the "away" player. This distinction is often important in the team games where the so-called home-field advantage may play a role; in other competition such an effect may exist as well, like in chess, the player who starts the game may be considered a home player. We consider three possible game results: i) the home player wins; denoted as $\{i_{\mathrm{H},n} \succ i_{\mathrm{A},n}\}$ in which case $\{y_n = \mathsf{H}\}$; ii) the draw (or tie) $\{y_n = \mathsf{D}\}$, denoted also as $\{i_{\mathrm{H},n} \doteq i_{\mathrm{A},n}\}$; and finally, iii) $\{y_n = \mathsf{A}\}$, which means that the "away" player wins which we denote also as $\{i_{\mathrm{H},n} \prec i_{\mathrm{A},n}\}$.

For compactness of notation, useful in derivations, it is convenient to encode the categorical variable y_n into numerical indicators defined over the set $\{0, 1\}$

$$h_n = \mathbb{I}\big[y_n = \mathsf{H}\big], \quad a_n = \mathbb{I}\big[y_n = \mathsf{A}\big], \quad d_n = \mathbb{I}\big[y_n = \mathsf{D}\big], \tag{1}$$

with $\mathbb{I}\big[\cdot\big]$ being the indicator function: $\mathbb{I}\big[A\big] = 1$ if A is true and $\mathbb{I}\big[A\big] = 0$, otherwise. The mutual exclusivity of the win/loss/draw events guarantees $h_n + a_n + d_n = 1$.

Having observed the outcomes of the games, $y_l, l = 1, \ldots, n$, we want to *rate* the players, i.e., assign a *rating level*—a real number—θ_m to each of them. The rating level should represent the player's ability to win; for this reason it is also called

strength (Glickman, 1999) or *skill* (Herbrich and Graepel, 2006)(Caron and Doucet, 2012). The ability should be understood in the probabilistic sense: no player has a guarantee to win so the outcome y_n is treated as a realization of a random variable Y_n. Thus, the levels $\theta_m, m = 1, \ldots, M$ should provide a reliable estimate of the distribution of Y_n over the set $\{\mathsf{H}, \mathsf{A}, \mathsf{D}\}$. In other words, the formal rating becomes an expert system explaining the past– and predicting the future results.

A. Win-loss model

It is instructive to consider first the case when the outcome of the game is binary, $y_n \in \{\mathsf{H}, \mathsf{A}\}$, i.e., for the moment, we ignore the possibility of draws, D, and we consider them separately in Sec. IV. In this case we are looking to establish the probabilistic model linking the result of the game and the rating levels of the involved players. By far the most popular approach is based on the so-called linear model (David, 1963, Ch. 1.3)

$$\Pr\{i \succcurlyeq j | \theta_i, \theta_j\} = \Phi_{\mathsf{H}}(\theta_i - \theta_j), \tag{2}$$

where $\Phi_{\mathsf{H}}(v)$ is an increasing function which satisfies

$$\lim_{v \to -\infty} \Phi_{\mathsf{H}}(v) = 0, \quad \lim_{v \to \infty} \Phi_{\mathsf{H}}(v) = 1, \tag{3}$$

and thus we may set $\Phi_{\mathsf{H}}(v) = \Phi(v)$, where $\Phi(v)$ is a conveniently chosen cumulative density function (CDF). By symmetry, $\Pr\{i \succcurlyeq j\} + \Pr\{j \prec i\}$, we obtain

$$\Phi_{\mathsf{H}}(v) = \Phi(v), \quad \Phi_{\mathsf{A}}(v) = \Phi(-v) = 1 - \Phi(v), \tag{4}$$

where the last relationship comes from the law of total probability, $\Pr\{i \succcurlyeq j\} + \Pr\{i \prec j\} = 1$ (remember, we are dealing with binary-outcome games).

Indeed, (2) corresponds to our intuition: the growing difference between rating levels $\theta_i - \theta_j$ should translate into increasing probability of user i winning against the user j.

To emphasize that the entire model is defined by the CDF $\Phi(v)$, which affects both $\Phi_{\mathsf{H}}(v)$ and $\Phi_{\mathsf{A}}(v)$ via (4), we keep the separate notation $\Phi_{\mathsf{H}}(v)$ and $\Phi(v)$ even if they are the same in the case we consider.

A popular choice for $\Phi(v)$ is the logistic CDF (Bradley and Terry, 1952)

$$\Phi(v) = \frac{1}{1 + 10^{-v/\sigma}} = \frac{10^{0.5v/\sigma}}{10^{0.5v/\sigma} + 10^{-0.5v/\sigma}}, \tag{5}$$

where $\sigma > 0$ is a scale parameter.

We note that the rating is arbitrary regarding

- the origin—because any value θ_0 can be added to all the levels θ_m without affecting the difference $v = \theta_i - \theta_j$ appearing as the argument of $\Phi(\cdot)$ in (4),
- the scaling—because the levels θ_m obtained with the scale σ can be transformed into levels θ'_m with a scale σ' via multiplication: $\theta'_m = \theta_m \sigma'/\sigma$, and then the value of $\Phi(\theta_i - \theta_j)$ used with σ is the same value as $\Phi(\theta'_i - \theta'_j)$ used with σ';[1] and

[1]The rating implemented by FIFA uses $\sigma = 600$ (FIFA, 2019), while FIDE uses $\sigma = 400$

- the base of the exponent in (5); for example, $10^{-v/\sigma} = \mathrm{e}^{-v/\sigma'}$ with $\sigma' = \sigma \log_{10} \mathrm{e}$; therefore, changing from the base-10 to the base of the natural logarithm requires replacing σ with σ'.

III. Rating via maximum likelihood estimation

Using the results from Sec. II-A, the random variables, Y_n, and the rating levels are related through conditional probability

$$\Pr\{Y_n = \mathsf{H}|\boldsymbol{x}_n, \boldsymbol{\theta}\} = \Phi_{\mathsf{H}}(v_n) = \Phi(v_n), \tag{6}$$

$$\Pr\{Y_n = \mathsf{A}|\boldsymbol{x}_n, \boldsymbol{\theta}\} = \Phi_{\mathsf{A}}(v_n) = \Phi(-v_n), \tag{7}$$

$$v_n = \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta} = \theta_{i_{\mathrm{H},n}} - \theta_{i_{\mathrm{A},n}}, \tag{8}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]^{\mathrm{T}}$ is the vector which gathers all the rating levels, $(\cdot)^{\mathrm{T}}$ denotes transpose, v_n is thus a result of linear combiner $\boldsymbol{x}_n$ applied to $\boldsymbol{\theta}$, and $\boldsymbol{x}_n$ is the game-scheduling vector, i.e.,

$$\boldsymbol{x}_n = [0, \ldots, 0, \underbrace{1}_{i_{\mathrm{H},n}\text{-th pos.}}, 0, \ldots, 0, \underbrace{-1}_{i_{\mathrm{A},n}\text{-th pos.}}, 0, \ldots, 0]^{\mathrm{T}}. \tag{9}$$

We prefer the notation using the scheduling vector as it liberates us from somewhat cumbersome repetition of the indices $i_{\mathrm{H},n}$ and $i_{\mathrm{A},n}$ as in (8).

Our goal now, is to find the levels $\boldsymbol{\theta}$ at time n using the game outcomes $\{y_l\}_{l=1}^n$ and the scheduling vectors $\{\boldsymbol{x}_l\}_{l=1}^n$. This is fundamentally a parameter estimation problem (model fitting) and we solve it using the ML principle. The ML estimate of $\boldsymbol{\theta}$ at time n is obtained via optimization

$$\hat{\boldsymbol{\theta}}_n = \operatorname*{argmin}_{\boldsymbol{\theta}} J_n(\boldsymbol{\theta}) \tag{10}$$

where

$$J_n(\boldsymbol{\theta}) = -\log \Pr\left\{\{Y_l\}_{l=1}^n = \{y_l\}_{l=1}^n | \boldsymbol{\theta}, \{\boldsymbol{x}_l\}_{l=1}^n\right\}. \tag{11}$$

Further, assuming that conditioned on the levels $\boldsymbol{\theta}$, the outcomes Y_l are mutually independent, i.e., $\Pr\{\{Y_l\}_{l=1}^n = \{y_l\}_{l=1}^n|\boldsymbol{\theta}, \{\boldsymbol{x}_l\}_{l=1}^n\} = \prod_{l=1}^n \Pr\{Y_n = y_n|\boldsymbol{\theta}, \boldsymbol{x}_n\}$, we obtain

$$J_n(\boldsymbol{\theta}) = \sum_{l=1}^n L_l(\boldsymbol{\theta}) \tag{12}$$

$$L_l(\boldsymbol{\theta}) = -\log \Pr\{Y_l = y_l|\boldsymbol{\theta}\} \quad \text{given theta, x\_ell} \tag{13}$$

$$= -h_l \log \Phi_{\mathsf{H}}(\boldsymbol{x}_l^{\mathrm{T}}\boldsymbol{\theta}) - a_l \log \Phi_{\mathsf{A}}(\boldsymbol{x}_l^{\mathrm{T}}\boldsymbol{\theta}), \tag{14}$$

where we applied the model (6)-(7).

A. Stochastic gradient and Elo algorithm

The minimization in (10) can be done via steepest descent which would result in the following operations

$$\hat{\boldsymbol{\theta}}_n \leftarrow \hat{\boldsymbol{\theta}}_n - \mu \nabla_{\boldsymbol{\theta}} J_n(\hat{\boldsymbol{\theta}}_n) \tag{15}$$

iterated (hence the symbol "$\leftarrow$") till convergence for a given n; the gradient is calculated as

$$\nabla_{\boldsymbol{\theta}} J_n(\boldsymbol{\theta}) = \sum_{l=1}^{n} \nabla_{\boldsymbol{\theta}} L_l(\boldsymbol{\theta}), \tag{16}$$

and the step, μ, should be adequately chosen to guarantee the convergence. Moreover, since $J_n(\boldsymbol{\theta})$ is convex,[2] the minimum is global.[3]

From (14) we obtain

$$\nabla_{\boldsymbol{\theta}} L_l(\boldsymbol{\theta}) = -h_l \boldsymbol{x}_l \psi(\boldsymbol{x}_l^{\mathrm{T}} \boldsymbol{\theta}) + a_l \boldsymbol{x}_l \psi(-\boldsymbol{x}_l^{\mathrm{T}} \boldsymbol{\theta}) \tag{17}$$

$$= -\boldsymbol{x}_l e_l(v_l) \tag{18}$$

where, directly from (5) we have

$$\psi(v) = \frac{\mathrm{d}}{\mathrm{d}v} \log \Phi(v) = \frac{\Phi'(v)}{\Phi(v)} = \frac{1}{\sigma'} \Phi(-v), \tag{19}$$

where $\sigma' = \sigma \log_{10} \mathrm{e}$, and, using $\Phi(-v) = 1 - \Phi(v)$ we have

$$e_l(v_l) = h_l \psi\big(v_l\big) + (h_l - 1) \psi\big(-v_l\big) \tag{20}$$

$$= \frac{1}{\sigma'} [h_l - \Phi(v_l)]. \tag{21}$$

The solution obtained in (15) is based on the model (6)-(7) which requires $\boldsymbol{\theta}$ to remain constant throughout the time $l = 1, \ldots, n$. Since, in practice, the levels of the players may vary in time (the abilities evolve due to training, age, coaching strategies, fatigue, etc.), it is necessary to track $\boldsymbol{\theta}$.

To this end, arguably the simplest strategy relies on the stochastic gradient (SG) which differs from the steepest descent in the following elements: i) at time n only one iteration of the steepest descent is executed, ii) the gradient is calculated solely for the current observation term $L_n(\hat{\boldsymbol{\theta}}_n)$, and iii) the available estimate $\hat{\boldsymbol{\theta}}_n$ is used as the starting point for the update

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n - \mu \nabla_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}_n + \mu \boldsymbol{x}_n e_n(v_n) \tag{22}$$

$$= \hat{\boldsymbol{\theta}}_n + \mu \boldsymbol{x}_n [h_n - \Phi(v_n)] \tag{23}$$

$$= \hat{\boldsymbol{\theta}}_n - \mu \boldsymbol{x}_n [a_n - \Phi(-v_n)], \tag{24}$$

where μ is the adaptation step; with abuse of notation the fraction $\frac{1}{\sigma'}$ from (21) is absorbed by μ in (23)-(24).

In the rating context, $\boldsymbol{x}_n$ has only two non-zero terms, and therefore only the level of the players $i_{\mathrm{H},n}$ and $i_{\mathrm{A},n}$ will be modified. By inspection, the update (23)-(24) may be written as a single equation for any player $i \in \{i_{\mathrm{H},n}, i_{\mathrm{A},n}\}$

$$\hat{\theta}_{n+1,i} = \hat{\theta}_{n,i} + K\big[s_i - \Phi(\Delta_i)\big] \tag{25}$$

where $\Delta_i = \hat{\theta}_{n,i} - \hat{\theta}_{n,j}$ and j is the index of the player opposing the player i, i.e., $j \neq i, j \in \{i_{\mathrm{H},n}, i_{\mathrm{A},n}\}$; $s_i = \mathbb{I}\big[i \succ j\big]$ indicates if the player i won the game. Since the variables s_i and Δ_i are intermediary, on purpose we do not index them with

[2]The convexity comes from the fact that $-\log \Phi_{\mathsf{H}}(v)$ is convex in v (easy to demonstrate by hand) and thus $\log \Phi_{\mathsf{H}}(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\theta})$, being a concatenation of a convex and linear functions is also convex (Tsukida and Gupta, 2011, Appendix A).

[3]While the minimum is global, it is not unique due to the ambiguity of the origin θ_0 we mentioned at the end of Sec. II-A.

n.

We also replaced μ with K so that (25) has the form of the Elo algorithm as usually presented in the literature (Elo, 2008)(Langville and Meyer, 2012, Ch. 5). Thus, the Elo algorithm implements the SG to obtain the ML estimate of the levels $\boldsymbol{\theta}$ under the model (4). This has been noted before, e.g., in (Király and Qian, 2017).

We also note that, in the description of the Elo algorithm (Elo, 2008), s_i is defined as a numerical "score" attributed to the game outcome H or A. In a sense, it is a legacy of rating methods which attribute numerical value to the game result. On the other hand, in the modelling perspective we adopted, attribution of numerical values to the categorical variables H and A is not required.

Task: understand how ELO updates ratings after a game?

IV. Draws

We want to address now the issue of draws (ties) in the game outcome. We ignored it for clarity of development, but draws are important results of the game and must affect the rating, especially in sports when they occur frequently, such as international football, chess and many others sports and competitions (Langville and Meyer, 2012, Ch. 11). Some approaches in the literature go around this problem by ignoring the draws, other count them as partial wins/losses with fractional score $s_i = \frac{1}{2}$ (Langville and Meyer, 2012, Ch. 11)(Glickman and Hennessy, 2015). Such heuristics, while potentially useful, do not show explicitly how to predict the results of the games from the rating levels.

Thus, the preferred approach is to model the draws explicitly; we must, therefore, augment our model to include the conditional probability of draws

$$\Pr\{i \doteq j | \theta_i, \theta_j\} = \Phi_{\mathsf{D}}(\theta_i - \theta_j), \tag{26}$$

where by axiomatic requirement $\Phi_{\mathsf{D}}(v)$ should be decreasing with the absolute value of its argument, and be maximized for $v = 0$. The justification is that large absolute difference in levels increases the probability of win or loss, while the rating levels proximity, $\theta_i \approx \theta_j$, should increase the probability of a draw.

By the law of total probability we require now

$$\Phi_{\mathsf{H}}(v) + \Phi_{\mathsf{A}}(v) + \Phi_{\mathsf{D}}(v) = 1, \tag{27}$$

which obviously implies that considering the draws, the functions $\Phi_{\mathsf{H}}(v)$ and $\Phi_{\mathsf{A}}(v)$ also must change with respect to those used when analyzing the binary (win/loss) game results.

A. Explaining draws in the Elo algorithm

The Elo algorithm also considers draws by setting $s_i = \frac{1}{2}$ in (25) (Elo, 2008, Ch. 1.6) (Langville and Meyer, 2012, Ch. 5). However, the function $\Phi_{\mathsf{D}}(v)$ is undefined which is quite perplexing: the draws are accounted for but the model, which would allow us to calculate their probability from the parameters $\boldsymbol{\theta}$, is lacking. Moreover, the description of algorithm (25) still indicates that $\Phi(\Delta_i)$ is the "expected score" which cannot be calculated without explicit definition of the probability of draw. Despite this logical gap, the algorithm is being widely used and is considered reliable.

Our objective here, is thus to "reverse-engineer" the Elo algorithm and explain what probabilistic model is compatible with the operation of the algorithm. This will bridge the gap providing formal basis to interpret the results.

Proposition 1 (The Elo algorithm with draws)**.** *The Elo algorithm* (25) *which assigns the score value* $s_i = 1$ *to a win,* $s_i = 0$ *to a loss and* $s_i = \frac{1}{2}$ *to a draw, implements SG to estimate the rating levels* $\boldsymbol{\theta}$ *using the ML principle for model defined by the following conditional probabilities*

$$\Phi_{\mathsf{H}}(v) = \Phi^2(v), \quad \Phi_{\mathsf{A}}(v) = \Phi^2(-v) \tag{28}$$

$$\Phi_{\mathsf{D}}(v) = 2\Phi(v)\Phi(-v). \tag{29}$$

Proof. We start by squaring the equation of the total probability law for the binary-outcome game, $\Phi(v) + \Phi(-v) = 1$, to obtain

$$\Phi^2(v) + \Phi^2(-v) + 2\Phi(v)\Phi(-v) = 1 \tag{30}$$

and thus, using the assignment (28)-(29), we satisfy (27). This may appear arbitrary but we have to recall that the whole model for the binary outcome is built on assumptions which reflect our idea about the loss/win probabilities and indeed, the draw probability function $\Phi_{\mathsf{D}}(v)$ has the behaviour we expected: it has a maximum for $v = 0$ and decreases with growing $|v|$.

Now, each function in the model, $\Phi_{\mathsf{H}}(v)$, $\Phi_{\mathsf{A}}(v)$, and $\Phi_{\mathsf{D}}(v)$, is a non-trivial transformation of $\Phi(v)$.

Using (28)-(29), we rewrite (13) as

$$L_l(\boldsymbol{\theta}) = -\log \Pr\{Y_l = y_l|\boldsymbol{\theta}\} \tag{31}$$

$$= -h_l \log \Phi_{\mathsf{H}}(v_l) - a_l \log \Phi_{\mathsf{A}}(v_l) - d_l \log \Phi_{\mathsf{D}}(v_l) \tag{32}$$

$$= -2h_l \log \Phi(v_l) - 2a_l \log \Phi(-v_l) - d_l \Big[\log \Phi(v_l) + \log \Phi(-v_l) \Big] \tag{33}$$

so the gradient is calculated as in (17)

$$\nabla_{\boldsymbol{\theta}} L_l(\boldsymbol{\theta}) = -2\tilde{h}_l \boldsymbol{x}_l \psi(v_l) + 2\tilde{a}_l \boldsymbol{x}_l \psi(-v_l) \tag{34}$$

$$= -2\boldsymbol{x}_l \tilde{e}_l(\boldsymbol{\theta}) \tag{35}$$

where $\tilde{h}_l = h_l + d_l/2$, $\tilde{a}_l = a_l + d_l/2 = 1 - \tilde{h}_l$, and

$$\tilde{e}_l(\boldsymbol{\theta}) = \tilde{h}_l - \Phi(v_l). \tag{36}$$

We thus recover the same equations as in the binary-result game, splitting the draw indicator, d_l, equally between the indicators of the home and away wins; we can reuse them directly in (23)-(23)

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \mu \boldsymbol{x}_n [\tilde{h}_n - \Phi(v_n)] \tag{37}$$

$$= \hat{\boldsymbol{\theta}}_n - \mu \boldsymbol{x}_n [\tilde{a}_n - \Phi(-v_n)], \tag{38}$$

which yields the same update as the Elo algorithm (25)

$$\hat{\theta}_{n+1,i} = \hat{\theta}_{n,i} + K\big[s_i - \Phi(\Delta_i)\big], \tag{39}$$

with the new definition of the score $s_i = \tilde{h}_n$ (for the home player) and $s_i = \tilde{a}_n$ (for the away player), and where for compatibility

of equations, the update step K absorbed the multiplication by 2 (the only difference between (35) and (18)). □

The following observations are in order:

1) We unveiled the implicit model behind the Elo algorithm thus, our findings do not affect the operation of the algorithm but rather clarify how to interpret its results. Namely, given the estimate of the levels $\hat{\boldsymbol{\theta}}_n$ the probability of the game outcomes should be estimated as

$$\Pr\left\{i \succcurlyeq j|\hat{\theta}_i, \hat{\theta}_j\right\} = \Phi^2(\hat{\theta}_i - \hat{\theta}_j) \tag{40}$$

$$\Pr\left\{i \preccurlyeq j|\hat{\theta}_i, \hat{\theta}_j\right\} = \Phi^2(\hat{\theta}_j - \hat{\theta}_i) \tag{41}$$

$$\Pr\left\{i \doteq j|\hat{\theta}_i, \hat{\theta}_j\right\} = 2\Phi(\hat{\theta}_i - \hat{\theta}_j)\Phi(\hat{\theta}_j - \hat{\theta}_i). \tag{42}$$

2) We emphasize that s_i is the indicator of the result but using (28)-(29) we can again calculate its expected value for $i = i_{\mathrm{H},n}$

$$\mathbb{E}_{Y_l|\hat{\theta}_{l,i},\hat{\theta}_{l,j}}[s_i(Y_l)] = \Pr\left\{i \succcurlyeq j|\hat{\theta}_i, \hat{\theta}_j\right\} + \frac{1}{2}\Pr\left\{i \doteq j|\hat{\theta}_i, \hat{\theta}_j\right\} \tag{43}$$

$$= \left[\Phi(\Delta_i)\right]^2 + \Phi(\Delta_i)\Phi(-\Delta_i) \tag{44}$$

$$= \Phi(\Delta_i)\left[\Phi(\Delta_i) + \Phi(-\Delta_i)\right] = \Phi(\Delta_i); \tag{45}$$

the same can be straightforwardly done for $i = i_{\mathrm{A},n}$.

Thus, indeed, the function $\Phi(\Delta_i)$ in the Elo update (25) has the meaning of the expected score. It has not been spelled out mathematically up to now—most likely—because the draws has been only implicitly considered. Nevertheless, with formidable intuition, the description of the Elo algorithm defines correctly the terms without making reference to the underlying probabilistic model.

We note, again, that the notion of expected score is not necessary in the development of the SG algorithm and the fact that the score takes the fractional value $s_i = \frac{1}{2}$ is a result of the particular form of the conditional probability (29) and our decision to make K absorb the multiplication by 2, see (35).

While the clarification we made regarding the meaning of the expected score is useful, the first observation above is the most important for the explicit interpretation of the results of the algorithm. Recall that, in the win-loss game, the function $\Phi(\Delta_i)$ has the meaning of the probability of winning the game, see (4). However, in the win-draw-loss model, such interpretation is incorrect because the probability of winning the game is given by (40) which we just derived. As we will see in the numerical examples, using the latter, however, provides poor results.

This surprising confusion persisted through time because the model we have shown in (28)-(29) is merely implicit in the Elo algorithm and the explicit derivation of the algorithm (Elo, 2008, Chap. 8) did not consider the draws in the formal probabilistic framework. Other works, e.g., (Glickman, 1999) (Lasek, Szlávik, and Bhulai, 2013), observed this conceptual difficulty before. In particular, (Glickman, 1999, Sec. 2) used $\Phi_{\mathsf{T}}(v) = \sqrt{\Phi(v)\Phi(-v)}$ but kept the legacy of the win-loss model, i.e., $\Phi_{\mathsf{H}}(v) = \Phi(v)$ and $\Phi_{\mathsf{A}}(v) = \Phi(-v)$, which leads to approximate solutions because (27) is violated.

The lesson learned is that, despite the apparent simplicity of the Elo algorithm, we should resist the temptation to tweak its parameters. While using the fractional score value $s_i = \frac{1}{2}$ for the draw is now explained, we cannot guarantee that modifying

s_i in arbitrary manner will correspond to a particular probabilistic model. Therefore, rather than tweaking the SG/Elo algorithm (25), the modification should start with the probabilistic model itself.

B. Generalization of the Elo algorithm

Having unveiled the implicit modeling of draws underlying the Elo algorithm we immediately face a new problem. Namely, considering the draws, we have three events (and thus two independent probabilities to estimate) but the Elo algorithm has no additional degree of freedom to take this reality into account. For example, using (40)-(42) the results of the game between the players with equal rating levels $\hat{\theta}_i = \hat{\theta}_j$, will always be predicated as $\Pr\left\{i \gtrdot j|\hat{\theta}_i, \hat{\theta}_j\right\} = 0.25$ and $\Pr\left\{i \doteq j|\hat{\theta}_i, \hat{\theta}_j\right\} = 0.5$. The Elo algorithm does that implicitly, but there is no real reason to stick to such a rigid solution which may produce an inadequate fit to the observed data, and a more general approach is necessary.

One of the workarounds proposed by (Rao and Kupper, 1967) and used later, e.g., (Fahrmeir and Tutz, 1994)(Herbrich and Graepel, 2006)(Király and Qian, 2017), modifies the model using a threshold value $v_0 \geq 0$

$$\Phi_{\mathsf{H}}(v) = \Phi(v - v_0), \quad \Phi_{\mathsf{A}}(v) = \Phi(-v - v_0), \quad \Phi_{\mathsf{D}}(v) = \Phi(v + v_0) - \Phi(v - v_0). \tag{46}$$

While (46) is definitely useful and solves formally the problem which is more general than the case of binary outcome game, we do not treat it as a generalization of the Elo algorithm itself, because there is no parameter v_0 which transforms (46) into (28)-(29) (which, as we demonstrated, is the model behind the Elo algorithm).

Here we propose to use the model of (Davidson, 1970) which can be defined as

$$\Phi_{\mathsf{H}}(v) = \Phi_\kappa(v) = \frac{10^{0.5v/\sigma}}{10^{0.5v/\sigma} + 10^{-0.5v/\sigma} + \kappa} \tag{47}$$

$$\Phi_{\mathsf{A}}(v) = \Phi_\kappa(-v) = \frac{10^{-0.5v/\sigma}}{10^{0.5v/\sigma} + 10^{-0.5v/\sigma} + \kappa} \tag{48}$$

$$\Phi_{\mathsf{D}}(v) = \kappa\sqrt{\Phi_{\mathsf{H}}(v)\Phi_{\mathsf{A}}(v)} = \frac{\kappa}{10^{0.5v/\sigma} + 10^{-0.5v/\sigma} + \kappa}, \tag{49}$$

where $\kappa \geq 0$ is a freely set draw parameter.

We hasten to say that the model (47)-(49) is not necessarily better in the sense of fitting to the data than (46). Our motivation to adopt (47)-(49) is the fact that these equations generalize previous models. Namely, for $\kappa = 0$ we obtain the win-loss model behind the Elo algorithm shown in (25), while using $\kappa = 2$ yields

$$\Phi_{\mathsf{H}}(v) = \frac{10^{0.5v/\sigma}}{\left(10^{0.25v/\sigma} + 10^{-0.25v/\sigma}\right)^2} = \Phi^2(v/2) \tag{50}$$

which, up to the scale factor σ, corresponds to the implicit win-draw-loss model behind the Elo algorithm we have shown in (28)-(29).

In other words, the implicit model for the Elo algorithm is based on the explicit modeling of draws proposed by (Davidson, 1970) if we set a particular value of the draw parameter ($\kappa = 2$).

1) Adaptation: We quickly note that the function $-\log \Phi_\kappa(v)$ is convex so the gradient-based adaptation will converge under adequate choice of the step μ.

To derive the adaptation algorithm we recalculate (31)

$$L_l(\boldsymbol{\theta}) = -\log \Pr\{Y_l = y_l|\boldsymbol{\theta}\} \tag{51}$$

$$= -h_l \log \Phi_{\mathsf{H}}(v_l) - a_l \log \Phi_{\mathsf{A}}(v_l) - d_l \log \Phi_{\mathsf{D}}(v_l) \tag{52}$$

$$= -\tilde{h}_l \log \Phi_{\mathsf{H}}(v_l) - \tilde{a}_l \log \Phi_{\mathsf{H}}(-v_l) \tag{53}$$

and the gradient is given by

$$\nabla_\theta L_l(\boldsymbol{\theta}) = -e_l(v_l)\boldsymbol{x}_l \tag{54}$$

where

$$e_l(v_l) = \tilde{h}_l \psi_\kappa(v_l) + (\tilde{h}_l - 1)\psi_\kappa(-v_l) \tag{55}$$

$$\psi_\kappa(v) = \frac{\Phi'_\kappa(v)}{\Phi_\kappa(v)} = \frac{1}{\sigma'} \frac{10^{-0.5v/\sigma} + \frac{1}{2}\kappa}{10^{0.5v/\sigma} + 10^{-0.5v/\sigma} + \kappa} = \frac{1}{\sigma'} F_\kappa(-v), \tag{56}$$

where, as before $\sigma' = \sigma \log_{10} \mathrm{e}$, and we define

$$F_\kappa(v) = \frac{10^{v/2} + \frac{1}{2}\kappa}{10^{v/2} + 10^{-v/2} + \kappa} = 1 - F_\kappa(-v), \tag{57}$$

and thus

$$e_l(v_l) = \frac{1}{\sigma'}\Big(\tilde{h}_l F_\kappa(-v_l) + (\tilde{h}_l - 1)F_\kappa(v_l)\Big) \tag{58}$$

$$= \frac{1}{\sigma'}\Big(\tilde{h}_l - F_\kappa(v_l)\Big). \tag{59}$$

Using (59) in (22) yields the same equations as in (39) except that $\Phi(v)$ must be replaced with $F_\kappa(v)$ and the division by σ' should be absorbed by the adaptation step. This yields a new κ-Elo rating algorithm

$$\hat{\theta}_{n+1,i} = \hat{\theta}_{n,i} + K\big[s_i - F_\kappa(\Delta_i)\big], \tag{60}$$

where as before i) $\Delta_i = \hat{\theta}_i - \hat{\theta}_j$ (j being the index of the player opposing the player i), ii) as in the Elo algorithm, K is maximum increase/decrease step, and iii) $s_i \in \{0, \frac{1}{2}, 1\}$ indicates the outcome of the game, i.e., the score.

The new κ-Elo algorithm is equally simple as the Elo algorithm, yet provides us with the flexibility to model the relationship between the draws and the wins via the draw parameter $\kappa \geq 0$. We recall that, in fact, the Elo algorithm is a particular version of κ-Elo for $\kappa = 2$. We provide numerical examples in Sec. V to illustrate its properties.

2) κ in κ-Elo algorithm: insights and pitfalls: Can we say something about the draw parameter, κ, without implementing and running the κ-Elo algorithm defined by (60)? The answer is yes, if we suppose that the fit we obtain is (almost) perfect, i.e., the average empirical probabilities averaged over a large time window

$$\overline{p}_{\mathsf{H}} = \frac{1}{N}\sum_{l=1}^{N} \mathbb{I}\big[y_l = \mathsf{H}\big], \quad \overline{p}_{\mathsf{A}} = \frac{1}{N}\sum_{l=1}^{N} \mathbb{I}\big[y_l = \mathsf{A}\big], \quad \overline{p}_{\mathsf{D}} = \frac{1}{N}\sum_{l=1}^{N} \mathbb{I}\big[y_l = \mathsf{D}\big], \tag{61}$$

can be deduced from the functions (47)-(49) using the estimated rating levels $\hat{\boldsymbol{\theta}}$.[4] If this is the case they should stay in the relationship prescribed by the model (49), i.e., $\overline{p}_{\mathsf{D}} \approx \kappa\sqrt{\overline{p}_{\mathsf{H}}\overline{p}_{\mathsf{A}}}$.

Denoting the difference between the frequency of home and away wins as $\overline{\delta} = \overline{p}_{\mathsf{H}} - \overline{p}_{\mathsf{A}}$, and from the law of total (empirical) probability we obtain $\overline{p}_{\mathsf{H}} = \frac{1}{2}(1 - \overline{p}_{\mathsf{D}} + \overline{\delta})$ and $\overline{p}_{\mathsf{A}} = \frac{1}{2}(1 - \overline{p}_{\mathsf{D}} - \overline{\delta})$ from which

$$\overline{p}_{\mathsf{D}} \approx \frac{\kappa}{2}\sqrt{(1 - \overline{p}_{\mathsf{D}})^2 - \overline{\delta}^2} \tag{62}$$

and thus, for the relatively small values of home/away imbalance, e.g., $\overline{\delta} < 0.2$ we can ignore the term $\overline{\delta}^2$ which allows us simply say what is implicit assumption about $\overline{p}_{\mathsf{D}}$ for arbitrary κ

$$\overline{p}_{\mathsf{D}} \approx \frac{\kappa}{2 + \kappa}. \tag{63}$$

Thus using $\kappa = 2$ (as done implicitly in the current rating of FIDE and FIFA), suggests that the $\overline{p}_{\mathsf{D}} \approx 0.5$. Since this is not the case in none of the competitions where these rating are used, we can expect that, when implementing the new rating algorithm with a more appropriate value of κ, FIDE and FIFA will improve the fit to the results in the sense of better estimation of the probabilities of win, loss, and draw.

We can also estimate the suitable value of κ as

$$\overline{\kappa} \approx \frac{2\overline{p}_{\mathsf{D}}}{1 - \overline{p}_{\mathsf{D}}}. \tag{64}$$

For example, using $\overline{p}_{\mathsf{D}} \approx 0.25$ (which was the average frequency of draws in English Premier Ligue football games over ten seasons, see Sec. V) we would find $\overline{\kappa} \approx 0.7$.

Is this value acceptable?

Before answering this question, we have to point to a particular problem that can arise in the modelling of the draws. Namely, the current formulations known in the literature (the threshold-based (46) or the one we used (47)-(49)), do not explicitly constrain the relationship between the predicted *values* of probabilities. Of course, we always keep the relationship $\Phi_{\mathsf{A}}(v) = \Phi_{\mathsf{H}}(-v)$. Thus, considering the case $v_l = \hat{\theta}_i - \hat{\theta}_j = \epsilon$ (where $\epsilon > 0$ is a small rating difference), we have $\Phi_{\mathsf{H}}(\epsilon) \geq \Phi_{\mathsf{A}}(\epsilon)$ and our intuition follows: it is more probable that a stronger home player wins than he looses.

On the other hand, it is not clear what should be said about the probability of the draw in such a case. Should we expect the probability of draw to be larger than the probability of home/away win? For example, is it acceptable to obtain the values $\Phi_{\mathsf{H}}(\epsilon) = 0.42$, $\Phi_{\mathsf{A}}(\epsilon) = 0.38$ and $\Phi_{\mathsf{D}}(\epsilon) = 0.20$? Nothing prevents such results in the model we use (and, to our knowledge, in other models used before) and the interpretation is counterintuitive: the stronger home player is more likely to loose than to draw.

Therefore, we might want to remove such results from the solution space: for equal-rating players we force the probability

[4]Such statistics may be obtained from previous seasons. While they do not change drastically through seasons and may be treated as a prior, in the case of on-line rating, they may also be estimated from the recent past. However, we do not follow this idea further in this work.

of the draw to be larger than the probability of home/away wins, we thus have to use κ which satisfies

$$\Phi_{\mathsf{D}}(0) > \Phi_{\mathsf{W}}(0) \tag{65}$$

$$\kappa \geq 1, \tag{66}$$

where the last inequality follows from (47)-(49). This is an important restriction and forces us to model the draws occurring with (a large) frequency $\overline{p}_{\mathsf{D}} \geq 0.33$, see (63). While it seems unsound to use the mismatched model, we don't know its impact on the prediction capability and yet, we have to remember, that the current version of the Elo algorithms uses $\kappa = 2$. We have no clear answer to this question and will seek more insight in the numerical examples.

V. Numerical Examples

We illustrate the operation of the algorithms using the results from the England Premier League football games available at (Football-data.co.uk, 2019). In this context, there are $M = 20$ teams playing against each other in one home- and one away-games. We consider one season at the time, thus $n = 1, \ldots, N$, index the games in the chronological order, and $N = M(M-1) = 380$.

Football (and other) games are known to produce the so-called home-field advantage, where the home wins $\{y_n = \mathsf{H}\}$ are more frequent than the away wins $\{y_n = \mathsf{A}\}$. In the rating context, this is modelled by artificially increasing the level of the home player, which corresponds, de facto, to left-shifting of the conditional probability functions

$$\Phi^{\text{hfa}}_{\mathsf{H}}(v) = \Phi_{\mathsf{H}}(v + \eta\sigma), \quad \Phi^{\text{hfa}}_{\mathsf{A}}(v) = \Phi_{\mathsf{A}}(v + \eta\sigma), \quad \Phi^{\text{hfa}}_{\mathsf{D}}(v) = \Phi_{\mathsf{D}}(v + \eta\sigma), \tag{67}$$

where home-field advantage parameter $\eta \geq 0$ should be adequately set; its value is independent of the scale thanks to multiplication by σ.[5]

As in FIFA rating algorithm, (FIFA, 2019), we set $\sigma = 600$; the levels are initialized at $\theta_{0,m} = 0$; as we said before these values are arbitrary. In what follows we always use the normalization $K = \tilde{K}\sigma$ which removes the dependence on the scale: for a given $\tilde{K}$ the prediction results will be exactly the same even if we change the value of σ.

An example of the estimated ratings $\theta_{n,m}$ for a group of teams is shown in Fig. 1 to illustrate the fact that quite a large portion of time in the beginning of the season is dedicated to the convergences of the algorithm; this is the "learning" period. Of course, using larger step $\tilde{K}$ we can accelerate the learning at the cost of increased variability of the rating. These well-known issues are related to the operation of SG but solving them is out of the scope of this work. We mention them mostly because, to evaluate the performance of the algorithms, we decide to use the second half of the season, where we assume the algorithms converged and the rating levels follow the performance of the teams. This is somewhat arbitrary of course, but our goal here is to show the influence of the draw-parameter and not to solve the entire problem of convergence/tracking in SG/Elo algorithms.

For concision, the estimated probability of the game result $\{\mathsf{H}, \mathsf{A}, \mathsf{D}\}$ calculated before the game at the time l using the rating levels $\hat{\boldsymbol{\theta}}_{l-1}$ obtained at the time $l-1$, is denoted as

$$\hat{p}_{l,\mathsf{H}} = \Phi_{\mathsf{H}}(\boldsymbol{x}_l^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{l-1}), \quad \hat{p}_{l,\mathsf{A}} = \Phi_{\mathsf{A}}(\boldsymbol{x}_l^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{l-1}), \quad \hat{p}_{l,\mathsf{D}} = \Phi_{\mathsf{D}}(\boldsymbol{x}_l^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{l-1}). \tag{68}$$

[5]We note that this version of the equation is slightly different from (Davidson and Beaver, 1977, Eq. 2.4); with our formulation, the relationship (64) is not affected by the home-field advantage parameter η.

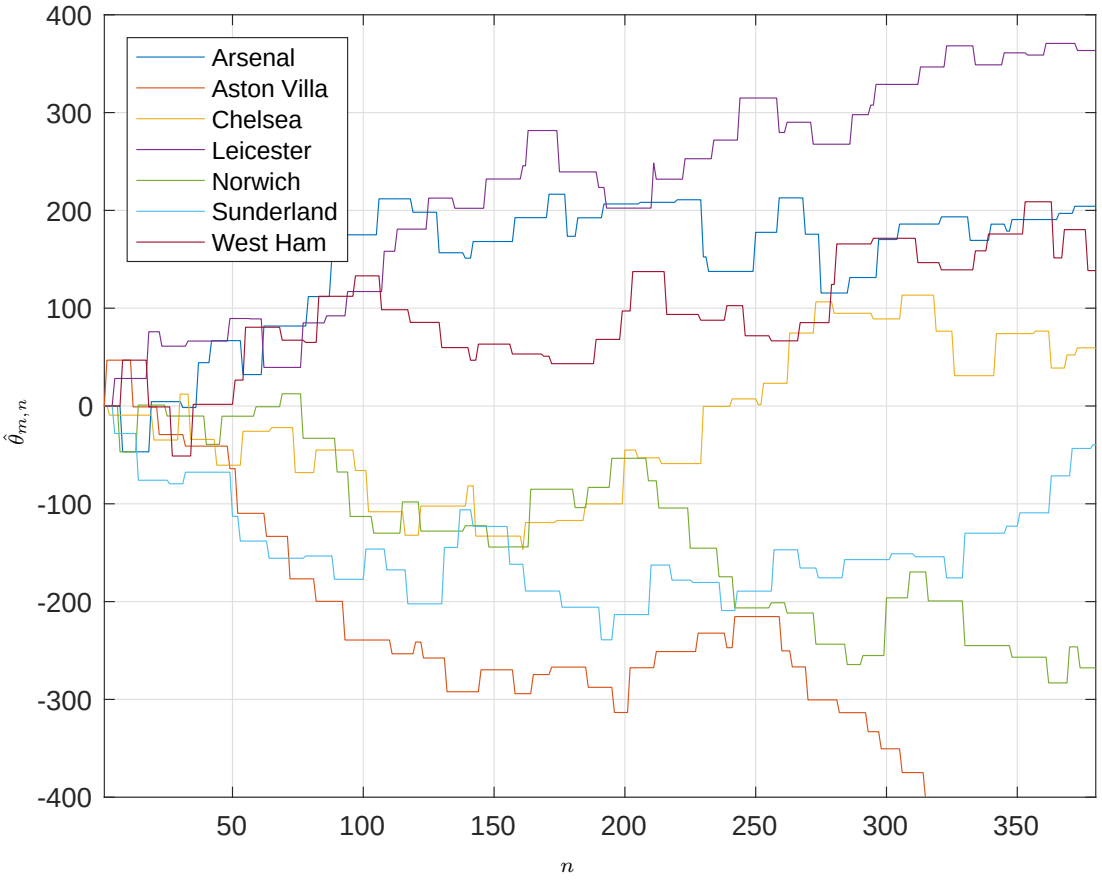

Fig. 1. Evolution of the rating levels $\hat{\theta}_{m,n}$ for selected English Premier League teams in the season 2015; $N = 380$, $\sigma = 600$, $\tilde{K} = 0.125$, $\eta = 0.3$, $\kappa = 0.7$. We assume that, the first half of the season absorbs the learning phase, and the tracking of the teams' levels in the second half is free of the initialization effect.

We show the (negative) logarithmic score (Gelman, Hwang, and Vehtari, 2014) averaged over the second-half of the season

$$\overline{\mathrm{LS}} = \frac{2}{N} \sum_{l=N/2+1}^{N} \mathrm{LS}_l. \tag{69}$$

where

$$\mathrm{LS}_l = -(h_l \log \hat{p}_{l,\mathsf{H}} + a_l \log \hat{p}_{l,\mathsf{A}} + d_l \log \hat{p}_{l,\mathsf{D}}). \tag{70}$$

We still have to define the prediction of the draw in the conventional Elo algorithm: we cannot set $\hat{p}_{l,\mathsf{D}}) = \Phi_{\mathsf{D}}(v) \equiv 0$, of course, because it would result in infinite logarithmic score. We thus follow the heuristics of (Lasek et al., 2013) which may be summarized as follows: the conventional Elo algorithm is used to find the rating levels (i.e., $\kappa = 2$ is used in κ-Elo), but the prediction is based on $\Phi_{\mathsf{H}}(v)$, $\Phi_{\mathsf{A}}(v)$, and $\Phi_{\mathsf{D}}(v)$ with a different value of the draw-parameter $\kappa = \check{\kappa}$. This may be seen as a model mismatch between estimation and prediction. We follow (Lasek et al., 2013) and apply $\check{\kappa} = 1$; this correspond to $\overline{p}_{\mathsf{D}} \approx 0.33$ and also is the minimum value of κ which guarantees (65).

We show in Fig. 2 the logarithmic score $\overline{\mathrm{LS}}$ for different values of the draw parameter κ, and normalized step $\tilde{K}$. We compare our predictions with those based on the probabilities inferred from the odds of the betting site Bets365 available,

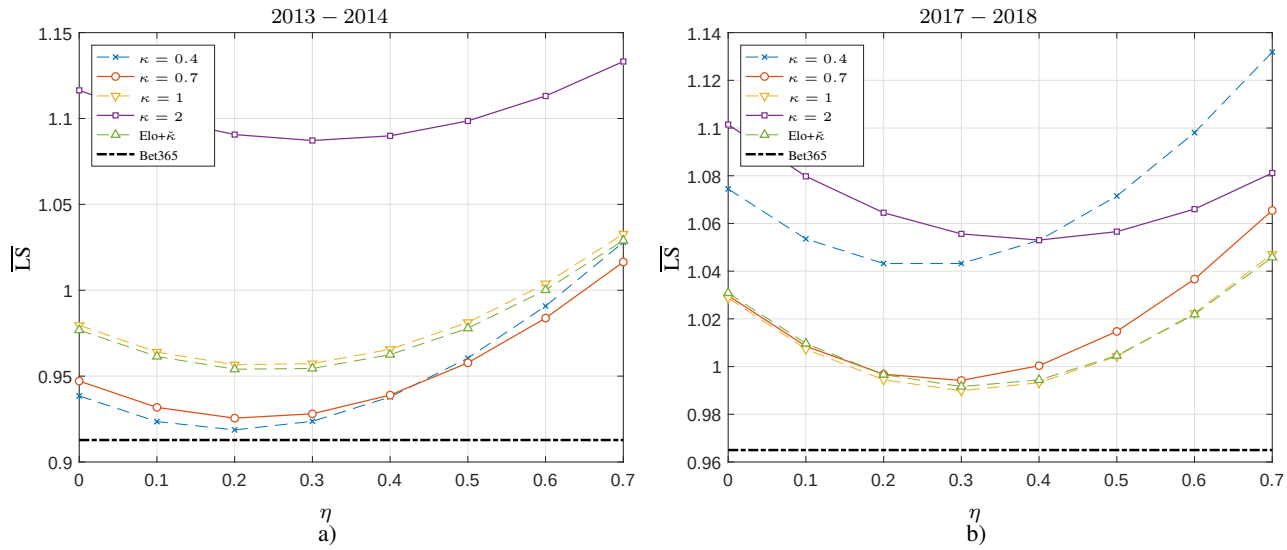

Fig. 2. Logarithmic score, (69), in second half of two seasons of English Premier League with $\sigma = 600$, $\tilde{K} = 0.125$, different values of κ indicated in the legend, and varying the home-advantage parameter, η; a) season 2013-2014, where $\overline{p}_{\mathsf{D}} = 0.17$ and thus using (64), we obtain $\overline{\kappa} \approx 0.40$; b) season 2017-2018, where $\overline{p}_{\mathsf{D}} = 0.26$, and thus $\overline{\kappa} \approx 0.7$. The results "Elo+$\check{\kappa}$" are obtained from the conventional Elo algorithm but $\check{\kappa} = 1$ is used in the prediction. The results "Bet365" are based on the probabilities inferred from the betting odds offered by the site Bet365.

together with the game results, at (Football-data.co.uk, 2019).[6] These are constant reference lines in Fig. 2 as they, of course, do not vary with the parameters we adjust.

We observe that introducing the draw parameter κ we improved the logarithmic score. On the other hand, using κ-Elo algorithm with $\kappa = 2$ yields particularly poor results if we explicit the model (and thus use $\kappa = 2$ for the prediction); a much better solution is to use a mismatched model and apply $\check{\kappa} = 1$; the results obtained are, in general very close to those obtained using κ-Elo algorithm especially when used with $\kappa = 1$. In the season 2013-2014, where frequency of draws was low, using the corresponding value $\kappa = \overline{\kappa}$ provided notable improvement comparing to large $\kappa \in \{1, 2\}$.

Finally, we show the comparison across various seasons in Table I where, beside the score $\overline{\mathrm{LS}}$ we also show the pseudo-credibility interval $(\overline{\mathrm{LS}}_{\mathrm{low}}, \overline{\mathrm{LS}}_{\mathrm{high}})$; this is the the minimum-length interval in which 95% of the data was found.[7] We observe that using $\kappa = 1$ does not incur a large penalty when compared to $\kappa = 0.7$ even if the latter matches closely the observed frequency of draws. The differences may be observed only for seasons where the low frequency of draws implies very small $\overline{\kappa}$, e.g., in 2013-2014 and 2018-2019.

On the other hand, the length of the credibility intervals is slightly smaller for $\kappa = 1$, indicating a better prediction "stability" across time. Similar average results may be obtained using the Elo algorithm with $\check{\kappa} = 1$ which produces also slightly larger credibility intervals.

The results obtained with this rather limited set of data stay in line with our previous theoretical discussion, indicating at the same time that no dramatic change in performance should be expected by using κ-Elo. Nevertheless, an improvement can be obtained by using the conservative value of $\kappa = 1$. This recommendation is motivated by the discussion in Sec. IV-B2 and comes at no implementation cost.

[6]This is done as in (Király and Qian, 2017): the published decimal odds for the three events, o_{H}, o_{A}, and o_{D}, are used to infer the probabilities, $\tilde{p}_{\mathsf{H}} \propto 1/o_{\mathsf{H}}$, $\tilde{p}_{\mathsf{A}} \propto 1/o_{\mathsf{A}}$, and $\tilde{p}_{\mathsf{D}} \propto 1/o_{\mathsf{D}}$; these are next normalized to make them sum to one (required as the betting odds are not "fair" and include the bookie's overhead, the so-called vigorish).

[7]We find it more informative than derivation of credibility intervals using unknown statistics.

Season	$\overline{\kappa}$	Bet365	κ-Elo, $\kappa = 0.7$	κ-Elo, $\kappa = 1$	Elo+$\check{\kappa}$
2009-2010	0.71	0.91 ∈ (0.16,1.68)	0.93 ∈ (0.19,1.64)	0.93 ∈ (0.26,1.68)	0.93 ∈ (0.20,1.62)
2010-2011	0.73	0.97 ∈ (0.23,1.64)	1.01 ∈ (0.29,1.69)	1.01 ∈ (0.35,1.66)	1.01 ∈ (0.32,1.70)
2011-2012	0.59	0.99 ∈ (0.16,1.87)	0.98 ∈ (0.16,1.70)	1.00 ∈ (0.22,1.73)	1.00 ∈ (0.17,1.72)
2012-2013	0.73	0.95 ∈ (0.24,1.66)	1.01 ∈ (0.22,1.82)	1.01 ∈ (0.26,1.90)	1.00 ∈ (0.22,1.92)
2013-2014	0.42	0.91 ∈ (0.14,1.98)	0.93 ∈ (0.17,1.86)	0.96 ∈ (0.21,1.82)	0.95 ∈ (0.20,1.92)
2014-2015	0.55	0.96 ∈ (0.21,1.66)	1.00 ∈ (0.22,1.88)	1.02 ∈ (0.27,1.94)	1.03 ∈ (0.23,2.00)
2015-2016	0.77	1.00 ∈ (0.27,1.73)	1.02 ∈ (0.22,1.77)	1.01 ∈ (0.27,1.78)	1.01 ∈ (0.24,1.86)
2016-2017	0.57	0.91 ∈ (0.15,1.91)	0.93 ∈ (0.19,1.92)	0.94 ∈ (0.19,1.78)	0.94 ∈ (0.15,1.82)
2017-2018	0.75	0.97 ∈ (0.14,1.91)	0.99 ∈ (0.18,1.78)	0.99 ∈ (0.23,1.78)	0.99 ∈ (0.19,1.86)
2018-2019	0.42	0.91 ∈ (0.17,1.91)	0.93 ∈ (0.17,1.80)	0.96 ∈ (0.21,1.87)	0.96 ∈ (0.17,1.96)

TABLE I

LOGARITHMIC SCORE $\overline{\mathrm{LS}}$, (69), IN TEN SEASONS OF ENGLISH PREMIER LEAGUE; $\sigma = 600$, $\tilde{K} = 0.125$, $\eta = 0.3$. THE RESULTS "ELO+$\check{\kappa}$" ARE OBTAINED FROM THE CONVENTIONAL ELO ALGORITHM BUT $\check{\kappa} = 1$ IS USED IN THE PREDICTION. THE RESULTS "BET365" ARE BASED ON THE PROBABILITIES INFERRED FROM THE BETTING ODDS OFFERED BY THE SITE BET365.

VI. CONCLUSIONS

In this paper we were mainly concerned with explaining the rationale and mathematical foundation behind the Elo algorithm. The whole discussion may be summarized as follows:

- We explained that, in the binary-outcome games (win-loss), the Elo algorithm is an instance of the well-known stochastic gradient algorithm applied to solve the ML estimation of the rating levels. This observation already appeared in the literature, e.g., (Király and Qian, 2017) so it was made for completeness but also to lay ground for further discussion.
- We have shown the implicit model behind the algorithm in the case of the games with draws. Although the algorithm has been used for decades in this type of games, the model of the draws has not been shown, impeding, de facto, the formal prediction of their probability. We thus filled this logical gap.
- We proposed a natural generalization of the Elo algorithm obtained from the well-known model proposed by (Davidson, 1970); the resulting algorithm, which we call κ-Elo, has the same simplicity as the original Elo algorithm, yet provides additional parameter to adjust to the frequency of draws. By extension, we revealed that the implicit model behind the Elo algorithm assumes that the frequency of draws is equal to 50%.
- We briefly discussed the constraints on the relationship between the values of draw and loss probabilities for the players with similar ratings; more precisely, we postulate that, in such a case, the draw probability should be larger than the probability of win/loss. While the discussion on such constraints has been absent from the literature, we feel it deserves further analysis to construct suitable models and algorithms for rating. Applying these constraints to the κ-Elo algorithm yields $\kappa \geq 1$. This is clearly a limitation which will produce a mismatch between the results and the model if the frequency of draws is less than 33%.
- To illustrate the main concepts we have shown numerical examples based on the results of the international football games in English Premier League.
- Finally, we conclude that, while in the past, the Elo algorithms has satisfied to a large extent the demand for simple rating algorithms, it is still possible to provide better, more flexible, and yet simple solutions. In particular the κ-Elo is better in a sense of taking the frequency of draws into account without compromising the complexity of implementation.

REFERENCES

Bradley, R. A. and M. E. Terry (1952): "Rank analysis of incomplete block designs: 1 the method of paired comparisons," *Biometrika*, 39, 324–345.

Caron, F. and A. Doucet (2012): "Efficient Bayesian inference for generalized Bradley–Terry models," *Journal of Computational and Graphical Statistics*, 21, 174–196, URL https://doi.org/10.1080/10618600.2012.638220.

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

David, H. (1963): *The Method of Paired Comparison*, Charles Griffin & Co. Ltd.

Davidson, R. R. (1970): "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments," *Journal of the American Statistical Association*, 65, 317–328, URL http://www.jstor.org/stable/2283595.

Davidson, R. R. and R. J. Beaver (1977): "On extending the Bradley-Terry model to incorporate within-pair order effects," *Biometrics*, 33, 693–702.

Elo, A. E. (2008): *The Rating of Chess Players, Past and Present*, Ishi Press International.

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 +http://dx.doi.org/10.1093/biomet/39.3-4.324.

FIFA (2019): "Fédération international de football association: men's ranking procedure," URL https://www.fifa.com/fifa-world-ranking/procedure/men.

Football-data.co.uk (2019): "Historical football results and betting odds data," URL https://www.football-data.co.uk/data.php.

Gelman, A., J. Hwang, and A. Vehtari (2014): "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, 24, 997–1016, URL https://doi.org/10.1007/s11222-013-9416-2.

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 http://dx.doi.org/10.1111/1467-9876.00159.

Glickman, M. E. and J. Hennessy (2015): "A stochastic rank ordered logit model for rating multi-competitor games and sports," *Journal of Quantitative Analysis in Sports*, 11.

Haykin, S. (2002): *Adaptive Filter Theory*, Prentice Hall, 4 edition.

Herbrich, R. and T. Graepel (2006): "Trueskill(TM): A Bayesian skill rating system," Technical report, URL https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system-2/.

Joe, H. (1990): "Extended use of paired comparison models, with application to chess rankings," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39, 85–93, URL http://www.jstor.org/stable/2347814.

Király, F. J. and Z. Qian (2017): "Modelling Competitive Sports: Bradley-Terry-Elo Models for Supervised and On-Line Learning of Paired Competition Outcomes," *arXiv e-prints*, arXiv:1701.08055.

Langville, A. N. and C. D. Meyer (2012): *Who's #1, The Science of Rating and Ranking*, Princeton University Press.

Lasek, J., Z. Szlávik, and S. Bhulai (2013): "The predictive power of ranking systems in association football," *International Journal of Applied Pattern Recognition*, 1, 27–46, URL https://www.inderscienceonline.com/doi/abs/10.1504/IJAPR.2013.052339, pMID: 52339.

Rao, P. V. and L. L. Kupper (1967): "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," *Journal of the American Statistical Association*, 62, 194–204, URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1967.10482901.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986): "Learning representations by back-propagating errors," *Nature*, 323, 533–536, URL https://doi.org/10.1038/323533a0.

Thurston, L. L. (1927): "A law of comparative judgement," *Psychological Review*, 34, 273–286.

Toulis, P. and E. M. Airoldi (2014): "Asymptotic and finite-sample properties of estimators based on stochastic gradients," *arXiv e-prints*, arXiv:1408.2923.

Tsukida, K. and M. Gupta (2011): "How to analyze paired comparison data," Technical report, University of Washington.

Wikipedia contributors (2019): "Wikipedia: elo rating system," URL https://en.wikipedia.org/wiki/Elo_rating_system.